2013 12th International Conference on Document Analysis and Recognition

ICDAR 2013 Robust Reading Competition

Dimosthenis Karatzas*, Faisal Shafait†, Seiichi Uchida‡, Masakazu Iwamura§, Lluis Gomez i Bigorda*,
Sergi Robles Mestre*, Joan Mas*, David Fernandez Mota*, Jon Almazàn Almazàn* and Lluis Pere de las Heras*
*Computer Vision Centre, Universitat Autònoma de Barcelona; *dimos@cvc.uab.es*
†The University of Western Australia, Australia; *faisal.shafait@uwa.edu.au*
‡Kyushu University, Japan; *uchida@ait.kyushu-u.ac.jp*
§Osaka Prefecture University, Japan; *masa@cs.osakafu-u.ac.jp*

***Abstract*—This report presents the final results of the ICDAR 2013 Robust Reading Competition. The competition is structured in three Challenges addressing text extraction in different application domains, namely born-digital images, real scene images and real-scene videos. The Challenges are organised around specific tasks covering text localisation, text segmentation and word recognition. The competition took place in the first quarter of 2013, and received a total of 42 submissions over the different tasks offered. This report describes the datasets and ground truth specification, details the performance evaluation protocols used and presents the final results along with a brief summary of the participating methods.**

I. Introduction

Text extraction from versatile text containers like born-digital images, real scenes and videos has been a continuous interest in the field for more than a decade. The series of Robust Reading Competitions addresses the need to quantify and track progress in this domain. The competition was initiated in 2003 by Simon Lucas, focusing initially on text localisation and text recognition in real scene images [1] [2]. The first edition was met with great success and was repeated in 2005 [3], creating a reference framework for the evaluation of text detection methods. In 2011, two challenges were organised under the ICDAR Robust Reading Competition, one dealing with text extraction from born-digital images [4], and the other from real scene images [5]. The 2013 edition of the Robust Reading Competition [6] marks a new milestone in the series. The two challenges on real scenes and born-digital images have been integrated further, unifying performance evaluation metrics, ground truth specification and the list of offered tasks. In parallel, a new challenge is established on text extraction from video sequences, introducing new datasets, tools and evaluation frameworks.

The 2013 Robust Reading Competition [6] brings to the Document Analysis community:

- An enhanced dataset on born-digital images
- New ground truth at the pixel level for real-scene images
- Improved and intuitive performance evaluation protocols
- A new dataset of video sequences obtained under various activities and with diverse capturing equipment
- Video based ground truth for text detection in video
- A single point of entry for all submissions to all challenges
- A comprehensive Web site allowing the continuous submission of new results (over and above the deadlines of the ICDAR 2013 competition), on-line performance evaluation tools and enhanced visualisation of results.

The competition consists of three Challenges: Reading Text in Born-digital Images (Challenge 1), Reading Text in Scene Images (Challenge 2), and Reading Text in Videos (Challenge 3). Each Challenge is based on a series of specific tasks. Challenges 1 and 2 offer tasks on Text Localisation, where the objective is to detect the existence of text and return a bounding box location of it in the image; Text Segmentation, where the objective is to obtain a pixel-level separation of text versus background and Word Recognition, where the objective is to automatically provide a transcription for a list of pre-localised word images. Challenge 3 at this time offers a single task on Text Localisation where the objective is to detect and track text objects in a video sequence.

This report presents the final results after analysing the submissions received. The presentation is following the structure of the competition and results are grouped by Challenge. For each Challenge the datasets used and ground truth specification are briefly described, the performance evaluation protocol is detailed and finally the results are presented and analysed. First, the competition protocol is briefly described in Section II. Section III is dedicated to Challenge 1, Section IV to Challenge 2 and Section V to Challenge 3. Overall conclusions are presented in Section VI. Finally, in the appendix the technical details of all participating methods are summarised.

II. Competition Protocol

The competition was run in open mode, meaning that results over the test set were requested by the authors, and not executables of their systems. At all times we relied on the scientific integrity of the authors to follow the rules of the competition. The authors were free to participate in as many Challenges and Tasks as they wished. They were allowed to make multiple submissions to the same task as well. In the case of submission of different methods from the same authors, each method is separately described and ranked in the final results presented here. In the case where the results of different variants of the same base method were submitted, only the best performing variant is shown in the ranking tables of this report. A full ranking including all different variants submitted will be available in the competition Web after ICDAR 2013.

In total, 52 submissions were made (42 after excluding variants of the same method) to the different Challenges and Tasks of the competition from 15 individual participants.

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.221

III. Challenge 1: Reading Text in Born-Digital Images (Web and Email)

Challenge 1 focuses on the extraction of textual content from born-digital images, such as the ones used in Web pages and email messages. Embedding text in images, rather than encoding it explicitly in the electronic document, is a frequent choice of content authors. Text embedded in images is practically invisible to any automatic processes, hindering applications such as indexing and retrieval, content filtering etc.

Challenge 1 of this competition aims to quantify the state of the art in text extraction from born-digital images, and to highlight progress since the last edition of the competition [4]. Notably, since the past edition of this challenge in 2011, a new body of work has been published on the topic [7], [8], [9], [10]. Comparing to the last edition of the challenge, this edition introduces an updated dataset as well as small changes in the evaluation protocol that make it more comprehensive and intuitive.

In the rest of this section we present the results over three independent tasks related to text extraction from born-digital images: text localization, text segmentation and word recognition. Section III-A describes the dataset and ground truth used for the competition, while Section III-B details the performance evaluation methodology followed. In Section III-C we present the results and make some key observations.

A. Dataset

The dataset used for this competition comprises images extracted from Web pages and email messages. We selected a representative sample of Web pages of different categories (news, personal, commercial, social, government, etc) and emails of different type (spam, newsletters, etc) in proportions that reflect their real world usage.

Overall, we analysed 315 Web pages, 22 spam and 75 ham emails, and extracted all the images that contained text. We selected a subset of 561 images with a minimum size of 100x100 pixels. The collection was split into a training set of 420 images (same as the 2011 edition) and a test set of 141 images (including new images compared to the 2011 edition). For the word recognition task only words with a length of 3 characters or more were considered. The dataset contains 5003 such words, out of which 3564 comprise the training set and 1439 the test set.

