ICDAR 2013 Robust Reading Competition

http://dag.cvc.uab.es/icdar2013competition/
ICDAR 2013 Robust Reading Competition

**Challenge 1:** “Reading Text in Born-Digital Images (Web and Email)"

**Challenge 2:** “Reading Text in Scene Images”

**Challenge 3:** “Reading Text in Videos”
Timeline

- **Web Site Online Registrations Open**: 15/1
- **Training Set Online**: 28/2
- **Test Set Available and Submission Open**: 29/3
- **Submission Closed**: 8/4

**Open mode participation**: Participants run their own algorithms and provided results.
## Participation in Numbers

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits of the Web Site (Unique Visitors)</td>
<td>2,724 (1,025)</td>
</tr>
<tr>
<td>Registered Users</td>
<td>124</td>
</tr>
<tr>
<td>Total Number of Individual Participants</td>
<td>15</td>
</tr>
<tr>
<td>Total Number of Submissions</td>
<td>52 (42 excluding variants)</td>
</tr>
<tr>
<td>Challenge 1 #Submissions (#Participants)</td>
<td>17 (8)</td>
</tr>
<tr>
<td>Challenge 2 #Submissions (#Participants)</td>
<td>22 (13)</td>
</tr>
<tr>
<td>Challenge 3 #Submissions (#Participants)</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>
Reading text in Born-Digital Images
Reading text in Scene Images

CHALLENGES 1 AND 2
Overview

**Born Digital Images**
- Low-Resolution
- Digitally Created Text
- Compression
- Anti-Aliasing

**Real Scene Images**
- High-Resolution
- Captured Text
- Illumination Artefacts
- Perspective
Structure

Challenges 1 and 2 were organised over three tasks:

Task 1 – **Text Localization**

**Objective**: To obtain a rough estimation of text areas in the image, in terms of bounding boxes corresponding to parts of text (words or text lines)

Task 2 – **Text Segmentation**

**Objective**: Pixel-level separation of text from the background

Task 3 – **Word Recognition**

**Objective**: Assuming known word bounding boxes, to obtain the correct text transcriptions
Datasets

**Updated datasets** compared to the 2011 edition:
- Extra test images in Challenge 1
- Segmentation ground truth and filtering of duplicate images in Challenge 2
- “Don’t Care” regions introduced in the ground truth

Datasets in Numbers

<table>
<thead>
<tr>
<th></th>
<th>Challenge 1</th>
<th>Challenge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Dataset (Full Images)</td>
<td>420</td>
<td>229</td>
</tr>
<tr>
<td>Test Dataset (Full Images)</td>
<td>141</td>
<td>233</td>
</tr>
<tr>
<td>Training Dataset (Word Images)</td>
<td>3564</td>
<td>848</td>
</tr>
<tr>
<td>Test Dataset (Word Images)</td>
<td>1439</td>
<td>1095</td>
</tr>
</tbody>
</table>
Ground Truth Task 1 – Text Localization

