

An on-line platform for ground truthing and performance evaluation of text extraction systems

Dimosthenis Karatzas, Sergi Robles, Lluís Gomez
Computer Vision Centre
Universitat Autònoma de Barcelona, Spain
Email: {dimos, srobles, lgomez}@cvc.uab.es

Abstract—This paper presents a set of on-line software tools for creating ground truth and calculating performance evaluation metrics for text extraction tasks such as localization, segmentation and recognition. The platform supports the definition of comprehensive ground truth information at different text representation levels while it offers centralised management and quality control of the ground truthing effort. It implements a range of state of the art performance evaluation algorithms and offers functionality for the definition of evaluation scenarios, on-line calculation of various performance metrics and visualisation of the results. The presented platform, which comprises the backbone of the ICDAR 2011 (challenge 1) and 2013 (challenges 1 and 2) Robust Reading competitions, is now made available for public use.

I. INTRODUCTION

Text extraction has received increasing attention over the past decade. End-to-end text extraction systems comprise a variable number of smaller tasks, including text localisation, text segmentation, character classification, and word recognition. In order to comprehensively evaluate the different facets of a text extraction system, a modular framework capable of evaluating a range of smaller research tasks is needed. The functioning of such a framework requires a significant amount of ground truthing effort addressing multiple text representation levels — a tedious and error prone exercise.

This paper introduces the *CVC Annotation and Performance Evaluation Platform for Text Extraction (APEP-te)*: a set of on-line software tools that facilitate ground truthing and streamline performance evaluation over a range of text extraction research tasks. The platform supports distributed ground truthing allowing multiple users to work in parallel, while maintaining centralised management and quality control of the process. It supports annotation at different text representation levels, from pixels to text lines. The platform supports the definition of evaluation scenarios for text localisation, text segmentation and word recognition tasks, and brings together implementations of state of the art performance evaluation algorithms, on-line calculation of performance metrics and per-image visualisation of results.

The APEP-te platform has been used extensively for ground truth creation, while it provides the submission management, performance evaluation and results visualisation functionality of the ICDAR 2011 (Challenge 1) and ICDAR 2013 (Challenges 1 and 2) Robust Reading competitions. It is now made available for public use as a service.

This paper describes the different parts of the platform, discusses the design principles and details the performance

evaluation methodologies implemented. In section II we describe the current best practices in text extraction performance evaluation, overview the datasets available, and make a case for the necessity of this platform. Section III gives an overview of the platform’s components and details for public access. Section IV discusses the ground truth specification and describes the set of functionalities related to image annotation and the management of the ground truthing process. In section V we describe the performance evaluation functionality offered, including the definition of evaluation scenarios and visualisation of results, while we provide a description of the evaluation algorithms implemented. Section VI concludes the paper.

II. BACKGROUND

Before introducing the new framework, it is of interest to examine the state of the art in terms of available datasets and ground truth data as well as current best practice when it comes to the evaluation of text extraction systems. Text extraction systems described in the literature refer to diverse pipelines that can have slightly different final objectives and target various application domains. More often than not, the target is text localisation in static real-scene images, and the preferred performance evaluation strategy is based on comparison of isothetic, axis-aligned bounding boxes at the word level.

The ICDAR robust reading competition dataset [1], [2] is the de-facto evaluation dataset in the community. A series of ICDAR competitions structured around this dataset have propelled its adoption as a community standard and created a long trend of consistent evaluation. The ICDAR dataset has served the community very well over the past decade — caution is nevertheless advised as there are a number of issues with its use including the inconsistent definition of performance metrics by various authors¹ and the existence of duplicate images in the dataset.²

Although text localisation is the most frequent task addressed, there is a variety of approaches in the literature that tackle different research tasks such as text segmentation [3], [4], character [5] or word recognition [6], word spotting [7] or combined localisation and recognition [8], [9]. In many cases, specialised datasets, custom ground truth and ad-hoc performance evaluation methodologies are used.

¹Precision, recall and f-score are variably calculated as averages over per-image results, or as overall values over text objects in the collection by different authors.