Ground truth information was created in an hierarchical way spanning all levels from pixel level labelling to text parts, atoms, words and text lines, as described in [11]. In cases where pixel level labelling is not possible (e.g. extreme anti-aliasing or very low resolution) word and text line bounding boxes are directly defined. All bounding boxes are defined as axis-aligned isothetic rectangles, a reasonable design choice as in born-digital images most text is horizontal.

B. Performance Evaluation

We use standard metrics for the evaluation of the different tasks. These are largely shared between Challenges 1 and 2, facilitating comparison between the two. A brief explanation of the evaluation protocols is given in this section.

1) Task 1 - Text Localization: For the evaluation of text localisation results we make use of the framework proposed by Wolf and Jolion [12]. The key principle of the scheme is that evaluation is done at the object (word bounding box) level over the whole collection, taking into account the quality of each match between detected and ground truth text boxes. Matches are first determined based on area overlapping, given certain minimum quality thresholds. Then different weights for one-to-one, one-to-many and many-to-one matches are used when pooling together the results.

Two thresholds on the area precision (t_p) and area recall (t_r)control the way matches between ground truth rectangles are determined. For this challenge, we use the default values suggested in [12] for these thresholds, namely $t_r = 0.8$ and $t_p = 0.4$. For calculating the overall object *Precision* and *Recall* the method considers all matches over the whole collection. One-to-many and many-to-one matches can be assigned different weights allowing the uneven penalisation of different behaviours of the method under question. In this implementation, we give a lower weight to one-to-many matches than the rest. The rationale behind this decision is that we make use of the word level of our ground truth for evaluation; therefore, although we want to penalise methods that produce multiple rectangles for a single ground truth word, we do not want to penalise methods designed to produce results at the text-line level, detecting many ground truth words of the same line with a single rectangle. Hence, for many-to-one matches we do not inflict any penalty, while for one-to-many matches we use the suggested fixed weight of 0.8. We encourage the interested reader to review [12] for further details.

2) Task 2 - Text Segmentation: For the evaluation of text segmentation results we make use of the framework proposed by Clavelli and Karatzas [11]. The units for the evaluation are the atoms defined in the ground truth. An atom is the smallest combination of text parts (connected components) that can be assigned a transcription; therefore an atom may comprise one or multiple text parts. See [11] for an extensive discussion on atoms and their definition. The quality of a segmentation method is measured by the degree to which it is able to produce regions that preserve the morphological properties of the ground-truth atoms, as opposed to simply counting the number of mislabelled pixels. In simple terms, the framework is more permissive to segmentation methods that produce text parts that are slightly more dilated or more eroded versions of the ground truth ones, while it penalises methods that produce text parts that have distinctly different shapes from the ground truth ones, even if the same number of mislabelled pixels is involved in both cases.

In order to establish a match between ground truth and detected text parts, two conditions are to be satisfied. The *Minimal Coverage* condition is satisfied if the detected text part covers a minimum set of pixels of the ground truth text part. In the implementation used here this is defined as a percentage of the area of the text part and controlled by a parameter T_{min}. The maximal coverage criterion is satisfied if no pixel of the detected text part lies outside a maximal area, defined as a dilated version of the ground truth text part. In the current implementation, a parameter T_{max} controls the amount of dilation considered as a function of the stroke width of the component. For this evaluation, T_{min} was set to 0.5 and T_{max} to 0.9. Each of the atoms in the ground truth is classified as *Well Segmented*, *Merged*, *Broken*, *Broken and Merged* or *Lost*, while *False Positive* responses are also counted. Atom *Precision*, *Recall* and *F-score* metrics are calculated over the whole collection summarising the performance of the method. Word boxes that are directly defined at the word level in the ground truth, and do not contain atoms or text parts, are treated as "don't care" regions and discounted from the evaluation. The interested reader is encouraged to read [11] for more details. Apart from atom level metrics, we also report *Precision*, *Recall* and *F-score* results at the pixel level for completeness.

3) Task 3 - Word Recognition: For the evaluation of Word Recognition results we implemented a standard edit distance metric, with equal costs for additions, deletions and substitutions. For each word we calculate the normalized edit distance between the ground truth and the submitted transcription and we report the sum of normalised distances over all words of the test set. The normalisation is done by the length of the ground truth transcriptions. The comparison we performed is case sensitive. To assist qualitative analysis, we also provide statistics on the number of correctly recognised words.

C. Results and Discussion

Overall, 17 methods from 8 different participants (excluding variants of the same method) were submitted in the various tasks of this challenge. In the sections below we provide the final ranking tables as well as a short analysis of the results for each task. The participating methods are referred to by name in the ranking tables and in the text, please see Appendix VI for technical details on the participating methods.

Similarly to the past edition of the competition, we have included the performance of an out-of-the-box commercial OCR software package as a baseline for text localisation (task 1) and word recognition (task 3). For this purpose we have used the ABBYY OCR SDK (version 10) [13]. Factory default parameters were used for pre-processing with the following exceptions. First, we enabled the option to look for text in low resolution images, since the resolution of born-digital images is typically below $100DPI$. This is not exclusive, meaning that text in high resolution images is also looked for with this parameter enabled. Second, we set the OCR parameters *FlexiForms* and *FullTextIndex* to true, to force detection of text inside images. For text localization (task 1) we have used the reported location of individual words, while for word recognition (task 3) the returned transcription.

Please note that a direct comparison of the results reported here to the 2011 edition of the challenge (and the additional submissions received online between 2011 and 2013 [14]) is not straightforward as the test dataset has been updated to include more images, while small changes have been introduced in the evaluation protocol.

1) Task 1 - Text Localization: The final results for the Text Localisation task are shown in Table I. The ranking metric used for the Text Localisation task is the *F-score* (column in grey) calculated according to the methodology described in section III-B1. All metrics are calculated cumulatively over the whole test set (all detections over all 141 images pooled together).

TABLE I. RANKING OF SUBMITTED METHODS TO TASK 1.1

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	**82.38**	**93.83**	**87.74**
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

TABLE III. ANALYSIS OF TASK 1.2 RESULTS.

Method	Well Segmented	Merged	Broken	Lost	False Positives
USTB_FuStar	**6258**	920	56	**587**	622
I2R_NUS	5051	1584	30	1151	758
OTCYMIST	5143	1420	34	1223	1170
I2R_NUS_FAR	4619	1474	12	1716	**185**
Text Detection	3883	2716	36	1187	358

As it can be easily observed, most participating methods rank below the baseline method, indicating that existing commercial solutions are doing reasonably well with these images. Notably, at least 3 methods rank better than the baseline one, with 2 of them (USTB_TexStar and TH-TexLoc) doing significantly better than the commercial solution.