Dataset Images

- How to Find the Perfect HDTV
- The Photo Specialists
- HIGHER SAVINGS RATES

Ground Truth (text files)

```
11, 15, 42, 28, "How"
41, 16, 61, 28, "to"
8, 33, 36, 46, "Find"
41, 32, 64, 46, "the"
11, 50, 61, 65, "Perfect"
16, 69, 56, 82, "HDTV"
```

```
22 249 113 286 "The"
142 249 287 286 "Photo"
326 245 620 297 "Specialists"
```

```
0, 1, 177, 32, "\"HIGHER"
11, 35, 182, 63, "SAVINGS"
12, 67, 183, 103, "RATES\"
50, 114, 56, 120, "A"
60, 114, 91, 120, "reward"
96, 114, 108, 120, "for"
112, 116, 126, 120, "our"
130, 114, 164, 120, "current"
...```

Ground Truth Visualisation

- Visualisation of text localization in the images
Performance Evaluation Task 1 – Text Localization

Performance Evaluation Methodology

• Methodology proposed by Wolf and Jolion [1]
• Takes into account both
  • Bounding Box area overlapping and
  • Precision at the level of detection counts
• Possible to create meaningful cumulative results over many images
• Ways to deal with one-to-many and many-to-one cases
• Set up to penalise over-segmentation (words split to parts), but no under-segmentation (group of words detected as text line)

Baseline Method

• Commercial OCR package (ABBYY OCR SDK v.10)

## Results: Task 1.1 – Text Localization

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USTB_TexStar</td>
<td>82.38%</td>
<td>93.83%</td>
<td>87.73%</td>
</tr>
<tr>
<td>TH-TextLoc</td>
<td>75.85%</td>
<td>86.82%</td>
<td>80.97%</td>
</tr>
<tr>
<td>I2R_NUS_FAR</td>
<td>71.42%</td>
<td>84.17%</td>
<td>77.27%</td>
</tr>
<tr>
<td>Baseline</td>
<td>69.21%</td>
<td>84.94%</td>
<td>76.27%</td>
</tr>
<tr>
<td>Text Detection</td>
<td>73.18%</td>
<td>78.62%</td>
<td>75.80%</td>
</tr>
<tr>
<td>I2R_NUS</td>
<td>67.52%</td>
<td>85.19%</td>
<td>75.33%</td>
</tr>
<tr>
<td>BDTD_CASIA</td>
<td>67.05%</td>
<td>78.98%</td>
<td>72.53%</td>
</tr>
<tr>
<td>OTCYMIST</td>
<td>74.85%</td>
<td>67.69%</td>
<td>71.09%</td>
</tr>
<tr>
<td>Inkam</td>
<td>52.21%</td>
<td>58.12%</td>
<td>55.01%</td>
</tr>
</tbody>
</table>

![Graph showing precision and recall for various methods]
## Results: Task 2.1 – Text Localization

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USTB_TextStar</td>
<td>66.45%</td>
<td>88.47%</td>
<td>75.90%</td>
</tr>
<tr>
<td>Text Spotter</td>
<td>64.84%</td>
<td>87.51%</td>
<td>74.49%</td>
</tr>
<tr>
<td>CASIA_NLPR</td>
<td>68.24%</td>
<td>78.89%</td>
<td>73.18%</td>
</tr>
<tr>
<td>Text_Detector_CASIA</td>
<td>62.85%</td>
<td>84.70%</td>
<td>72.16%</td>
</tr>
<tr>
<td>I2R_NUS_FAR</td>
<td>69.00%</td>
<td>75.08%</td>
<td>71.91%</td>
</tr>
<tr>
<td>I2R_NUS</td>
<td>66.17%</td>
<td>72.54%</td>
<td>69.21%</td>
</tr>
<tr>
<td>TH-TextLoc</td>
<td>65.19%</td>
<td>69.96%</td>
