Department of Electrical Engineering
Faculty of Engineering & Technology
Jamia Millia Islamia, New Delhi, India

INDICON 2015

IEEE Delhi Section

IEEE India Council

Paper ID: 1570186597

imPlag: Detecting Image Plagiarism Using Hierarchical Near Duplicate Retrieval

Siddharth Srivastava, Prerana Mukherjee, Brejesh Lall
Indian Institute of Technology, Delhi

Introduction

The key characteristic of image plagiarism is that it may involve the reproduction of the original image using an entirely different mode such as hand made sketches. Image Plagiarism can be posed as a superset of image copy detection problems.

Fig. 1. (a) Original Image (b) Plagiarised image (reproduction of the source image) (c) Copied image (considered as strong attack by copy detection algorithms but an expected case for Image Plagiarism)

Problems ?

- Detection of similar images – Huge Databases, Interactive Time
- Plagiarism brings in innovation

- Stitched from 3888 images
- One column/row pixel from each image

So knowing your limits is necessary

- Hence involves both Research and Engineering Challenges

Image Courtesy:Eirik Solheim
(Image has been used for demonstrating the extent of deformation possible in images)

KEY CONTRIBUTIONS

- Development of a hierarchical feature extraction and feature indexing technique.
- Evaluation of recent feature extraction techniques against simple, moderate and extreme deformations.
- Dataset construction for testing image plagiarism algorithms.