Using the performance of the baseline method as a yardstick (largely the same between the two competitions), an indirect comparison to the results of 2011 is possible. During the past edition [4] no method performed substantially better than the baseline method, with the highest ranking methods (including the baseline) yielding results within 1% of each other. Since the past edition many new methods were submitted online in the Challenge Web site [14], indicating an improvement of up to 10% over the 2011 results had been achieved. This is confirmed with this edition, with the top ranking method performing 11.5% better than the baseline method.

2) Task 2 - Text Segmentation: The final results for the Text Segmentation task are shown in Table II. The ranking metric used for the Text Segmentation task is the atom *F-Score* (column in grey) calculated as explained in section III-B2. All metrics are calculated cumulatively over the whole test set. For completeness, apart from the atom based metrics, table II also shows the pixel *Precision*, *Recall* and *F-score*.

Table III summarises the classification of the atoms as well as the number of false positives produced by each of the methods. It can be easily observed that the USTB_FuStar produces significantly more *Well Segmented* atoms than the rest of the methods. At the same time it produces less *Merged* atoms and quite a lot more *Broken* ones. Compared to this performance, all the rest of the methods seem to have a tendency to over-merge.

In terms of *Lost* atoms (atoms that were not detected at all, or were only partially detected), again USTB_FuStar seems to be performing the best. In most of the cases, the participating methods fail to detect the dots of the letters "i" and "j", hence missing most of the atoms corresponding to such characters. Similarly, punctuation points seem to create problems to most

TABLE II. Ranking of submitted methods to Task 1.2

Method	Pixel Level			Atom Level		
	Recall (%)	Precision (%)	F-score	Recall	Precision	F-score
USTB_FuStar	87.21	**78.84**	**82.81**	**80.01**	**83.31**	**81.62**
I2R_NUS	**87.95**	73.88	80.31	64.57	72.67	68.38
OTCYMIST [7]	81.82	70.42	75.69	65.75	70.79	68.18
I2R_NUS_FAR	82.56	73.67	77.86	59.05	79.64	67.82
Text Detection [15], [16]	78.68	67.97	72.93	49.64	67.67	57.27

TABLE IV. Ranking of submitted methods to Task 1.3

Method	Total Edit Distance	Correctly Recognised Words (%)
PhotoOCR	**105.5**	**82.21**
MAPS [17]	196.2	80.4
PLT [18]	200.4	80.26
NESP [19]	214.5	79.29
Baseline	409.4	60.95

of the methods. Finally, in terms of *False Positives*, the method I2R_NUS_FAR is the best performing one, as is also indicated by the high precision it yields in table II.

3) Task 3 - Word Recognition: The results for the Word Recognition task are shown in Table IV. Two metrics are shown, namely the total edit distance, which is also used for the final ranking (column in grey) and the percentage of correctly recognised words as explained in section III-B3. Overall, we can note a significant advancement of the state of the art since the 2011 edition. In the previous edition of the competition there was a single participating method with performance very close to the baseline, recognising correctly a 61.54% of the words. The best performing methods in this edition show an important improvement and an increment of 20% in terms of correctly recognised words.

IV. Challenge 2: Reading Text in Scene Images

In this edition, a new task of Text Segmentation (Task 2.2) is introduced in Challenge 2. For this task, a pixel-level ground-truth has been created for scene images in the dataset. The participating methods are asked to provide a pixel-level segmentation of the given real scene images such that all pixels belonging to text are marked as fore-ground and other pixels as background.

A. Dataset

The scene image dataset used for this competition is almost the same as the dataset at ICDAR2011 Robust Reading Competition. The difference from the ICDAR2001 dataset is revision of ground-truth texts at several images. In addition, a small number of images duplicated over training and test sets were excluded. Accordingly, ICDAR2013 dataset is a subset of ICDAR2011 dataset. The number of images of ICDAR2013 dataset is 462, which is comprised of 229 images for the training set and 233 images for the test set.

A new feature of ICDAR2013 dataset is pixel-level ground-truth. Human operators carefully observed each scene image and painted character pixels manually. (Precisely speaking, they used a semi-automatic segmentation software and then made further manual revision.) Every character is painted by a unique color. Consequently, the pixel-level ground-truth can be used for ground-truth of individuial character segmentation as well as ground-truth of character detection. The operators have made a double-check for more reliable ground-truth.

One differece from Challenge 1 on attaching pixel-level ground-truth is that a single connected component showing multiple characters (e.g., the connected component showing "oca" in the Coca-Cola logo) is segmented into individual characters and then painted mutiple colors. (Thus, the "oca" will be colored by three different colors.) This segmentation is done rather intuitively.

The other difference is that "don't care" regions are introduced in the ground-truth. Scene images often contain illegible tiny text regions. In ICDAR2011 dataset, each legible text region was marked by a bounding box, whereas illegible text regions were left unmarked. We have introduced the concept of "don't care" regions in ICDAR2013 dataset, where tiny text regions are marked with a special colored rectangle. To keep consistensy with ICDAR2011 competition as much as possible, only those text regions are marked as don't care that were not marked as text in the previous dataset.

B. Performance Evaluation

The performance evaluation schemes for all tasks in Challenge 2 follow the same protocols as those in Challenge 1. The only difference is in handling of "don't care" regions. Don't care regions are treated during a pre-processing phase. Before the core evaluation starts overlapping with don't care regions is checked and any detected regions that fall inside don't cares are removed from the results list. This makes sure that no false alarms are reported if an algorithm marks a don't care region as text. On the other hand, another algorithm that does not consider a don't care region as text also does not get penalized.

