Word Image Retrieval Using Bag of Visual Words

Ravi Shekhar and C.V. Jawahar

CVIT, IIIT-Hyderabad, INDIA

Motivation

Why Recognition free Retrieval?

- Robust OCRs are unavailable for many non-latin languages.
- These languages have rich heritage and there is a need for content level search.
- Word Spotting based methods are too slow for real time system.
- Most of the existing retrieval methods are memory intensive.
- Scalability is an immediate challenge.

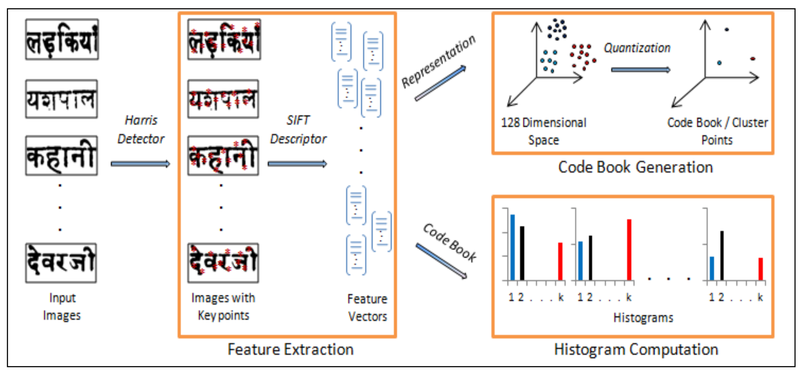
Why Bag of Visual Words?

- Bag of Words (BoW) representation is the most popular representation for text retrieval.
- BoW based efficient system like Lucene are publically available.
- Bag of Visual Words (BoVW) performs excellently for image and video retrieval.
- BoVW based system is flexible, powerful and scalable to Billions of images.

Bag of Visual Words

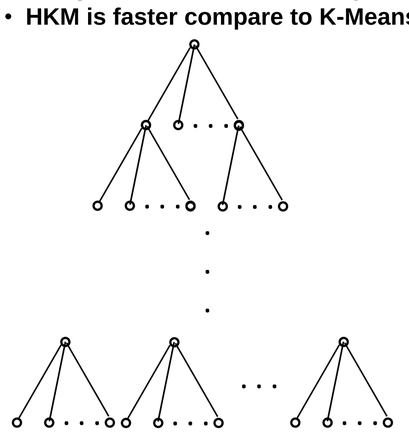
Bag of Visual Words Representation

- Represents word image as histogram of visual words
- Ignores the spatial relationships between visual words



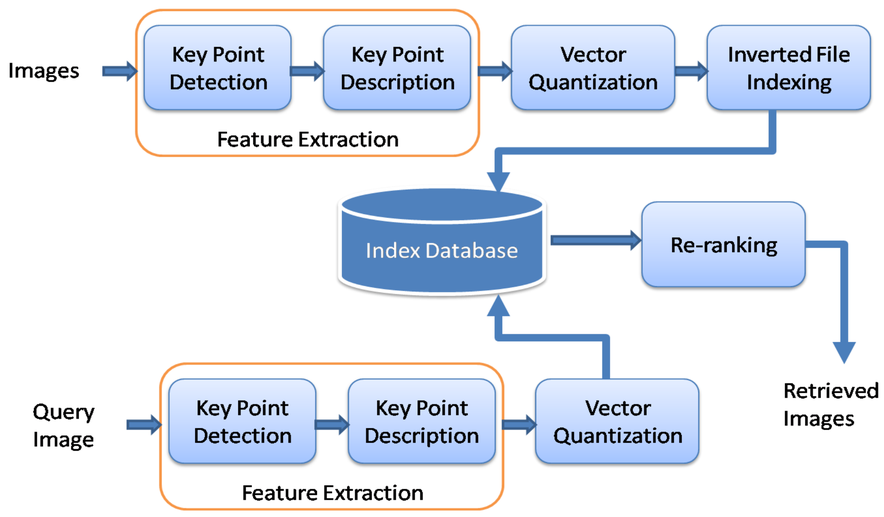
Code Book generation

- Using Hierarchical K-Means (HKM)
- HKM is faster compare to K-Means



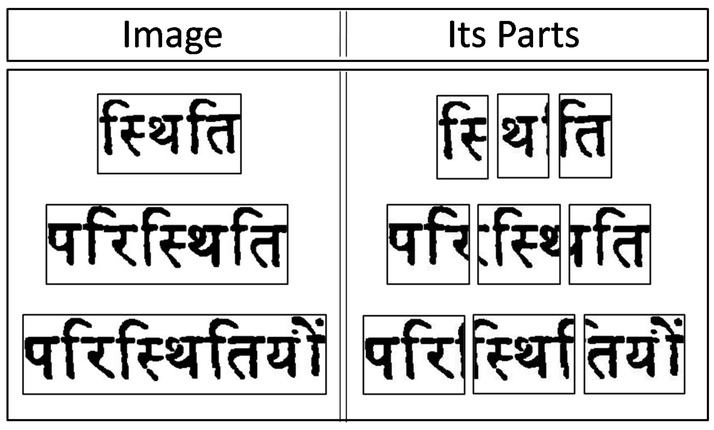
System Overview

Indexing and Retrieval



Spatial Verification

- To provide spatial order/structure of characters in word



Re-ranking

- Based on number of SIFT matches
- Higher the *Total Score*, better the match

$$Score(I_i, I_j) = \frac{\#Match\ Points}{\#\ SIFT\ in\ I_i + \#\ SIFT\ in\ I_j}$$