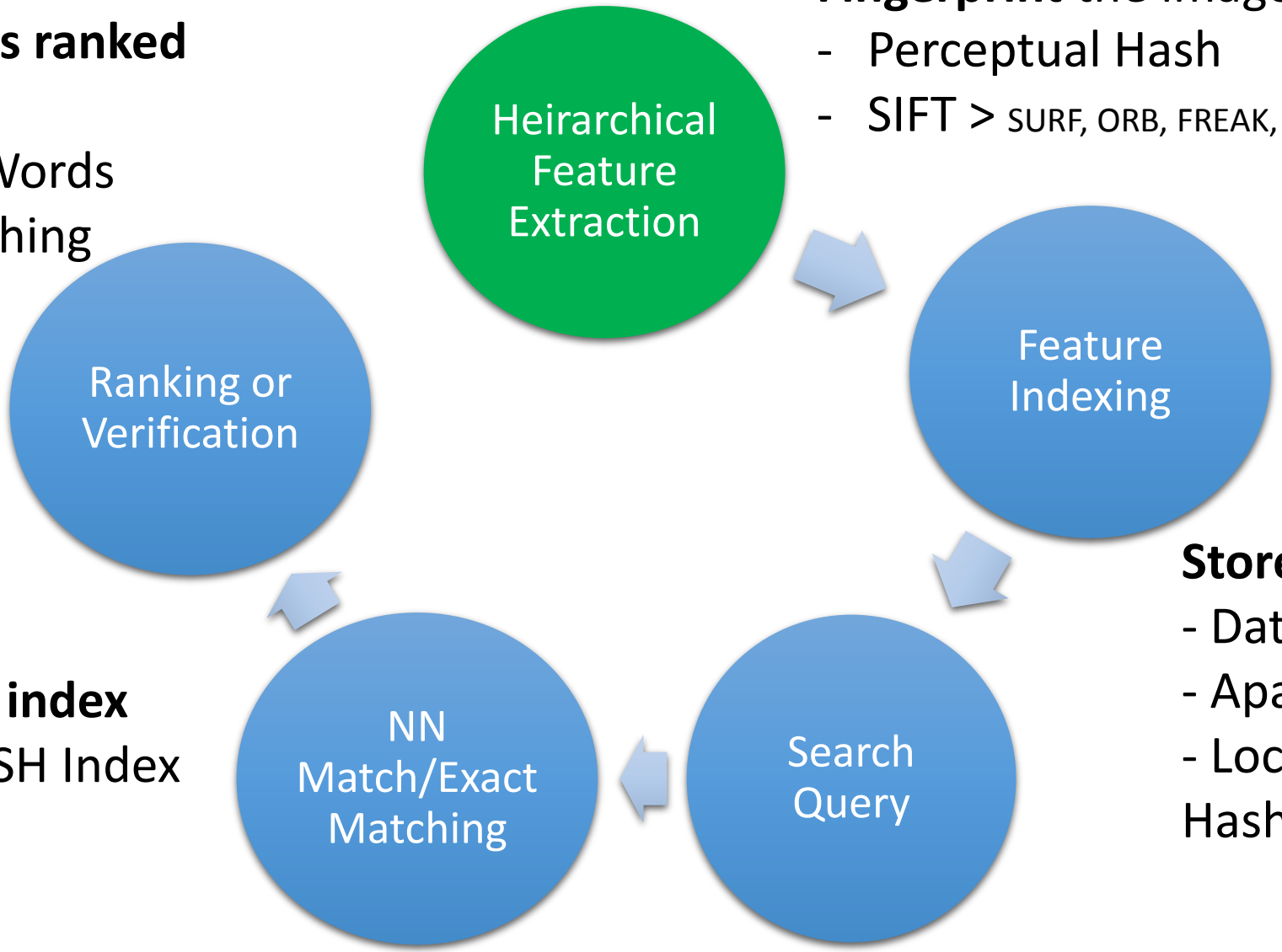
Dataset

- Natural Images – mountains, rivers, animals, birds etc.
- Actual scenario – too many images can be similar but might not be plagiarized (synthetically transformed)
- So for evaluation, dataset was created since detecting image plagiarism is not really only Content Based Image Retrieval
  - Search for images on Flickr, ukbench dataset
  - Find similar images using Google Reverse Image Search (Google doesn't index Flickr !!)
- Transformed Images – Affine, Grayscale, Color channel separation etc. (30 transformations)

