NEOCR: A Configurable Dataset for Natural Image Text Recognition

Robert Nagy, Anders Dicker, Klaus Meyer-Wegener

Chair for Computer Science 6 (Data Management)
Friedrich-Alexander-University of Erlangen-Nürnberg

22.09.2011
Contents

- motivation
- related work
- NEOCR dataset
- experiments
- conclusion
Motivation

- applications for OCR in natural images
  - help for visually impaired
  - mobile applications
  - object classification
  - image annotation
  - driving assistant systems
Differences to Scanned Document OCR

- background
- blurredness
- camera position
- character arrangement
- colors
- contrast
- font size

- font type (diversity in document)
- font type (in general)
- noise
- number of lines
- occlusion
- rotation
- surface
- **current datasets for OCR in natural images**
  - ICDAR 2003 (Robust Reading) Dataset
  - Chars74K
  - Street View Text Dataset
  - Microsoft Text Detection Database

- **limitations of current datasets**
  - bounding boxes parallel to the axes
  - not all text visible in the images annotated
  - limited scope of natural image variations
  - no additional metadata
### Related Work

- **Comparison of natural image text recognition datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># images</th>
<th># boxes</th>
<th>avg. # char/box</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICDAR 2003</td>
<td>509</td>
<td>2263</td>
<td>6.15</td>
</tr>
<tr>
<td>Chars74K</td>
<td>312</td>
<td>2112</td>
<td>6.47</td>
</tr>
<tr>
<td>MS Text DB</td>
<td>307</td>
<td>1729</td>
<td>10.76</td>
</tr>
<tr>
<td>Street View Text</td>
<td>350</td>
<td>904</td>
<td>6.83</td>
</tr>
<tr>
<td><strong>NEOCR</strong></td>
<td><strong>659</strong></td>
<td><strong>5238</strong></td>
<td><strong>17.62</strong></td>
</tr>
</tbody>
</table>

**NEOCR Dataset 22.09.2011**
NEOCR Dataset

- 659 images, 5238 textfields
- distortion quadrangles
- rich annotation (XML)
  - optical,
  - geometrical and
  - typographical characteristics

22.09.2011
NEOCR Dataset
Distortion Quadrangles

**distorted**

**straightened**

22.09.2011
NEOCR Dataset
Experiments

- configurable dataset
- special test scenarios, e.g.:
  - character arrangement
  - contrast / inversion
  - distortion
  - impact of vocabularies
  - combinations of these and more

- results using popular open source and commercial OCR tools
  - only bounding boxes were used
  - Levenshtein (edit-)distance

- similarity: $s(x_n, y_m) = 1 - \frac{d(x_n, y_m)}{\max(n, m)}$
Character Arrangement

- A
- B
- C
- D
- E
- F
- G
- H
- I

horizontal  vertical  circular

0  0.1  0.2  0.3  0.4  0.5
Conclusions and Future Work

- **NEOCR dataset**
  - all visible textfields annotated
  - rich additional metadata
  - distortion quadrangles

- **future work**
  - extend dataset with further images
  - open access to adapted annotation tool
Datasets

- NEOCR Dataset. 

- Chars74K Dataset. 
  [http://www.ee.surrey.ac.uk/CVSSP/demos(chars74k](http://www.ee.surrey.ac.uk/CVSSP/demos(chars74k)

- ICDAR 2003 Robust Reading Dataset. 
  [http://algoval.essex.ac.uk/icdar/Datasets.html](http://algoval.essex.ac.uk/icdar/Datasets.html)

- Microsoft Text Detection Database. 

- Street View Text Dataset. [http://vision.ucsd.edu/~kai/svt](http://vision.ucsd.edu/~kai/svt)
Brightness

![Brightness Histogram](image)

- **Brightness (Y-channel mean)**
- **Number of images**
- **Typical scanned text document**

22.09.2011
NEOCR Dataset
Languages

![Languages Chart](image)

- **Languages**
  - German: 1800 images
  - Numbers: 700 images
  - Company: 600 images
  - Abbreviation: 400 images
  - Spanish: 300 images
  - Person: 200 images
  - French: 200 images
  - Hungarian: 100 images
  - Italian: 100 images
  - Latin: 100 images
  - Roman: 100 images
  - Czech: 100 images
  - Unknown: 100 images
  - Turkish, Portuguese, Greek, Swedish, Russian: 1 image

**NEOCR Dataset**

22.09.2011