$$Total\ Score(I_i, I_j) = Score(I_i, I_j) + \frac{1}{3}\sum_{k=1}^{3} Score(I_i^k, I_j^k)$$

where, $Score(I_i, I_j)$: Score for entire image

$Score(I_i^k, I_j^k)$: Score for k^{th} part of the image

Retrieval Results

Sample Outputs

Query Image	AP	Retrieved Images (Rank)
# Occurrence: 86	0.677	47, 51, 56, 58
# Occurrence: 22	0.654	1, 2, 8, 10
# Occurrence: 23	0.875	2, 7, 13, 16
# Occurrence: 15	0.961	1, 2, 3, 9

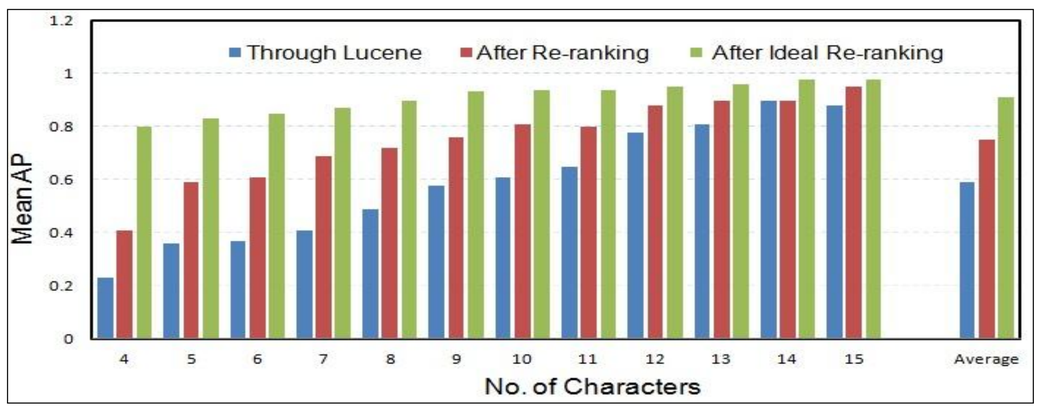
Performance Statistics

- Query set contains images which appear at least 15 times in the ground truth.

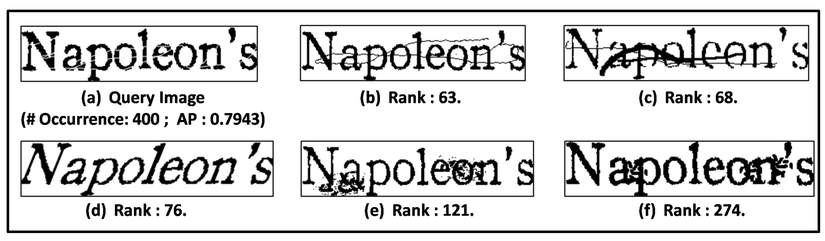
Language	# Books	# Images	# Query	Prec@10	mAP	Prec@10 after Spatial Verification	mAP after Spatial Verification
Hindi	4	112677	138	0.8437	0.6808	0.8770	0.7865
Hindi	32	1008138	138	0.8059	0.5894	0.8543	0.7062
Malayalam	6	108767	101	0.7668	0.6962	0.8581	0.8188
Telugu	5	131156	131	0.8507	0.6483	0.8830	0.7495
Bangla	3	124584	125	0.8498	0.7806	0.9182	0.8947

mAP Vs no. of Characters in Query

- Larger the length of query, Higher the MeanAP



Sample Output for Noisy Images



Retrieval Time

# Images	Retrieval Time	Index Size
25K	50ms	28MB
100K	209ms	130MB
0.5M	411ms	550MB
1M	700ms	1.2GB

Retrieval Time for 1M Words ~ 700msec

Occurrence of Images in Ground Truth Vs mAP

- Lesser the # Occurrence of Images in Ground Truth, Higher the mAP

# Occurrence of Images in Ground Truth	5	10	15	20	25	25+
mAP	0.8351	0.8293	0.8037	0.7942	0.7854	0.7813

Implementation Details

Keypoint Detection :
Harris Corner Detection – Invariant to rotation, scale and image noise

Keypoint Description :
Scale Invariant Feature Transform (SIFT) on third scale and zero degree

Indexing :
Inverted File Index using Lucene – Popular, reliable and open source search engine

Contributions

Language independent system : Demonstrated on 5 different languages

Scalable to huge datasets : Demonstrated on 1 Million images

Handles noisy document images : Demonstrated on dataset for which Commercial OCRs fail

Future Work

- Learning document-specific local descriptors
- Use of noisy OCR outputs along with BoVW
- Improve/remove the re-ranking methodology
- Mining character specific pattern to support text query

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad, India

http://cvit.iiit.ac.in/