Methodology

Relevant Results ranked at the top
- Bag of Visual Words Histogram matching

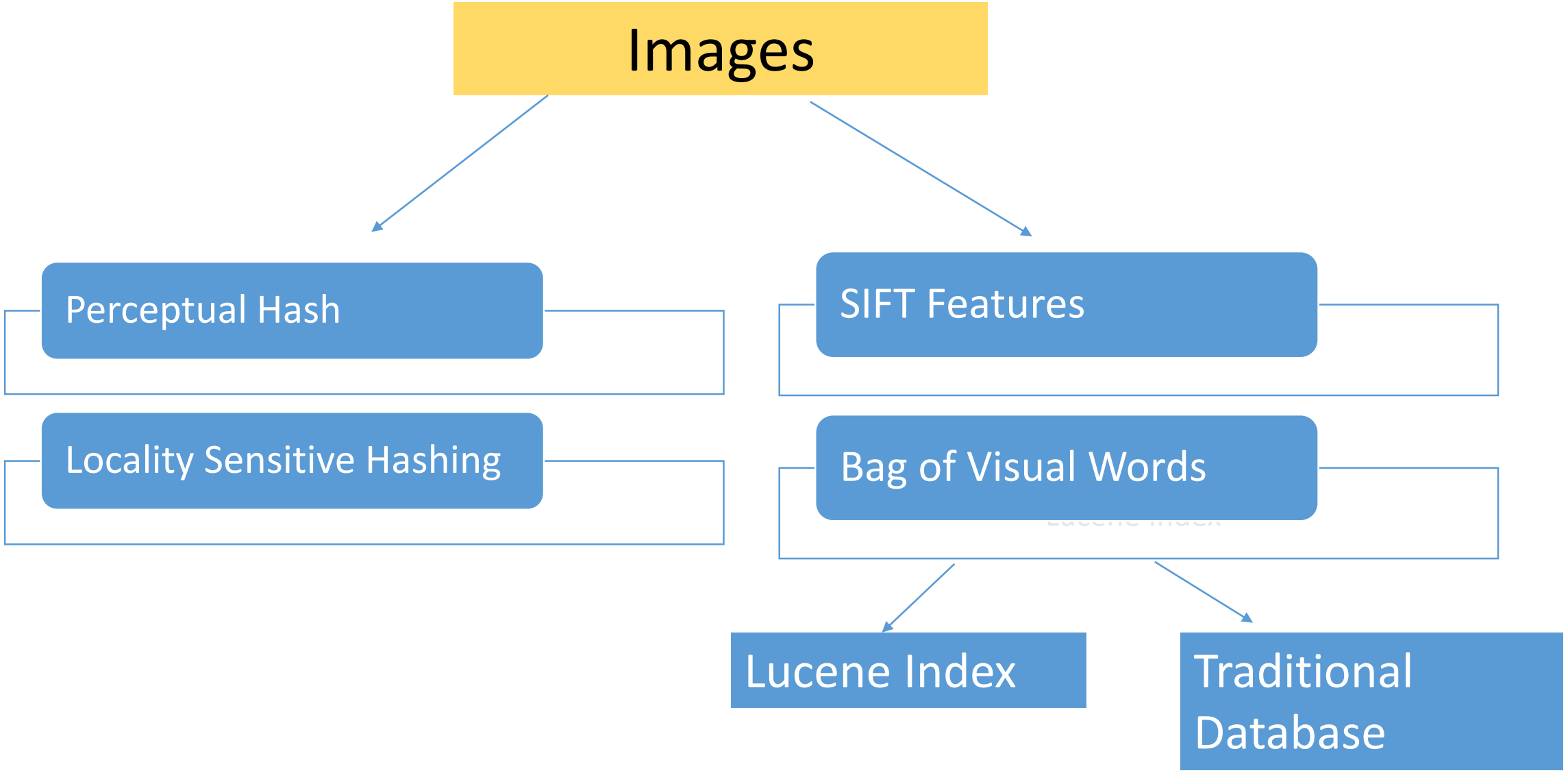
Fingerprint the image
- Perceptual Hash
- SIFT > SURF, ORB, FREAK, PCA-SIFT



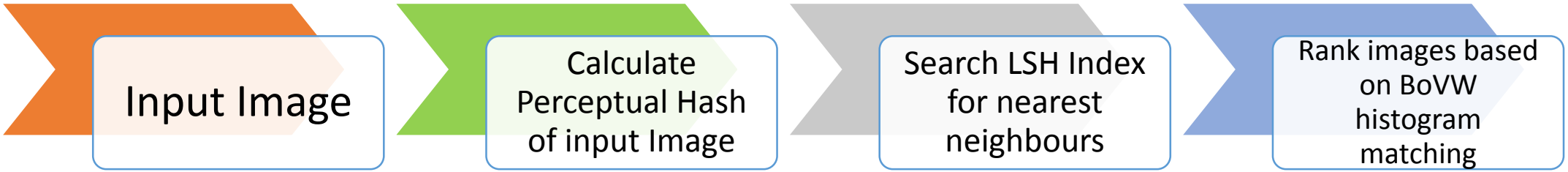
Store for retrieval
- Database
- Apache Lucene
- Locality Sensitive Hashing

