Image Retrieval Using Textual Cues

Anand Mishra [1] Karteek Alahari [2] C.V. Jawahar [1]

[1] CVIT, IIIT Hyderabad, India [2] Inria, France

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
[Neumann-Matas, CVPR'12, Epshtein et al., CVPR'10, Mishra et al., CVPR'12]

- 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

Retrieval Results on SVT

Method	mAP
Neumann-Matas	0.23
Epshtein et al. + Mishra et al.	0.19
Wang et al.*	0.21
Wang et al.**	0.52

End-to-End Scene Text Recognition
[Wang et al., ICCV'11]

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

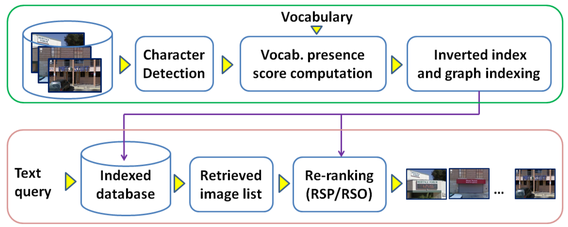
Video Google [Sivic et al., ICCV'03]

- Suitable for instance retrieval with *image as a query*
- Representing text queries is not trivial

Our contributions

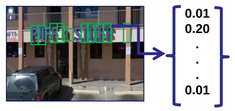
- **Query driven approach** for scene text based retrieval
- **Category as well as instance** retrieval
- Three new datasets, including one with **1 million** images

Scene Text Indexing and Retrieval



Character Detection

- Affine transformed (AT) samples
- HoG-13 [Felzenszwalb 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

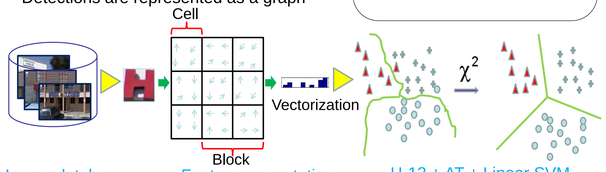


Character likelihood

$$\begin{bmatrix} 0.01 \\ 0.20 \\ \cdot \\ \cdot \\ \cdot \\ 0.01 \end{bmatrix}$$

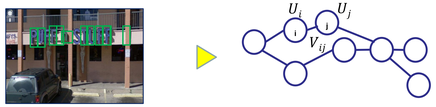
Character detection

Significant improvement in **character classification accuracy** as compared to Wang et al. [ICCV'11], Mishra et al. [CVPR'12]



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
- U_i: 36-dimensional character likelihood
- V_{ij}: matrix containing joint probabilities of character pairs



Vocabulary Presence Score Computation

- $\{\omega_1, \omega_2, \ldots, \omega_k\}$: Set of vocabulary words
- $\{I_1, I_2, \ldots, I_m\}$: Images in database
- A vocabulary of length p is represented by its characters as
  $\omega_k = \omega_{k1}\, \omega_{k2} \ldots \omega_{kp}$
- Vocabulary presence score for image I_m and word ω_k is given by

Over all the horizontal strips of height=30 pixels

Maximum character likelihood over all the detection windows

$$S(I_m, \omega_k) = \max_h \sum_{l=1}^{p} min\left(\max_j P(\omega_{kl} | hog_j), \tau \right)$$

Truncation parameter

For all the p characters

Inverted Index File

- Dataset is indexed for vocabulary words
- Initial retrievals are obtained using inverted index
- Does not guarantee spatial constraints on characters

Re-ranking Schemes

Spatial ordering (RSO)

- ψ_{total} : set of all the character pairs in query word
- ψ_{present} : set of spotted character pairs

$$S_{so}(I_m, \omega_k) = \frac{\psi_{\text{present}} \cap \psi_{\text{total}}}{\psi_{\text{total}}}$$

Spatial positioning (RSP)

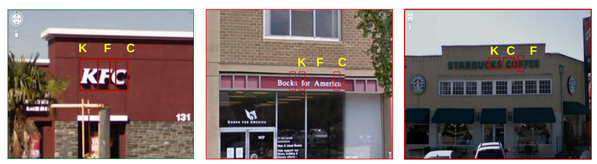
- The score is modified to

$$S_{sp}(I_m, \omega_k) = \sum_{l=1}^{p} min(\max_i U_i(\omega_{kl}), \tau) + \sum_{l=1}^{p-1} \max_{ij} V_{ij}(\omega_{kl}, \omega_{kl+1})$$

Likelihood of characters of query word

Joint probability of character pairs of query word

Example:



S_{so} : high, S_{sp}: high — S_{so} :high, S_{sp}: low — S_{so} :low, S_{sp}: low

Datasets

- Public scene text datasets not very ideal for evaluating retrieval
- Image **(IIIT-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.iiit.ac.in/projects/STR/

Quantitative results

Mean average precision (mAP) on Image datasets

Dataset	Char. spotting	RSO	RSP
SVT	0.17	0.46	**0.56**
ICDAR'11	0.24	0.58	**0.65**
IIIT-STR	0.22	0.36	**0.43**

Precision@20 on large video datasets

Dataset	Char. spotting	RSO	RSP
Sports	0.24	0.38	**0.43**
TV Series	0.39	0.57	**0.59**

Qualitative results

Text query: MOTEL

Text query: RESTAURANT

Text query: INDIAN

Text query: CENTRAL

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