²A number of such duplicate images have been detected and removed in the ICDAR 2013 Robust Reading competition

As a result, numerous new datasets and associated ground truth have been published recently, addressing different community needs. Datasets oriented towards text localisation define ground truth at different granularities: the ICDAR Robust Reading [1] and the CHAR-74K [5] datasets define bounding boxes at character and word levels, the Microsoft Text DB [3] and the MSRA-TD500 [10] at text line level, while the NEOCR dataset [11] at text field level. Transcription information is usually provided at the same granularity. Most of the above datasets define isothetic axis-aligned boxes and are not suitable for evaluating localisation of non-horizontal text. Notable exceptions are the NEOCR dataset and the MSRA-TD500 which account for perspective transformations.

The strategy for the collection of images is also varied substantially. Datasets like ICDAR, MSRA-TD500 or KAIST [12] originate from targeted text shooting, typically resulting in high resolution images with centred text content. Others, like NEOCR, the Microsoft Text DB or Street View Text DB [7], comprise images obtained in less uncontrolled conditions. A limited number of datasets, e.g. KAIST and MSRA-TD500, include text images in multiple scripts or languages.

Pixel level segmentation information is rarely available in the ground truth. Examples include the KAIST dataset and datasets produced using the platform presented here (i.e. the ICDAR 2011 [13] and 2013 [14] born-digital images dataset).

Overall, there is a lack of standardisation in terms of annotation information. We consider that one of the key reasons for this is the lack of a unified framework for ground truthing and performance evaluation. Such a framework should be flexible enough to address a variety of text extraction facets. At the same time it should permit the definition of ground truth information at multiple text representation levels and capture the full hierarchy from pixels to text lines if so required. The platform presented in this paper implements such a framework and helps to address this gap.

III. OVERVIEW OF THE PLATFORM

The APEP-te platform is an online collection of tools and processes, integrated as a public Web service. The platform is based on a synchronised database/filesystem where datasets are physically created as separate folders in a network-accessible resource. The database is responsible for maintaining metadata for each image in the dataset while a synchronisation process ensures consistency with the physical storage. All public software tools are implemented as HTML5 interfaces, while specialised processing (e.g. the calculation of performance evaluation metrics) takes place on the server side. Key features of the platform include:

- Comprehensive ground truthing tools
- Centralised management of the ground truthing process
- Quality control and versioning
- Streamlined definition of evaluation scenarios
- On-line calculation of performance evaluation metrics
- In-depth results visualisation

In the next two sections the main aspects of the platform are described, grouped into two functionality blocks: tools and processes for the creation of ground truth information, and for the evaluation of different text extraction aspects.

The on-line framework is available for public use. An installation pack that allows local deployment of the Web portal is available through the Web site of the authors.³

IV. IMAGE ANNOTATION FUNCTIONALITY

The ground truthing strategy is based on a scheme previously introduced by the authors [15]. The platform provides updated specifications and tools implementations that adhere to this scheme.

A. Basic Concepts

The evaluation of different aspects of a text extraction system requires annotation at several text representation levels. The ground truthing tool allows for annotation at the pixel level (i.e. areas and skeletons of individual text parts) and at a sequence of higher semantic levels from text parts to atoms, words and text lines. Levels from atoms upwards can be assigned transcriptions.

The ground truthing process has been designed to be flexible based on the understanding that not all levels of ground truth are always necessary. Obviously, the more ground truth information available, the more diverse a set of evaluation scenarios can be ultimately defined. The ground truth XML specification is defined to reflect this flexibility.

The lowest representation level captures pixel-level information about basic text structures. *Text Parts* can be thought of as connected components and are the most primitive structures supported by the framework. Text Parts group together pixels that should be segmented as a single region by a perfect segmentation process (i.e. pixels that were created in the same text production step). Text Parts might correspond to single characters, parts of characters (e.g. the two parts of character ‘i’), or even multiple characters (e.g. in the case of cursive text all connected characters created by the same stroke would be part of a single Text Part). Text Parts are represented by their area, and optionally by their skeleton.

Atoms are defined as the minimum set of text parts that can be assigned a transcription. It is quite usual for Atoms to correspond to single Text Parts (especially for Latin script text), but they might comprise more than one Text Part (e.g. two-part characters, digital-7 fonts etc). Individual characters, when they have been produced individually, are implicitly defined as the subset of Atoms with a transcription of a single character. The concept of Atoms and the rationale for their use is explained in more detail in [15].

Groups of Atoms give rise to *Words*, which in turn can be grouped into *Text Lines*. Words and Text Lines are represented by their isothetic bounding box. If they are the result of a grouping of lower level entities (Words made from Atoms, Text Lines made from Words), their bounding boxes are automatically calculated. Alternatively, Word and Text Line bounding boxes can be defined explicitly by the user. Words that comprise Atoms and Words that are defined directly at the bounding box level can co-exist in the ground truth definition.