Search the index
- Search LSH Index

Hierarchical Indexing

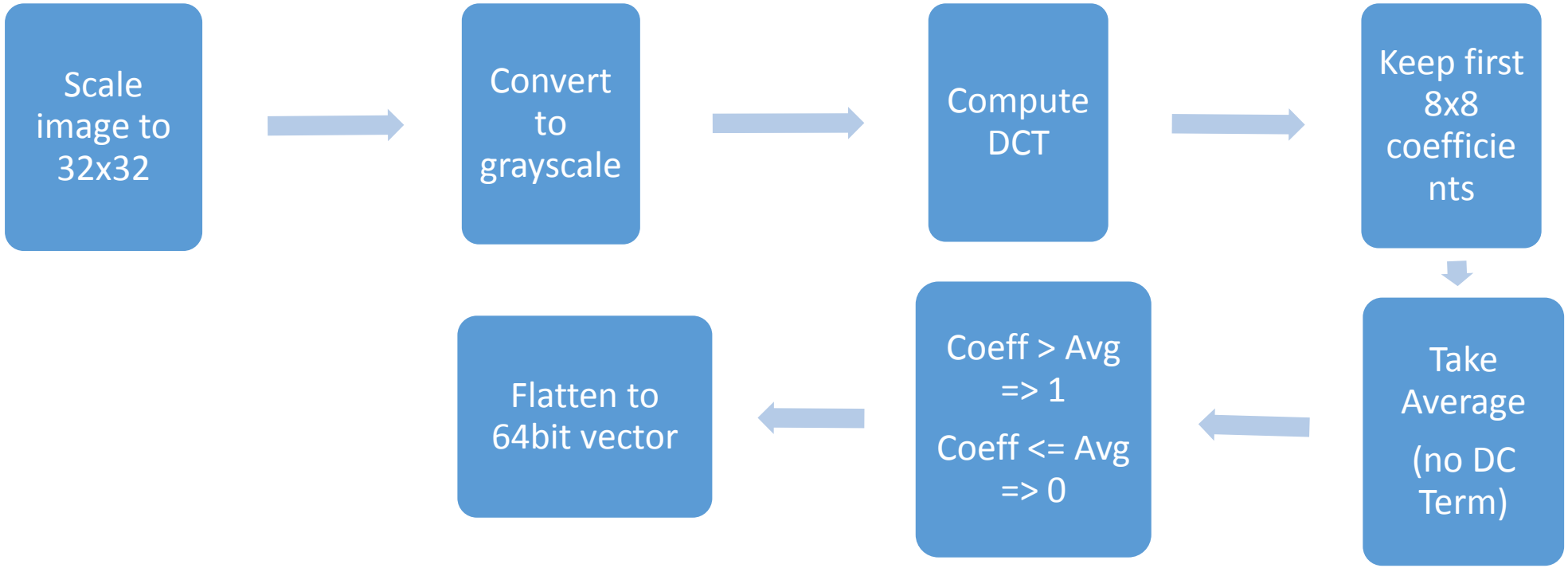


Layered Retrieval



Perceptual Hash

- Can be used for multimedia content (audio, video, images)
- Similar images have similar hash values



Bag of Visual Words

- SIFT features converted to Bag of Visual Words
- More efficient than direct keypoint matching
- Observations:
  - Large vocabulary size may increase false negatives
  - Small vocabulary size may increase false positives
- Though there is no definite pattern on what should the vocabulary size be

Results

- Accuracy: 81%
- Scalability

Conclusion

- We perform evaluations to choose best criteria and techniques for detecting image plagiarism.
- A method is proposed, consisting of perceptual hashing and SIFT with hierarchical approximate matching scheme.
- This scheme was able to maintain the tradeoff between time and accuracy.

References

- E. Chalom, E. Asa, and E. Biton, “Measuring image similarity: an overview of some useful applications,” Instrumentation & Measurement Magazine, IEEE, vol. 16, no. 1, pp. 24–28, 2013.
- C. Zauner, M. Steinebach, and E. Hermann, “Rihamark: perceptual image hash benchmarking,” in IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2011, pp. 78 800X–78 800X.
- V. Voronin, V. Frantc, V. Marchuk, and K. Egiazarian, “Fast texture and structure image reconstruction using the perceptual hash,” Image Processing: Algorithms and Systems XI, 2013.
- A. Kumar, A. Anand, A. Akella, A. Balachandran, V. Sekar, and S. Seshan, “Flexible multimedia content retrieval using infonames,” ACM SIGCOMM Computer Communication Review, vol. 41, no. 4, pp. 455–456, 2011
- A. Gionis, P. Indyk, R. Motwani et al., “Similarity search in high dimensions via hashing,” in VLDB, vol. 99, 1999, pp. 518–529.
- V. Christlein, C. Riess, J. Jordan, and E. Angelopoulou, “An evaluation of popular copy-move forgery detection approaches,” Information Forensics and Security, IEEE Transactions on, vol. 7, no. 6, pp. 1841– 1854, 2012.
- M. Lux and S. A. Chatzichristofis, “Lire: lucene image retrieval: an extensible java cbir library,” in Proceedings of the 16th ACM international conference on Multimedia. ACM, 2008, pp. 1085–1088.