C. Results and Discussion

Challenge 2 of the competition received 22 entries from 13 different participants (excluding variants of the same method). In this edition of the competition, we have included the performance of an out-of-the-box commercial OCR software package as a baseline to compare against for text localisation (task 1) and word recognition (task 3) inline with Challenge 1 of the competition. As in Challenge 1, ABBYY OCR SDK (version 10) [13] was used as the baseline. Although the same dataset was used in this edition of the competition as that in 2011 [5], some corrections were made in the ground-truth as well as a few duplicated images were removed. Hence, a direct comparison with the results of previous competition is not straightforward. However, since the corrections were minor, we

TABLE V. RANKING OF SUBMITTED METHODS TO TASK 2.1

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	66.45	**88.47**	**75.89**
Text Spotter [20], [21], [22]	64.84	87.51	74.49
CASIA_NLPR [23], [24]	68.24	78.89	73.18
Text_Detector_CASIA [25], [26]	62.85	84.70	72.16
I2R_NUS_FAR	**69.00**	75.08	71.91
I2R_NUS	66.17	72.54	69.21
TH-TextLoc	65.19	69.96	67.49
Text Detection [15], [16]	53.42	74.15	62.10
Baseline	34.74	60.76	44.21
Inkam	35.27	31.20	33.11

can still safely comment on the overall trends in performance between the two latest editions of the competition.

1) Task 1 - Text Localization: The final results for the Text Localization task are shown in Table V. The ranking metric used for the Text Localisation task is the *F-score* (column in grey) calculated according to the methodology described in Section III-B1. In contrast to Challenge 1, most of the participating methods rank above the baseline method. One possible reason for this effect might be that the commercial OCR software does not expect, and hence is not trained for the distortions introduced by camera-captured images (background clutter, perspective distortions, . . .).

One can see from Table V that there are a number of methods achieving an F-score of more than 70%, whereas only one method was able to achieve this level in the last contest (refer to Table I in [5]). The two method that participated both times, TH-TextLoc and Text Spotter (named as Neumann's method previously) have shown significant imporvements to their previous results. This indicates a promising overall advancement in the state-of-the-art. The winning method USTB_TexStar achieved an F-score of 75.9%, implying that the competition dataset still poses significant challenges to state-of-the-art methods today.

2) Task 2 - Text Segmentation: This was a newly introduced task for Challenge 2 in this edition of the competition. The final results for this task are given in Table VI. The ranking metric used is the atom *F-Score* (column in grey) calculated in a similar fashion as that for Challenge 1 (see Section III-B2). Besides atom based metrics, Table VI also shows the pixel *Precision*, *Recall* and *F-score* values. The results show that the winning method I2R_NUS_FAR achieved the best performance both with respect to the atom-level as well as pixel-level metrics.

Further analysis of the results in Table VII shows that the winning method I2R_NUS_FAR not only obtained the highest number of well segmentated atoms, but also the least number of lost atoms. This explains the overall best performance of this method both in pixel level metrics and in atom level metrics.

3) Task 3 - Word Recognition: It is for the first time in the history of Robust Reading competitions that this challenge has received more than three entries, clearly indicating renewed interest on the topic. The results for the Word Recognition task based on the total edit distance (the ranking metric) and the percentage of correctly recognized words are shown in Table VIII. Like in Challenge 1, we can note a huge improvement in the performance of the state of the art since the 2011 edition. The winning method PhotoOCR by Google Inc. was able to correctly recognize over 82% of the words, which is double than the performance achieved by the winner in 2011 (41%). Hence, we can safely conclude that the state of research in scene text recognition has now advanced to the level where it can be used in practical applications – though room of improvement still remains.

TABLE VII. ANALYSIS OF TASK 2.2 RESULTS.

Method	Well Segmented	Merged	Broken	Lost	False Positives
I2R_NUS_FAR	**4027**	297	11	**1532**	368
NSTextractor	3719	106	10	2033	**338**
USTB_FuStar	3992	279	12	1585	941
I2R_NUS	3540	637	7	1684	353
NSTsegmentator	3989	148	25	1706	2819
Text Detection	3640	314	27	1884	1888
OTCYMIST	2451	276	23	3118	4573

TABLE VIII. RANKING OF SUBMITTED METHODS TO TASK 2.3

Method	Total Edit Distance	Correctly Recognised Words (%)
PhotoOCR	**122.7**	**82.83**
PicRead [27]	332.4	57.99
NESP [19]	360.1	64.20
PLT [18]	392.1	62.37
MAPS [17]	421.8	62.74
Feild's Method	422.1	47.95
PIONEER [28], [29]	479.8	53.70
Baseline	539.0	45.30
TextSpotter [20], [21], [22]	606.3	26.85

V. CHALLENGE 3: READING TEXT IN VIDEOS

Challenge 3 focuses on text localization in videos. The amount of videos available is rapidly increasing. This has been led on one hand by the pervasive use of camera phones and hand-held digital cameras which allow easy video capture, and on the other hand by video-sharing websites such as YouTube[1] and Nico Nico Douga[2]. In addition to the videos actively captured by people, a new kind of video feed, passively captured by wearable cameras (e.g. Memoto[3] and Google Glass[4]), are used to record and observe the user's daily life (e.g. [30]). For both actively and passively captured videos, analyzing occurences of text in videos is highly demanded.

The aim of this challenge is to provide the impetus for the development of methods that take advantage of video sequences to localize text in the depicted scene. The objective of the text localisation task is to obtain the locations of words in the video sequence in terms of their affine bounding boxes. The task requires that words are both localised correctly in every frame and tracked correctly over the video sequence.

Since a video is a collection of static images, one may think that this challenge is not essentially different from Challenge 2. Although an existing method for static images could in principle be used in this challenge, videos are of different nature compared to static images. Consecutive frames present in general a small differences, while images from videos are

[1] https://www.youtube.com/

[2] http://www.nicovideo.jp/

[3] http://memoto.com/

[4] http://www.google.com/glass/start/

TABLE VI. RANKING OF SUBMITTED METHODS TO TASK 2.2

Method	Pixel Level			Atom Level		
	Recall (%)	Precision (%)	F-score	Recall	Precision	F-score
I2R_NUS_FAR	**74.73**	**81.70**	**78.06**	**68.63**	80.36	**74.03**
NSTextractor	60.71	76.28	67.61	63.38	**83.70**	72.14
USTB_FuStar	69.58	74.45	71.93	68.03	72.79	70.33
I2R_NUS	73.57	79.04	76.21	60.33	76.69	67.53
NSTsegmentator	68.41	63.95	66.10	67.98	54.14	60.28
Text Detection [15], [16]	64.74	76.20	70.01	62.03	57.40	59.63
OTCYMIST [7]	46.11	58.53	51.58	41.77	31.49	35.91

typically worse than static images due to motion blur and out of focus issues, and video compression introduces further artifacts.