<td>67.49%</td>
</tr>
<tr>
<td>Text Detection</td>
<td>53.42%</td>
<td>74.15%</td>
<td>62.10%</td>
</tr>
<tr>
<td>Baseline</td>
<td>34.72%</td>
<td>60.76%</td>
<td>44.21%</td>
</tr>
<tr>
<td>Inkam</td>
<td>35.27%</td>
<td>31.20%</td>
<td>33.11%</td>
</tr>
</tbody>
</table>

![Graph showing precision and recall for various methods](image)
Ground Truth Task 2 – Text Segmentation

Dataset Images

Ground Truth (images)

okcupid

How to Find the Perfect HDTV

The Photo Specialists

okcupid

How to Find the Perfect HDTV

The Photo Specialists
Performance Evaluation Task 2 – Text Segmentation

Performance Evaluation Methodology

• Framework proposed by Clavelli et al [2]
• Measures the degree to which morphological properties of the text are preserved, not simply the number of misclassified pixels
• As a secondary evaluation scheme we implemented standard pixel level precision and recall (compatibility with other results)

(a)  
(b)  
(c)  

Results: Task 1.2 – Text Segmentation

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USTB_FuStar</td>
<td>80.01%</td>
<td>86.20%</td>
<td>82.99%</td>
</tr>
<tr>
<td>I2R_NUS</td>
<td>64.57%</td>
<td>73.44%</td>
<td>68.72%</td>
</tr>
<tr>
<td>OTCYMIST</td>
<td>65.75%</td>
<td>71.65%</td>
<td>68.57%</td>
</tr>
<tr>
<td>I2R_NUS_FAR</td>
<td>59.05%</td>
<td>80.04%</td>
<td>67.96%</td>
</tr>
<tr>
<td>Text Detection</td>
<td>49.64%</td>
<td>69.46%</td>
<td>57.90%</td>
</tr>
</tbody>
</table>
Results: Task 2.2 – Text Segmentation

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I2R_NUS_FAR</td>
<td>68.64%</td>
<td>80.59%</td>
<td>74.14%</td>
</tr>
<tr>
<td>NSTextractor</td>
<td>63.38%</td>
<td>83.57%</td>
<td>72.09%</td>
</tr>
<tr>
<td>USTB_FuStar</td>
<td>68.03%</td>
<td>72.46%</td>
<td>70.18%</td>
</tr>
<tr>
<td>I2R_NUS</td>
<td>60.33%</td>
<td>76.62%</td>
<td>67.51%</td>
</tr>
<tr>
<td>NSTsegmentator</td>
<td>68.00%</td>
<td>54.35%</td>
<td>60.41%</td>
</tr>
<tr>
<td>Text Detection</td>
<td>62.03%</td>
<td>57.43%</td>
<td>59.64%</td>
</tr>
<tr>
<td>OTCYMIST</td>
<td>41.79%</td>
<td>31.60%</td>
<td>35.99%</td>
</tr>
</tbody>
</table>
Ground Truth Task 3 – Word Recognition

Dataset Images

1.png, "okcupid"
2.png, "How"
3.png, "to"
4.png, "Find"
5.png, "the"
6.png, "Perfect"
7.png, "HDTV"
8.png, "Creative"
9.png, "Printing"
10.png, "Community"
11.png, "KIDS:
12.png, "LEARN"
13.png, "ENGLISH"
14.png, "ONLINE"
15.png, "\"HIGHER\
16.png, "Games,"
17.png, "stories,"
18.png, "and"
19.png, "more"
20.png, "RATES\"

Ground Truth (single text file for all the words)
Performance Evaluation Task 3 – Word Recognition

Performance Evaluation Methodology

- **Edit distance** (normalised to the length of the ground truth transcription)