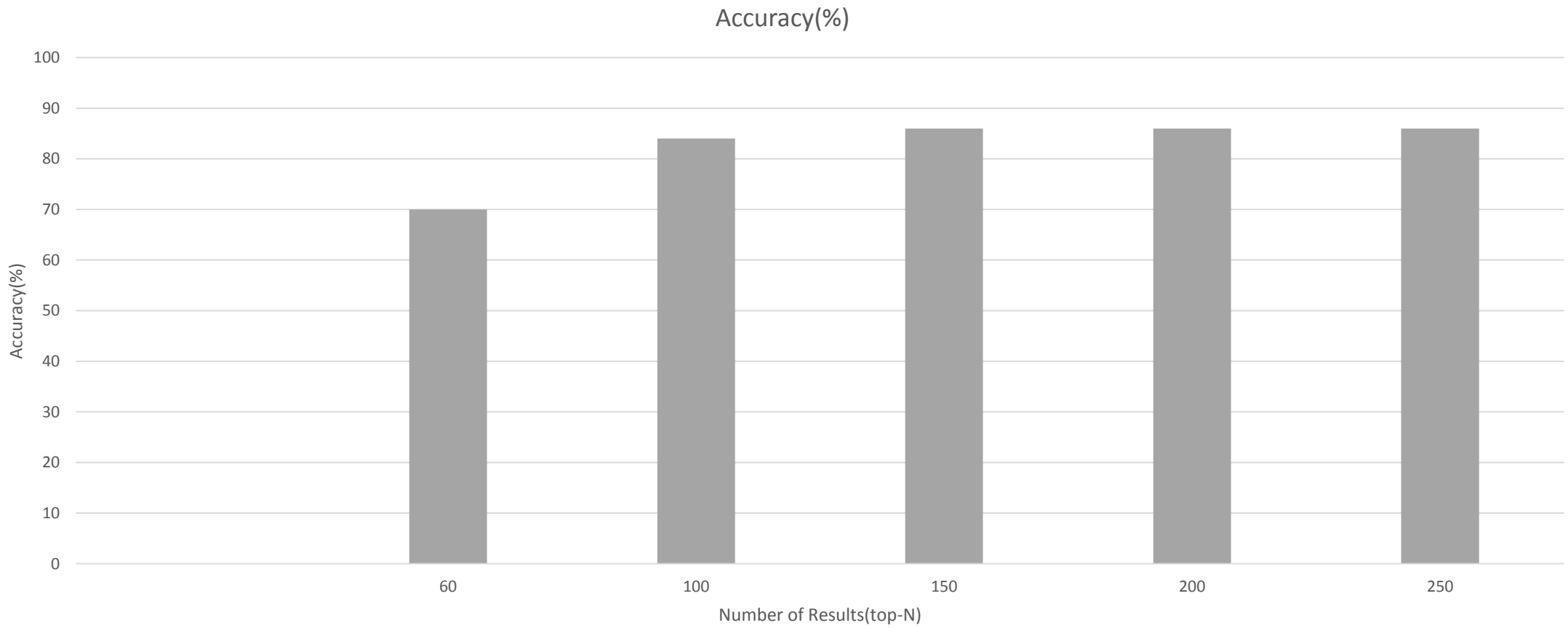
THANKYOU

Appendix: Dataset Images

Perceptually Similar ?