Applying text detection in a frame-by-frame fashion makes little sense for video sequences as it ignores any temporal cues. Combining detection with tracking would be a reasonable approach here for a few reasons. First, most text detection methods in real scenes are far from working in a real-time fashion, hence it is necessary to propagate hypotheses until a new observation is made. Another important reason for a tracking approach is the presence of occlusions, or equivalently the variable quality of the video sequence that might make detection in certain frames impossible. This also enables a sophisticated rejection, which can cope with excessive amounts of false positives (see for example the performance of the baseline method we employ in this competition in section V-C).

A. Dataset

The challenge is based on various short sequences (around 10 seconds to 1 minute long) selected so that they represent a wide range of real-life situations (high-level cognitive tasks), using different types of cameras. The dataset was collected by the organisers in different countries, in order to include text in various languages (Spanish, French and English[5]). The video sequences correspond to certain tasks that we asked users to perform, like searching for a shop in the street or finding their way inside a building. The tasks were selected so that they represent typical real-life applications, and cover indoors and outdoors scenarios. We used different cameras for different sequences, so that we also cover a variety of possible hardware used; these include mobile phones, hand-held cameras and head-mounted cameras.

We provided 28 videos in total; 13 videos comprised the training set and 15 the test set. The training set was divided into two parts (A and B). The only difference between them was the time of release. The videos were captured with 4 kinds of cameras ((A).Head-mounted camera, (B).Mobile Phone, (C).Hand-held camcorder and (D).HD camera) and categorized into 7 tasks ((1).Follow wayfinding panels walking outdoors, (2).Search for a shop in a shopping street, (3).Browse products in a super market, (4).Search for a location in a building, (5).Driving, (6).Highway watch, (7).Traing watch). See the summary of the videos in Table IX.

TABLE IX. SUMMARY OF VIDEOS IN THE DATASETS IN CHALLENGE 3.

		Video ID	Task	Camera type	No. of Frames	Duration
training set	A	7	(5)	(A)	264	00:11
		8	(5)	(A)	240	00:10
		37	(2)	(C)	456	00:19
		42	(2)	(C)	504	00:21
		51	(7)	(D)	450	00:15
		52	(7)	(D)	450	00:15
		54	(7)	(D)	480	00:16
	B	13	(4)	(A)	576	00:24
		19	(5)	(A)	408	00:17
		26	(5)	(B)	387	00:16
		36	(2)	(C)	336	00:14
		40	(2)	(C)	408	00:17
		41	(2)	(C)	528	00:22
test set		1	(1)	(B)	602	00:20
		5	(3)	(B)	542	00:18
		6	(3)	(B)	162	00:05
		11	(4)	(A)	312	00:13
		17	(3)	(A)	264	00:11
		20	(5)	(A)	912	00:38
		23	(5)	(B)	1020	00:34
		24	(5)	(B)	1050	00:35
		32	(2)	(C)	312	00:13
		35	(2)	(C)	336	00:14
		39	(2)	(C)	438	00:20
		44	(6)	(D)	1980	01:06
		48	(6)	(D)	510	00:17
		49	(6)	(D)	900	00:30
		53	(7)	(D)	450	00:15

B. Performance Evaluation

There are numerous evaluation frameworks proposed for multiple object tracking systems [31], [32], an elaborated review, including their use for text tracking in videos, can be found in Kasturi *et al.* [31]. In this competition, we selected to use the CLEAR-MOT [32] and VACE [31] metrics adapted to the specificities of text detection and tracking, extending the CLEAR-MOT code[6] provided by Bagdanov *et al.* [33].

The CLEAR-MOT [32] evaluation framework provides two overall performance measures: the Multiple Object Tracking Precision (MOTP), which expresses how well locations of words are estimated, and the Multiple Object Tracking Accuracy (MOTA), which shows how many mistakes the tracker system made in terms of false negatives, false positives, and ID mismatches. On the other hand, the Average Tracking Accuracy (ATA) VACE metric [31] provides a spatio-temporal measure that penalizes fragmentations while accounting for the number of words correctly detected and tracked, false negatives, and false positives.

Given a video sequence an ideal text tracking method

[5] Japanese data are also planned to be released in the near future.

[6] http://www.micc.unifi.it/masi/code/clear-mot/

should be able to detect all text words present at every frame and estimate their bounding boxes precisely; additionally it should also keep consistent track of each word over time, by assigning a unique ID which stays constant throughout the sequence (even after temporary occlusion, etc).

For every time frame t a text tracking system outputs a set of hypotheses $\{h_1^t, ..., h_n^t\}$ for a set of words in the ground-truth $\{w_1^t, ..., w_m^t\}$. Those frame level objects can be grouped by their unique identifiers into sequence level hypotheses $\{H_1, ..., H_p\}$ and ground-truth words $\{W_1, ..., W_q\}$ (that typically span more than one frame). For a distinctive notation we refer as H_i^t and W_i^t to the frame level objects at frame t in H_i and W_i respectively.

The evaluation procedure is based on a mapping list of word-hypothesis correspondences. At the frame level we have a mapping M_t for each frame t in the video sequence made up with the set of pairs (w_i^t, h_j^t) for which the sum of overlap(w_i^t, h_j^t) is maximized, where overlap$(\cdot)$ is a function overlap$(w_i^t, h_j^t) = \frac{a(w_i^t \cap h_j^t)}{a(w_i^t \cup h_j^t)}$ of the intersection area $(a(\cdot))$ of their bounding boxes. Additionally, a pair (w_i^t, h_j^t) is considered a valid correspondence iff overlap$(w_i^t, h_j^t) > 0.5$. At the sequence level we have a unique mapping M of word-hypothesis correspondences (W_i, H_j), but in this case maximizing the spatio-temporal overlap of all possible (W_i, H_j) combinations.

The two CLEAR-MOT metrics are calculated using the frame level mappings as:

$$MOTP = \frac{\sum_{i,t} o_t^i}{\sum_t c_t} \tag{1}$$

where o_t^i refers to the overlapping ratio of the ith correspondence in the mapping M_t and c_t is the number of correspondences in M_t; and:

$$MOTA = 1 - \frac{\sum_t (fn_t + fp_t + id\_sw_t)}{\sum_t g_t} \tag{2}$$

where fn_t, fp_t, id_sw_t, and g_t refer respectively to the number of false negatives, false positives, ID switches, and ground-truth words at frame t.