Atoms, Words and Text Lines can be qualified with the special tag “Don’t care”, which indicates to subsequent performance evaluation algorithms that they should not penalise

³<http://www.cvc.uab.es/apep>

methods that fail to detect them, and should ignore methods that have actually detected them. This is useful for example in the case of low-quality unreadable text which lies out of the scope of automatic detection.

B. Specification

The standard representation of the ground truth for a given image comprises a set of three files. The main part (required) is an XML file which encodes the whole hierarchy from Atoms upwards. Pixel level information (optional) is stored in two image files that are referenced in the XML part. The first image file is a colour-coded image where the text parts of each atom are represented in a different colour. The second image file is a bi-level image that encodes the skeletons for all text part, if they have been specified. The schema of the XML file, as well as an online XML verification tool can be found online.⁴ The specification is significantly updated compared to [15] to include new concepts and decouple pixel level information from the XML file.

C. Ground Truthing Tool

Figure 1 shows the ground truthing tool. Choosing to edit an image reserves the said image for a short period, so that many users can work in parallel without conflicts. Every editing round creates a new version of the image ground truth, while the user can revert to previous ground truth versions.

The edit screen presents the target image on the right and a tree structure representing the hierarchy of textual content on the left. The user can edit pixel level information in various ways, from individual pixel labelling to adjustable flood fill operations. To accelerate the marking of horizontal words / text lines that comprise single text-part characters (the most common case), the user can define the text parts of the entire word / text line in a single step and the system will intelligently create the structure of text parts, atoms words and text lines by analysing their relative location and parsing the transcription given. Optionally, the user can define the skeleton of each text part; in this case, the software automatically calculates a first skeleton approximation that the user can then edit.

As explained before, the user can opt to define directly words or text lines at the bounding box level, and skip the pixel level all together. Finally, existing ground truth information can be imported to the platform in various ways, making it relatively easy to convert existing datasets to the format of the APEP-te platform.

D. Managing the Ground Truthing Effort

Figure 2 shows a screenshot of the ground truthing management tool. The platform presents a searchable list to the ground truth manager that allows one to keep track of the overall progress, respond to specific comments that ground truthers make and assign a quality rating to each image. Using the same tool, the user can assign images to the training and test subsets that are subsequently used for defining evaluation scenarios.

V. PERFORMANCE EVALUATION FUNCTIONALITY

Performance evaluation is structured around the definition of evaluation scenarios. For each evaluation scenario a simplified ground truth (containing the necessary subset of the full information relevant to the task at hand) is created according to the defined training and test sets. Results over the test set, in the same simplified format can then be submitted to the platform, and performance evaluation metrics are automatically calculated. Results are reported in terms of comparative tables, while in depth per-image visualisation is also supported.

A. Research Tasks and Performance Evaluation Algorithms

The tasks of text localisation, text segmentation and text recognition can be evaluated using the APEP-te platform. These tasks reflect the principal needs of the international community, and are the ones targeted by the past two editions of the ICDAR Robust Reading competition. Note that the ground truth constructed permits the evaluation of many more text extraction research tasks. Support for word spotting and end-to-end (combined localisation and recognition) evaluation scenarios is planned for the near future.

Text Localisation evaluation scenarios can be defined at the word level or the text line level. The performance evaluation method implemented is the one proposed by Wolf and Jolion [16]. The said method uses a combined area overlap and object count based precision and recall metric, which can be adapted to separately penalise one-to-many and many-to-one relationships. This is important in the case of mismatches between the localisation granularity of the ground truth and the tested method. The interested reader should study [16] for more details.

For the evaluation of *Text Segmentation* methods two performance metrics are implemented in the platform. The first is an overall pixel classification mismatch, which is widely used in the community. The above metric only allows for global evaluation and is strongly biased by the size of text parts (big characters count more than small ones). A second metric, designed to evaluate segmentation quality in terms of its optimality for a later recognition step, is defined at the level of atoms. This metric checks the degree to which a segmentation method produced corresponding segments that preserve the overall shape the atoms and their text parts. The interested reader is referred to [15] for more details.