Appendix: Nature is not always greenish

Appendix: Accuracy



Appendix: Results

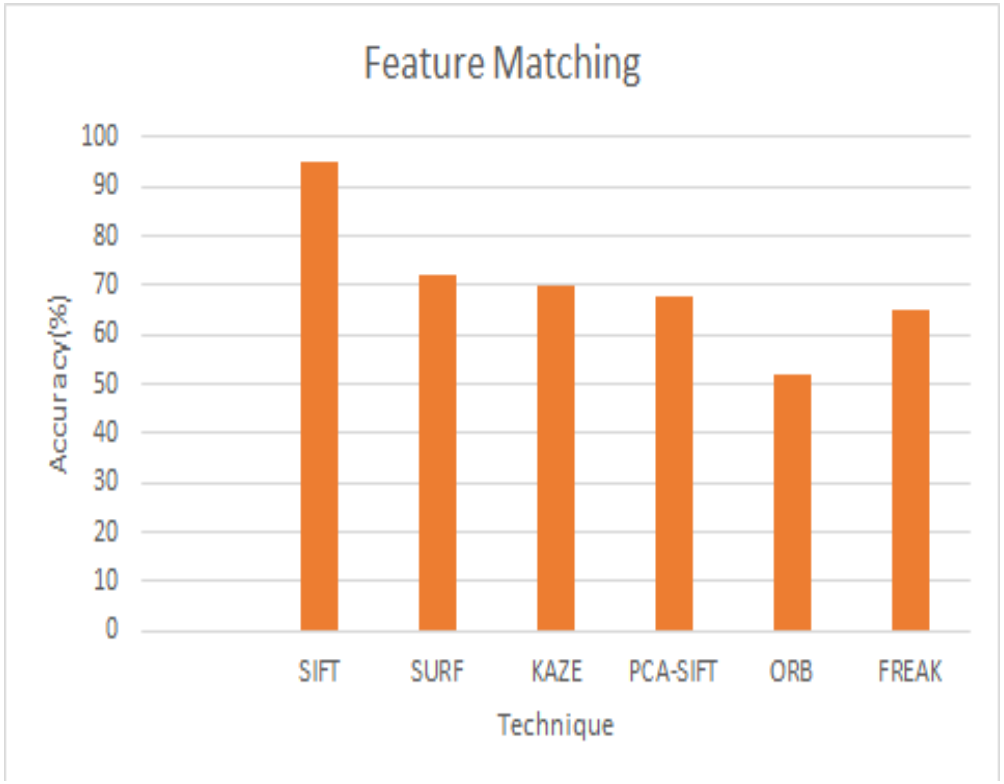


Fig 2. Comparison of Feature matching techniques

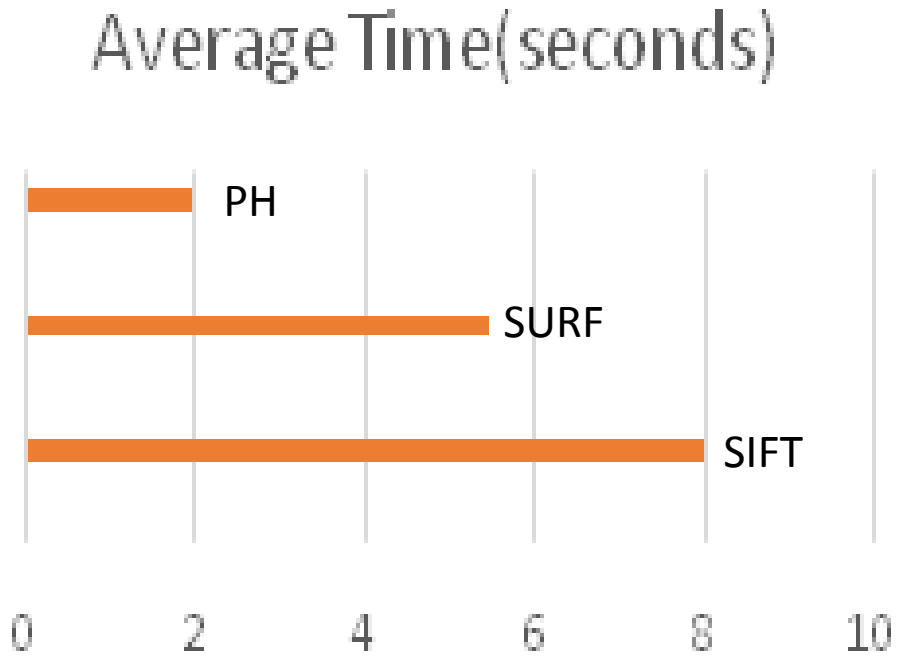


Fig 3. Average time taken by SIFT, SURF and Perceptual Hash

Appendix: Results

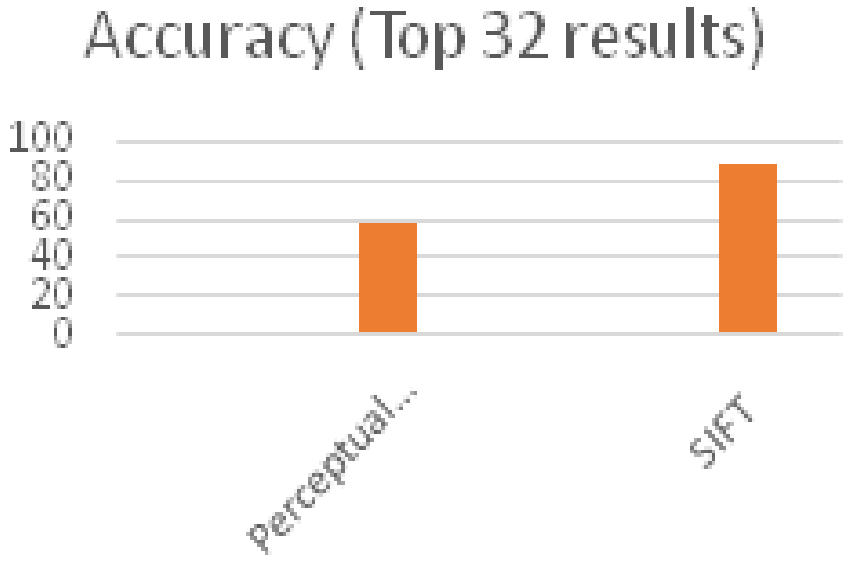


Fig 4. Comparison of ranked retrieval

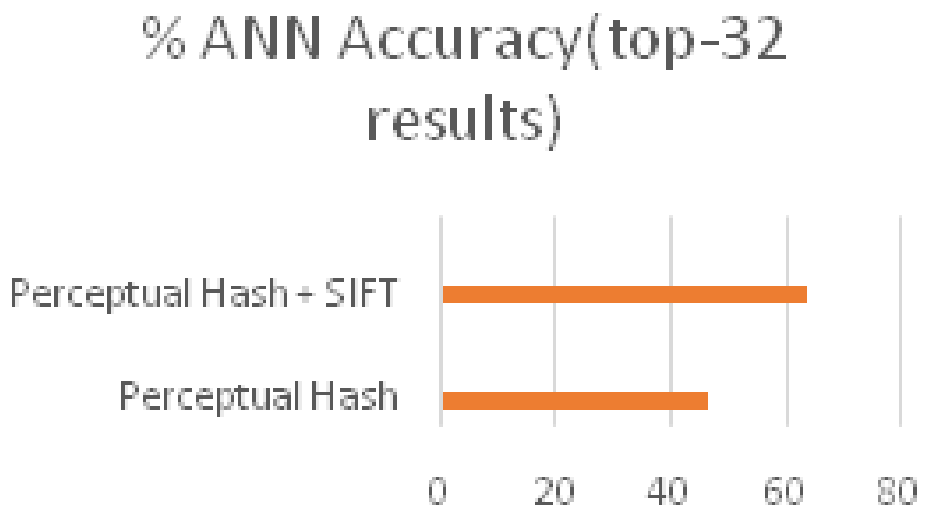