The Sequence Track Detection Accuracy (STDA) is calculated by means of the sequence level mapping M as:

$$STDA = \sum_{i=1}^{N_M} \frac{\sum_t m(W_i^t, H_i^t)}{N_{W_i \cup H_i \neq \emptyset}} \tag{3}$$

where N_M is the number of correspondences in M, $N_{W_i \cup H_i \neq \emptyset}$ is the number of frames where either W_i or H_i exist, and $m(W_i^t, H_i^t)$ takes a value of 1 iff overlap$(W_i^t, H_i^t) > 0.5$ or 0 otherwise.

The STDA is a measure of the tracking performance over all of the objects in the sequence and thus can take a maximum value of N_W, which is the number of ground-truth words in the sequence. The Average Tracking Accuracy (ATA), which is the normalized STDA per object, is defined as:

$$ATA = \frac{STDA}{\left[\frac{N_W + N_H}{2}\right]} \tag{4}$$

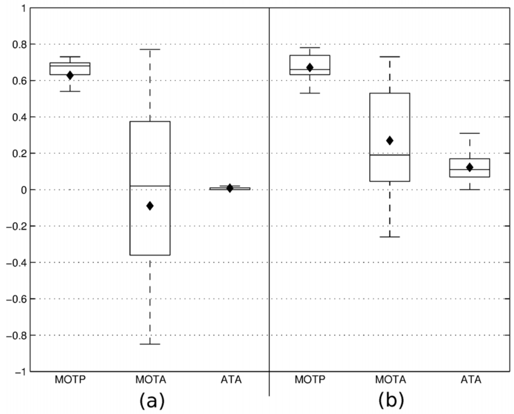

Fig. 1. CLEAR-MOT and ATA metrics for (a) the baseline algorithm and (b) *TextSpotter* [20], [21], [22] method. Diamond markers indicate average values.

C. Results and Discussion

Figure 1 shows the box-plots of the detection and tracking scores for comparison of the baseline algorithm and the only participant in this challenge.

The baseline algorithm consists in a detection stage performed using the ABBYY OCR SDK (with same set-up as in the other challenges of the competition - see section III), and a tracking stage where each detected word is assigned the identifier of the previously detected word with the best overlapping ratio (at least 0.5) searching backwards in a buffer of the 7 prior frames. In order to prevent an excessive number of false positives the algorithm takes a conservative strategy and words are not reported in the output unless there is a matching word detected in the immediate previous frame.

Figure 1(a) shows the baseline algorithm performance in terms of the metrics described in the previous section. The mean tracking precision (MOTP) was 0.63, and the mean tracking accuracy (MOTA) was -0.09, much lower due to the high number of false positives. Notice that in equation (2) negative values of MOTA are obtained when the evaluated method counts more false positives and/or ID-switches per frame than the actual number of words in the ground-truth. The ATA score obtained by the baseline algorithm remained close to zero in all video sequences of the test set, as a consequence of the high fragmentation in temporal overlapping produced by such a simplistic tracking strategy.

The *TextSpotter* [20], [21], [22] method (see Figure 1(b)) significantly outperforms the baseline algorithm in the MOTA and ATA scores, while achieving a slightly better MOTP. The mean for MOTP and MOTA metrics where 0.67 and 0.27 respectively, and the mean ATA was 0.12. In this comparison it is clear that the TextSpotter method is able to detect more words, while generating less false positives and ID-switches than the baseline algorithm.

The obtained results show coherency with the ones reported in [31] for text detection and tracking in video while being lower in general, however is worth to notice that the datasets are very different in nature and also the evaluated tasks differ: here the evaluation is done at the word level while in [31] it was done at the text block level.

VI. Conclusions

An overview of organization and the results of ICDAR 2013 Robust Reading Competition was given in this paper. Besides the two challenges on extracting text from still images, a new challenge of text extraction from scene videos was introduced. Although limited participation was made in the newly introduced challenge, we hope this competition will trigger research on this exciting topic. It is planned to present the detailed results of the competition in a journal paper due to space restrictions here. Further results, as well as a visualisation of the results of each method on an image by image basis will be given in the Web site of the competition at [6]. Also, continuous submission in all three challenges is encouraged and open.

Acknowledgment

This work was supported by the Spanish research projects TIN2011-24631, RYC-2009-05031 and JST CREST project.

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, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto *et al.*, “ICDAR 2003 robust reading competitions: entries, results, and future directions,” *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 105–122, 2005.

[3] 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.

[4] 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.

[5] A. Shahab, F. Shafait, and A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1491–1496.

[6] ICDAR 2013 Robust Reading Competition. [Online]. Available: http://dag.cvc.uab.es/icdar2013competition/

[7] D. Kumar and A. Ramakrishnan, “OTCYMIST: Otsu-Canny minimal spanning tree for born-digital images,” in *Proc. Int. Workshop on Document Analysis Systems*, 2012.

[8] B. Su, S. Lu, T. Q. Phan, and C. L. Tan, “Character extraction in web image for text recognition,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3042–3045.

[9] H. Yang, B. Quehl, and H. Sack, “A framework for improved video text detection and recognition,” *Multimedia Tools and Applications*, pp. 1–29, 2012.

[10] A. González, L. M. Bergasa, J. J. Yebes, and S. Bronte, “Text location in complex images,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 617–620.

[11] 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.

[12] 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 (IJDAR)*, vol. 8, no. 4, pp. 280–296, 2006.

[13] Abbyy FineReader. [Online]. Available: finereader.abbyy.com/

[14] ICDAR 2011 Robust Reading Competition - Challenge 1: ”Reading Text in Born-Digital Images (Web and Email)”. [Online]. Available: http://www.cvc.uab.es/icdar2011competition/

[15] J. Fabrizio, B. Marcotegui, and M. Cord, “Text segmentation in natural scenes using toggle-mapping,” in *Proc. Int. Conf. on Image Processing*, 2009.

[16] ——, “Text detection in street level images,” *Pattern Analysis and Applications*, 2013, (Accepted for publication).

[17] D. Kumar, M. Anil, and A. Ramakrishnan, “MAPS: midline analysis and propagation of segmentation,” in *Proc. Indian Conf. on Vision, Graphics and Image Processing*, 2012.