The following changes are introduced compared to [15]. First, segmentation results can be submitted at the image level (a bi-level image of text vs non-text pixels) or the atom level (a colour coded image with atoms corresponding to different colours). Depending on the submitted result, the framework uses either user submitted atom information (colour codes), or it automatically calculates the best atom matches. Furthermore, an additional option has been introduced for the calculation of the minimal coverage criterion (see [15] for an explanation). Apart from using the skeleton, as per the original publication, an area-based alternative definition is introduced as per [14].

For *Text Recognition* the implemented performance evaluation method addresses the word level (character and text line levels will be supported soon). The platform automatically creates training and test sets based on cropped word images

⁴<http://dag.cvc.uab.es/tools/?com=gt>



Fig. 1. A screenshot of the ground truthing tool, the hierarchy of textual content and the defined text parts (areas and skeletons) are visible over the image.




Image	width	height	Subset	Quality	gt	comments	creator
 img_106049628.jpg	2304	3072	use test	regular	2	Check unreadable text	DemoUser
 img_111981198.jpg	1024	768	use training	regular	2	Check product names	DemoUser
 img_11267413.jpg	2592	1944	use training	good	3		dimos

Fig. 2. A reduced screenshot of the ground truth management tool.

with associated descriptions, and requires a single transcription per word image as result. The platform allows the inclusion of surrounding context in the cropped word images by relaxing the tight bounding boxes by an amount of pixels defined by the user. Statistics on correctly recognised words and on accumulated edit distance are reported.

B. Definition of Evaluation Scenarios

The user can create evaluation scenarios for selected tasks through an on-line interface. For each of the tasks addressed in the evaluation scenario, the user can adjust the parameters of the corresponding performance evaluation algorithm. The platform makes use of the latest ground truth version available for each image, and of the assignment of images into training and test sets provided to construct an evaluation portal through which results over the test set can be submitted.

C. Evaluation Results and Visualisation

Once evaluation scenarios have been defined, results can be uploaded over the test set for each task offered in predefined, simple formats. To obtain a better idea of the process, the reader can examine the structure of the research tasks of the ICDAR 2013 Robust Reading competition⁵ (Challenges 1 and 2) — note that the competition is making use of this platform.

⁵<http://dag.cvc.uab.es/icdar2013competition>

The framework automatically analyses submitted results and presents performance metrics for the submitted methods in comparative tables and graphs. An example is shown in Figure 3. Furthermore, tables with per-image results are produced, while a range of different per-image visualisation options are offered. An example for text localisation results is shown in Figure 4, while similar visualisation tools are available for text segmentation and word recognition.

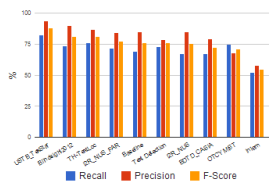
In many cases it is difficult to specify a single realistic evaluation scenario by manually fixing the corresponding parameters for each research task. Hence, apart from specifying a fixed evaluation scenario, a full validation can be run over a range of values for the evaluation parameters. Figure 5 shows such a validation over a range of Area Recall and Area Precision values in a text localisation task. Similar surfaces can be created for text segmentation tasks. The volume under surface can then be used as an alternative performance metric.

VI. CONCLUSION

This paper presented a set of on-line software tools that permit the creation of comprehensive ground truth and the performance evaluation of different text extraction tasks. The platform has been tested in real-life conditions in the context of the ICDAR 2011 and ICDAR 2013 Robust Reading competitions, and is now made available for public use.

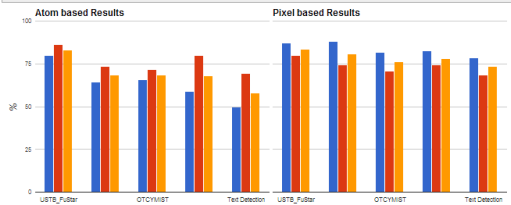
Ranking for Task 1 - Text Localization

Method	Recall	Precision	Fmean
USTB_TextStar	82.38%	93.83%	87.74%
Blindsght2012	73.81%	90.11%	81.15%
TH-TextLoc	75.85%	86.82%	80.96%
IQR_NUS_FAR	71.42%	84.17%	77.27%
Baseline	69.21%	84.94%	76.27%
Text Detection	78.18%	78.62%	78.41%
IQR_NUS	67.52%	85.19%	75.34%
BOTO_CASIA	67.05%	78.98%	72.53%
OTCVIMST	74.85%	67.69%	71.09%
Inkam	52.21%	58.12%	55.00%



