Motivation
Can an image retrieval scheme retrieve these images together?

- Visual cues are not enough
- Tags are often noisy and incomplete

Goal: Text to image retrieval from large image/video databases

Category retrieval
Instance retrieval

Plausible Approaches
Exact text localization and recognition
- State-of-the-art text localization and recognition methods
- Text is localized and then recognized
- Retrieval becomes equivalent to that of text retrieval
- Failures are irreversible
- Not query driven

End-to-End Scene Text Recognition
- Spots the words (detects and recognizes)
- Small lexicon size (~50)
- Word spotting setting (**)
- Their Character detection module + our retrieval scheme (**)

Video Google [Shi et al., ICCV’11]
- Suitable for instance retrieval with image as a query
- Representing text queries is not trivial

Our contributions
- Query driven approach for scene text based retrieval
- Category as well as instance retrieval
- Three new datasets, including one with 1 million images

Scene Text Indexing and Retrieval

Character Detection
- Affine transformed (AT) samples
- HoG-13 ([Potamianos et al., TPAMI’10] ) 9 orientation features + 4 overall gradient energy
- Explicit feature map
- Linear SVM for speedup
- Multi-scale sliding window detection
- Character-specific NMS
- Detections are represented as a graph

Graph construction
- Graph is used for pruning false windows and in re-ranking stage
- Character windows = nodes, spatial proximity = edge
- Contextual information to prune some nodes, edges
- $\theta_3$: 36-dimensional character likelihood
- $P_{ij}$ matrix containing joint probabilities of character pairs

Vocabulary Presence Score Computation
- $\omega_0, \omega_1, ..., \omega_p$: Set of vocabulary words
- $(I_0, I_1, ..., I_p)$ Images in database
- A vocabulary of length $p$ is represented by its characters as:
  $\omega = \omega_0 \omega_1 ... \omega_p$
- Vocabulary presence score for image $I_m$ and word $\omega$ is given by:

$$S(I_m, \omega) = \max_{\omega' \in \omega} \min_{i,j} \max_i P(\omega'_i | h_i)$$

Inverted index and graph indexing

Retrieval Results on SVT

Method
- Neumann-Matas: 0.23
- Epishkin et al. : 0.19
- Mishra et al. : 0.21
- Wang et al. : 0.52

Datasets
- Public scene text datasets not very ideal for evaluating retrieval
- Image (IIT-STR) and video datasets (Sports-10K and TV Series-1M)
- Video dataset with 1M frames
- Multiple occurrences, many fonts, styles, views
- Annotated to say if an image contains a query text or not
- Available at http://cvit.iit.ac.in/projects/STR/

Qualitative results

Mean average precision (mAP)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Char. spotting</th>
<th>RSP</th>
<th>RSVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVT</td>
<td>0.06</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>KDAR11</td>
<td>0.24</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>IIT-STR</td>
<td>0.22</td>
<td>0.36</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Precision@20 on large video datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Character spotting</th>
<th>RSP</th>
<th>RSVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>0.24</td>
<td>0.38</td>
<td>0.43</td>
</tr>
<tr>
<td>TV Series</td>
<td>0.39</td>
<td>0.57</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Example:

- $S_m$: High, $S_m$: High
- $S_m$: High, $S_m$: Low
- $S_m$: Middle, $S_m$: Low
- $S_m$: Low, $S_m$: Low

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad, INDIA
http://cvit.iit.ac.in/projects/STR/