[18] D. Kumar and A. Ramakrishnan, “Power-law transformation for enhanced recognition of born-digital word images,” in *Proc. 9th Int. Conf. on SPCOM*, 2012.

[19] D. Kumar, M. Anil, and A. Ramakrishnan, “NESP: nonlinear enhancement and selection of plane for optimal segmentation and recognition of scene word images,” in *Proc. SPIE Document Recognition and Retrieval XX*, 2013.

[20] L. Neumann and J. Matas, “A method for text localization and recognition in real-world images,” in *Proc. Asian Conf. on Computer Vision*, 2010, pp. 2067–2078.

[21] ——, “Real-time scene text localization and recognition,” in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3538–3545.

[22] ——, “On combining multiple segmentations in scene text recognition,” in *Proc. Int. Conf. on Document Analysis and Recognition*, 2013.

[23] Y.-M. Zhang, K.-Z. Huang, and C.-L. Liu, “Fast and robust graph-based transductive learning via minimum tree cut,” in *Proc. Int. Conf. on Data Mining*, 2011.

[24] C.-L. L. B. Bai, F. Yin, “Scene text localization using gradient local correlation,” in *Proc. Int. Conf. on Document Analysis and Recognition*, 2013.

[25] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, “Scene text recognition using part-based tree-structured character detections,” in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2013.

[26] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao, “Scene text detection using graph model built upon maximally stable extremal regions,” *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107–116, 2013.

[27] N. Tatiana, O. Barinova, P. Kohli, and V. Lempitsky, “Large-lexicon attribute-consistent text recognition in natural images,” in *Proc. European Conf. on Computer Vision*, 2012, pp. 752–765.

[28] J. J. Weinman, E. Learned-Miller, and A. Hanson, “A discriminative semi-Markov model for robust scene text recognition,” in *Proc. Intl. Conf. on Pattern Recognition*, 2008.

[29] J. Weinman, Z. Butler, D. Knoll, and J. Feild, “Toward integrated scene text reading,” *IEEE Trans. on Pattern Anal. Mach. Intell.*, 2013, to be published.

[30] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, “Passive capture and ensuing issues for a personal lifetime store,” in *Proc. 1st ACM workshop on Continuous archival and retrieval of personal experiences (CARPE’04)*, 2004, pp. 48–55.

[31] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.

[32] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” *EURASIP Journal on Image and Video Processing*, vol. 2008, May 2008.

[33] A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi, “Compact and efficient posterity logging of face imagery for video surveillance,” *IEEE Multimedia*, vol. 19, no. 4, pp. 48–59, 2012.

Appendix: Description of the Participating Methods

Method Name	Authors	Affiliation	Brief Description	Participation in Challenge 1			Participation in Challenge 2		
				TaskT1T	TaskT2T	TaskT3T	TaskT1T	TaskT2T	TaskT3T
Text_detector_CASIAT	CunzhaoTShi,TYangTZhang,TChunhengTWang,TBaihuaTXiao,TSongTGao,TJinlongTHuT	InstituteTofTAutomation,TChineseTAcademyTofTSciences,TBeijing,TChinaT	First,TMSERsTareTdetectedTasTtheTbasicTconnectedTcomponents.TThen,TaTregionTbasedTclassifierTisTusedTtoTexcludeTsomeTnon-textTMSERsTandTtheTleftTonesTareTgroupedTintoTcandidateTtextTregionsTaccordingTtoTtheTposition,TsizeTandTcolorTofTeachTMSER.TFinally,Ttree-structuredTcharacterTmodelsT(TSM)TareTappliedTonTtheTtextTregionsTtoTsearchTtheTmissingTcharactersTandTeliminateTtheTfalseTpositives.T	T	T	T	✓T	T	T
USTB_TexStarT	XuwangTYin[1],TXu-ChengTYin[1]TandTHong-WeiTHao[2]T	[1]TUniversityTofTScienceTandTTechnology,TBeijing,TChinaT [2]TInstituteTofTAutomation,TTChineseTAcademyTofTSciences,TBeijing,TChinaT	HierarchicalTstructureTofTMSERsTisTusedTtoTextractTcharacterTcandidates.TCharacterTcandidatesTareTmergedTtoTformTtextTregionsTusingTaTsingle-linkTclusteringTalgorithm.TFinally,Tnon-textTandTtextTclassifiersTareTsequentiallyTappliedTtoTgetTfinalTtextTregions.T	✓T	T	T	✓T	T	T
USTB_FuStarT			TextTregionsTobtainedTusingTUSTB_TexStarTareTthresholdedTbasedTonTMSERs.TPost-processingTisTperformedTtoTrecoverTsmallTtextTcomponentsTthatTmightTbeTmissedTbyTtheTpreviousTstage.T	T	✓T	T	T	✓T	T
Feild'sTMethodT	JacquelineTFeildTandTErikTLearned-MillerT	UniversityTofTMassachusetts,TAmherst,TUSAT	ATregressionTbasedTtextTsegmentationTtechniqueTisTusedTtoTsegmentTtextTfromTbackground.TWordTrecognitionTisTperformedTusingTaTConditionalTRandomTFieldTfollowedTbyTaTlanguageTmodelTtrainedTfromToverT13TmillionTwords.T	T	T	T	T	T	✓T
PhotoOCRT	AlessandroTBissacco,TMarkTCummins,TYuvalTNetzer,THartmutTNevenT	GoogleTInc.,TUSAT	ClassicalTover-segmentationTandTbeamTsearchTarchitectureTisTused.TATdeepTneuralTnetworkTservesTasTtheTcharacterTclassifier,TwhereasTcharacter-ngramsTareTusedTforTlanguageTmodelling.TTopThypothesisTfromTtheTbeamTsearchTareTre-rankedTusingTaTword-ngramTmodel.TT	T	T	✓T	T	T	✓T
BDTD_CASIAT	YangTZhang,CunzhaoTShi,TSongTGao,TJinlongTHu,TChunhengTWang,TBaihuaTXiaoT	InstituteTofTAutomation,TChineseTAcademyTofTSciences,TBeijing,TChinaT	ColorTgradientsTareTusedTtoTbinarizeTtheTimage.TTextTregionsTareTextractedTusingTseveralTheuristicTrulesTincorporatingTmorphology,Tdensity-basedTfiltering,TandTsize,Talignment,TandTcolorTconsistency.TFinallyTSVMsTareTusedTtoTfurtherTrefineTtheTresultsTbyTremovingTfalseTalarms.T	✓T	T	T	T	T	T
PicReadT	TatianaTNovikova[1],TOlgaTBarinova[1],TSergeyTMilyaev[1],TVladimirTKirichenkov[1],TAlexanderTSapatov[1],TPushmeetTKohli[2],TVictorTLempitsky[3]T	[1]TLomonosovTMoscowTStateTUniversity,TMoscow,TRussiaTT [2]TMicrosoftTResearch,TCambridge,TUKT [3]TSkolkovoTInstituteTofTScienceTandTTechnology,TMoscow,TRussiaT	WordTrecognitionTisTperformedTinTaTunifiedTprobabilisticTframeworkTusingTmaximumTaTposterioriT(MAP)TinferenceTbasedTonTweightedTfiniteTstateTtransducers.TTheTmodelTenforcesTbothTtheTlanguageTconsistency,TandTtheTconsistencyTofTtheTattributesTofTlettersTthatTconstituteTaTword.T	T	T	T	T	T	✓T
NSTsegmentorT	SergeyTMilyaev[1],TOlgaTBarinova[1],TTatianaTNovikova[1],TPushmeetTKohli[2],TVictorTLempitsky[3]T		ImageTbinarizationTincorporatesTlocalTcues,TsuchTasTlocalTbinarizationTmapTandTimageTLaplacian,TintoTaTglobalToptimizationTframework.TTextline-candidatesTareTgeneratedTbasedTonTsimilarityTofTtheTsizeTandTcolorTofTfilteredTconnectedTcomponents.TText-lineTcandidatesTwithTlowerTclassifierTscoresTareTfilteredTout.T	T	T	T	T	✓T	T
NSTextractorT			ThisTmethodTisTsimilarTtoTNSTsegmentatorTexceptTthatTfilteringTofTtext-lineTcandidatesTalsoTtakingTintoTaccountTtheTnumberTofTbackgroundTsegments.T	T	T	T	T	✓T	T
CASIA_NLPRT	BoTBaiT	InstituteTofTAutomation,TChineseTAcademyTofTSciences,TBeijing,TChinaT	UsingTgradientTlocalTcorrelation,TtheTdensityTofTpairwiseTedgesTofToppositeTdirections,TandTtheTstrokeTwidthTconsistencyTareTbothTcharacterizedTtoTcomputeTaTtextTconfidenceTmap.TTextTcandidatesTareTobtainedTusingTaTfastTsemi-supervisedTlearningTmethodT[4]TandTareTfurtherTprunedTusingTanTSVMTclassifier.TFeaturesTsuchTasTcolorTandTstrokeTwidthTareTusedTtoTobtainTtext-linesTandTsubsequentlyTwords.T	T	T	T	✓T	T	T