Ranking for Task 2 - Text Segmentation

Method	Pixel Results			Atom based Results									
	Recall	Precision	F-Score	Well s.	Merged	Broken	Br-Mer.	Lost	False p.	Detected	Recall	Precision	Fscore
USTB_TextStar	87.21%	79.58%	83.44%	6258	920	56	1	587	370	7260	80.01%	86.20%	82.99%
IQR_NUS	87.95%	74.40%	80.61%	5051	1584	30	6	1151	685	6878	64.57%	73.44%	68.72%
OTCVIMST	81.82%	71.00%	76.03%	5143	1420	34	2	1223	1083	7178	65.75%	71.65%	68.57%
IQR_NUS_FAR	82.56%	74.31%	78.22%	4619	1474	12	1	1716	156	5771	59.05%	80.04%	67.96%
Text Detection	78.68%	68.63%	73.32%	3883	2716	36	0	1187	210	5590	49.64%	69.46%	57.90%



Ranking for Task 3 - Word Recognition

Method	Total Edit distance	Correctly Recognised Words	T.E.D. (upper)	C.R.W. (upper)
PhotoOCR	105.5	82.21%	88.8	85.41%
MAPS	196.2	80.4%	186.4	81.51%
PLT	200.4	80.26%	190.9	81.38%
NESP	214.5	79.29%	198.2	80.75%
Baseline	409.4	60.95%	400.1	61.57%

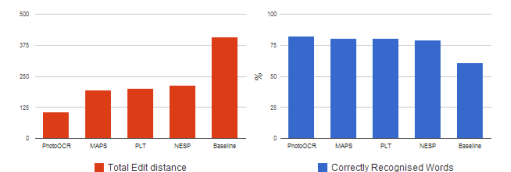


Fig. 3. Comparative tables and graphs for a text extraction task of the ICDAR 2013 Robust Reading competition, run on the APEP-te platform.



Fig. 4. Visualisation of localisation results for an image of the ICDAR 2013 Robust Reading competition. Different colours indicate correctly matched (green), many-to-one (blue), one-to-many (orange), missed and false positive (red) detections. Further visualisation options are available.

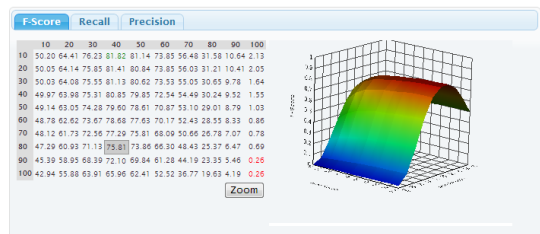


Fig. 5. Visualisation of full validation results for a text localisation task.

ACKNOWLEDGMENT

This work has been supported by the research project "Text and the City" (TIN2011-24631) and the fellowship RYC-2009-

REFERENCES

- [1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, vol. 2, 2003, pp. 682–687.
- [2] S. M. Lucas, "ICDAR 2005 text locating competition results," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 80–84.
- [3] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.
- [4] L. Gómez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proceedings of the 12th Int. Conf. on Document Analysis and Recognition (ICDAR 2013)*, vol. 1, 2013.
- [5] K. Sheshadri and S. Divvala, "Exemplar driven character recognition in the wild," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, pp. 13.1–13.10.
- [6] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 10, pp. 1733–1746, 2009.
- [7] K. Wang and S. Belongie, "Word spotting in the wild," in *Computer Vision—ECCV 2010*. Springer, 2010, pp. 591–604.
- [8] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3538–3545.
- [9] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1457–1464.
- [10] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1083–1090.
- [11] R. Nagy, A. Dicker, and K. Meyer-Wegener, "Neocr: A configurable dataset for natural image text recognition," in *Camera-Based Document Analysis and Recognition*. Springer, 2012, pp. 150–163.
- [12] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, "Scene text extraction with edge constraint and text collinearity," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3983–3986.
- [13] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, "Icdar 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email)," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1485–1490.
- [14] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "Icdar 2013 robust reading competition."
- [15] A. Clavelli, D. Karatzas, and J. Lladós, "A framework for the assessment of text extraction algorithms on complex colour images," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM, 2010, pp. 19–26.
- [16] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal of Document Analysis and Recognition (IJ DAR)*, vol. 8, no. 4, pp. 280–296, 2006.