- Equal weights for deletions, additions, substitutions
- We also report statistics on correctly recognised words (case sensitive)

Baseline method

- Commercial OCR package (ABBYY OCR SDK v.10)
## Results: Task 1.3 – Word Recognition

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Total Edit Distance</th>
<th>Correctly Recognized Words (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhotoOCR</td>
<td>105.5</td>
<td>82.21</td>
</tr>
<tr>
<td>MAPS</td>
<td>196.2</td>
<td>80.4</td>
</tr>
<tr>
<td>PLT</td>
<td>200.4</td>
<td>80.26</td>
</tr>
<tr>
<td>NESP</td>
<td>214.5</td>
<td>79.29</td>
</tr>
<tr>
<td>Baseline</td>
<td>409.4</td>
<td>60.95</td>
</tr>
</tbody>
</table>
# Results: Task 2.3 – Word Recognition

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Total Edit Distance</th>
<th>Correctly Recognized Words (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhotoOCR</td>
<td>122.7</td>
<td>82.83</td>
</tr>
<tr>
<td>PicRead</td>
<td>332.4</td>
<td>57.99</td>
</tr>
<tr>
<td>NESP</td>
<td>360.1</td>
<td>64.20</td>
</tr>
<tr>
<td>PLT</td>
<td>392.1</td>
<td>62.37</td>
</tr>
<tr>
<td>MAPS</td>
<td>421.8</td>
<td>62.74</td>
</tr>
<tr>
<td>Feild’s Method</td>
<td>422.1</td>
<td>47.95</td>
</tr>
<tr>
<td>PIONEER</td>
<td>479.8</td>
<td>53.70</td>
</tr>
<tr>
<td>Baseline</td>
<td>539.0</td>
<td>45.30</td>
</tr>
<tr>
<td>TextSpotter</td>
<td>606.3</td>
<td>26.85</td>
</tr>
</tbody>
</table>
Reading Text in Videos

CHALLENGE 3
Overview - Structure

Challenge 3 was organised over a single task:

**Task 1 – Text Localization**

**Objective:** The objective of this task is to obtain a localisation of the words in the video in terms of their *affine bounding boxes*. The task requires that words are both localised correctly in every frame and tracked correctly (assigned the same ID) over the video sequence.
Datasets

New Datasets created for this challenge

• Video sequences collected in different countries (English, French, Spanish)
• Users were given seven different tasks (e.g. Driving, follow way-finding panels, browse products in a super market, etc)
• Four different cameras were used (head-mounted, mobile, hand-held 720p, tripod high-definition 1080p)
• Sequences ranging from 10” to 1’06” each

Dataset in Numbers

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Videos in Training Set</td>
<td>13</td>
</tr>
<tr>
<td>Number of Frames in Training Set</td>
<td>5,486</td>
</tr>
<tr>
<td>Number of Videos in Test Set</td>
<td>15</td>
</tr>
<tr>
<td>Number of Frames in Test Set</td>
<td>9,790</td>
</tr>
<tr>
<td>Total Video Length</td>
<td>9’35”</td>
</tr>
</tbody>
</table>
Ground Truth

Frame Image

Ground Truth Visualisation

Ground Truth (XML file)