Fig 5. Ranked V/s Non Ranked Retrieval

Appendix: Results

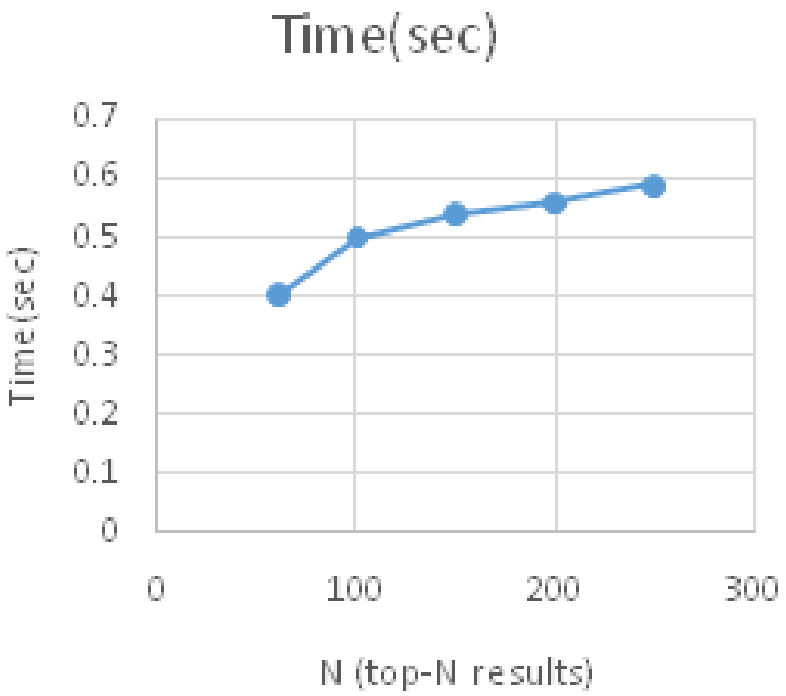


Fig 6. Time vs Number of results

Fig 7. Time vs Number of Images in the dataset

Appendix: Results

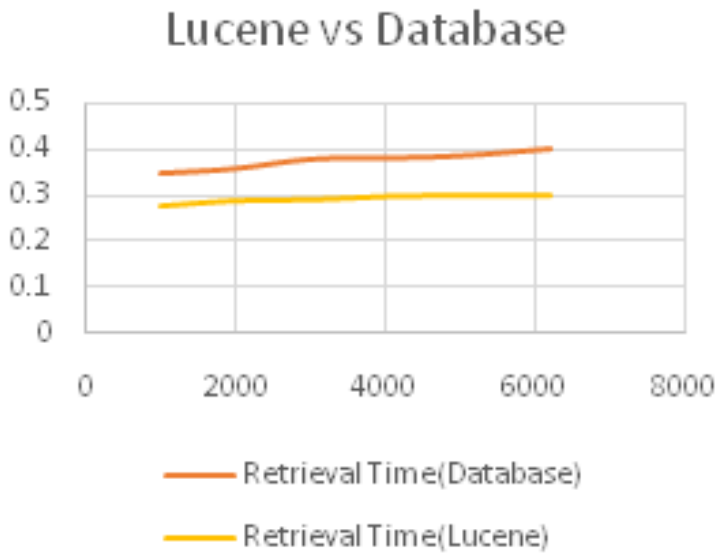


Fig 8. Lucene v/s Database Retrieval time

Locality Sensitive Hashing

- Similar features hashed to same hash values
- Parameters
  - No of bits (k)
  - No of tables (l)
  - Maximum Bucket capacity (usually unlimited)
- Empirical Analysis needed for determining parameters as per the dataset
- varying number of bits, varies bucket size (small hash, more collisions and vice versa)

Lucene

- Very efficient in document indexing and retrieval
- Bag of Visual words histograms are indexed
- Allows for random access of documents
- Histograms are fetched from Lucene index and ranked (Filtering)