APPENDIX CONTD.: DESCRIPTION OF THE PARTICIPATING METHODS

Method Name	Authors	Affiliation	Brief Description	Participation in Challenge 1			Participation in Challenge 2		
				Task 1	Task 2	Task 3	Task 1	Task 2	Task 3
TH-TextLoc	Cheng Yang, Changsong Liu, Xiaoqing Ding	Tsinghua University, Beijing, China	This is an improved version of the method that took part in ICDAR 2011 Robust Reading competition. The key enhancement is the use of Conditional Random Field for more precise text candidate selection.	✓			✓		
OTCYMIST	D. Kumar and A. G. Ramakrishnan	Indian Institute of Science Bangalore, India	Individual color channels are binarized separately using Otsu's threshold. Minimum spanning tree based grouping is performed on detected components to filter background components and group text candidates into words.	✓	✓			✓	
PLT			The gray scale is enhanced by applying power-law transform. Enhanced gray scale image is segmented using Otsu's threshold. Omnipage OCR is used for word recognition from the binarized image.			✓			✓
NESP			Fischer discrimination factor is calculated for various color planes with different power-law values. The plane with maximum discrimination value is selected for binarization.			✓			✓
MAPS			A local window based min-max thresholding criteria incorporating means and variances of thus generated foreground/background regions are used.			✓			✓
Text Detection	Raphaël Boissel[1], Ana Stefania Calarasanu[1], Jonathan Fabrizio[1], Séverine Dubuisson[2]	[1] EPITA Research and Development Laboratory (LRDE-EPITA), France [2] Université Pierre et Marie Curie (LIP6-UPMC), France	This method has two main steps: a hypothesis generation step to get potential text boxes and a hypothesis validation step to filter false detections. The hypothesis generation step process relies on the TMMS segmentation method, while the validation step is based on a texture-based SVM classification approach.	✓	✓		✓	✓	
Text Spotter	Lukas Neumann, Jiri Matas, Michal Busta	Czech Technical University, Czech Republic	Text Spotter is an unconstrained real-time end-to-end text localization and recognition method. The real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of Extremal Regions (ERs). ERs are grouped into word regions which are recognized using an approximate nearest-neighbor classifier operating on a coarse Gaussian scale-space pyramid. A demo of the software is available online: http://www.textspotter.org/				✓		✓
PIONEER	Jerod Weinman	Grinnell College, Iowa, USA	Each word image is over-segmented and logistic regression is used to classify text components. Polynomials fit to the tops and bottoms of binarized characters are normalized to a horizontal, linear orientation with a thin-plate spline. A discriminative semi-Markov model jointly segments and recognizes characters in the normalized image using steerable pyramid features, character bigrams, and a large lexicon.						✓
I2R_NUS_FAR	Lu Shijian[1], Tian Shangxuan[2], Lim Joo Hwee[1], Tan Chew Lim[2]	[1] Institute for Infocomm Research, A*STAR, Singapore [2] School of Computing, National University of Singapore	This method builds upon I2R_NUS by further reducing false alarms through a machine learning approach.	✓	✓		✓	✓	
I2R_NUS			This method makes use of perceptual saliency of texts in scenes/born digital images. In particular, four text-specific saliency features are designed that are consistently accompanied with texts in scene/web images. Combined with the text layout information, the saliency features are integrated to classify text and non-text objects accurately. After text localization, text boundary is extracted to get the binary text images.						
Inkam	None specified	None specified	None specified; Link to software: http://inkamocr.com/eng/	✓			✓		
Baseline	Organizers	N/A	ABBYY OCR SDK v10	✓		✓	✓		✓