```xml
<?xml version="1.0" encoding="us-ascii"?>
<Frames>
  <frame ID="1" />
  <frame ID="2" />
  <frame ID="3" />
  ...
  <frame ID="434">
    <object Transcription="Badia" ID="377001"
      Language="Catalan" Quality="MODERATE">
      <Point x="446" y="58" />
      <Point x="447" y="69" />
      <Point x="484" y="67" />
      <Point x="484" y="55" />
    </object>
    <object Transcription="Barbera" ID="377002"
      Language="Catalan" Quality="MODERATE">
      <Point x="502" y="81" />
      <Point x="447" y="85" />
      <Point x="446" y="73" />
      <Point x="502" y="70" />
    </object>
    ...
  </frame>
  ...
  <frame ID="1859" />
  <frame ID="1860" />
</Frames>
```
Performance Evaluation

- Evaluation based on CLEAR-MOT [3, 4], and VACE [5] metrics
  - Multiple Object Tracking Precision (MOTP), expresses how well locations of words are estimated
  - Multiple Object Tracking Accuracy (MOTA), how many mistakes the tracker has made (false positives, ID mismatches, etc), can be negative
  - Average Tracking Accuracy (ATA), comprehensive spatio-temporal measure.

Baseline Method

- Detection stage performed using the ABBYY OCR SDK, followed by a tracking stage where each detected word is assigned the identifier of the previously detected word with the best overlapping ratio

Results: Task 3.1 – Text Localization

<table>
<thead>
<tr>
<th>Method Name</th>
<th>MOTP</th>
<th>MOTA</th>
<th>ATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextSpotter</td>
<td>0.67</td>
<td>0.23</td>
<td>0.12</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.63</td>
<td>-0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>
WHAT’S NEXT?
What’s Next

The competition site is open! Register to download the datasets and upload new results

http://dag.cvc.uab.es/icdar2013competition

- The competition is now open on a **continuous mode**
- **Datasets** are freely available (on Competition Web site and soon on TC11)
- **Online performance evaluation functionality**
- **Ranking charts and tables**
- **Visualisation** of the results of all tasks, and all challenges
Online Evaluation and Ranking Tables

Results

Ranking for Task 1 - Text Localization

Ranking for Task 2 - Text Segmentation

Ranking for Task 3 - Word Recognition

ICDAR 2013 Robust Reading Competition

Challenge 2: Reading Text in Image
Visualisation of different Tasks

Task 1 - Text Localization

Task 2 - Text Segmentation

Task 3 - Word Recognition

Method comparison

images list

Method

<table>
<thead>
<tr>
<th>Word 1111</th>
<th>method: PhotoOCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1112</td>
<td>method: PhotoOCR</td>
</tr>
</tbody>
</table>

Method comparison

Prel results

Atom based results

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
<th>Well Seg</th>
<th>Merged</th>
<th>Broken</th>
<th>BR Margin</th>
<th>Lost</th>
<th>False Pos</th>
<th>Det. Atoms</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>USTB_TextStar</td>
<td>82.46</td>
<td>90.54</td>
<td>86.33</td>
<td>52</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>69</td>
<td>79.49</td>
<td>89.86</td>
<td>84.85</td>
</tr>
<tr>
<td>OTOMAT</td>
<td>86.18</td>
<td>78.02</td>
<td>85.78</td>
<td>85</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>53</td>
<td>48.87</td>
<td>68.04</td>
<td>55.44</td>
</tr>
<tr>
<td>Saliency_Text</td>
<td>84.38</td>
<td>74.61</td>
<td>83.43</td>
<td>24</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>44</td>
<td>30.77</td>
<td>54.55</td>
<td>40.54</td>
</tr>
<tr>
<td>I3R_NUS_V4</td>
<td>84.61</td>
<td>74.61</td>
<td>83.43</td>
<td>27</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>55</td>
<td>34.62</td>
<td>49.09</td>
<td>40.61</td>
</tr>
<tr>
<td>text_loc....</td>
<td>64.90</td>
<td>85.53</td>
<td>73.71</td>
<td>10</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>24.85</td>
<td>51.55</td>
<td>33.04</td>
</tr>
</tbody>
</table>
Task 1.1 – Text Localization in Born Digital Images
“USTB_TexStar”, Xuwang Yin¹, Xu-Cheng Yin¹ and Hong-Wei Hao²
¹ University of Science and Technology, Beijing, China
² Institute of Automation, Chinese Academy of Sciences, Beijing, China

Task 1.2 – Text Segmentation in Born Digital Images
“USTB_FuStar”, Xuwang Yin¹, Xu-Cheng Yin¹ and Hong-Wei Hao²
¹ University of Science and Technology, Beijing, China
² Institute of Automation, Chinese Academy of Sciences, Beijing, China

Task 1.3 – Word Recognition in Born Digital Images
“Photo OCR”, Alessandro Bissacco, Mark Cummins, Yuval Netzer, Hartmut Neven
Google Inc., USA

Task 2.1 – Text Localization in Real Scenes
“USTB_TexStar”, Xuwang Yin¹, Xu-Cheng Yin¹ and Hong-Wei Hao²
¹ University of Science and Technology, Beijing, China
² Institute of Automation, Chinese Academy of Sciences, Beijing, China

Task 2.2 – Text Segmentation in Real Scenes
“I2R_NUS_FAR”, Lu Shijian¹, Tian Shangxuan², Lim Joo Hwee¹, Tan Chew Lim²
¹ Institute for Infocomm Research, A*STAR, Singapore
² School of Computing, National University of Singapore

Task 2.3 – Word Recognition in Real Scenes
“PhotoOCR”, Alessandro Bissacco, Mark Cummins, Yuval Netzer, Hartmut Neven
Google Inc., USA

Task 3.1 – Text Localization in Videos (Mention)
“TextSpotter”, Lukas Neumann, Jiri Matas, Michal Busta
Czech Technical University, Czech Republic