SPATIAL RERANK-BASED BAG-OF-WORDS MODEL

Presenters: Bao Truong, Thuyen Phan
Instructor: Tiep Nguyen

June 27, 2015

VNUHCM - University of Science - Faculty of Information Technology

OUTLINE

1. Introduction
2. Background & Related Work
3. Method
    - Overview of Image Retrieval System
    - Preparation
        - Feature extraction
        - Codebook building
        - Quantization
    - Query
        - TF-IDF weighting
        - Query expansion (with geometric verification)
4. Experimental results
5. Conclusion

INTRODUCTION

MOTIVATION

Alongside the intensive growth of social networks such as Facebook or Twitter, the information that users want to share is not only text but also other complex types, especially images.

In January 2009, there are over 3 billion photos on Flickr.

In 2012, 250 million photos are uploaded to Facebook everyday.

MOTIVATION

Nowadays, the tendency to search not only by text but also some special features of other complex types such as images is currently one of the most popular concerns.

- Many text retrieval techniques are applied to image retrieval since search by text and by images share many common characteristics.
- There are still numerous differences between text and image retrieval in many criterion.

BACKGROUND & RELATED WORK

BAG-OF-WORDS

The early Bag-of-Words model is proposed by A.Zisserman and J. Sivic in 2003[1]. The main idea of their work is to represent an object as a set of invariant features, similar to words in a text document.

The model was then improved by J. Philbin et al[2][3] with some advanced techniques such as soft assignments and geometric verification which was applied by the authors' work.

---

[1] Sivic; J.; Zisserman; A.;
"Video Google: a text retrieval approach to object matching in videos;
" Computer Vision; 2003. Proceedings. Ninth IEEE International Conference on ; vol.;
no.; pp.1470; 1477 vol.2; 13-16 Oct. 2003.

[2] Philbin; J.; Chum; O.; Isard; M.; Sivic; J.; Zisserman; A.;
"Object retrieval with large vocabularies and fast spatial matching;
" Computer Vision and Pattern Recognition; 2007. CVPR '07. IEEE Conference on ; vol.;
no.; pp.1; 8; 17-22 June 2007.

[3] Philbin; J.; Chum; O.; Isard; M.; Sivic; J.; Zisserman; A.;
"Lost in quantization: Improving particular object retrieval in large scale image databases;
" Computer Vision and Pattern Recognition; 2008. CVPR 2008. IEEE Conference on ;
vol.; no.; pp.1; 8; 23-28 June 2008.

BAG-OF-WORDS

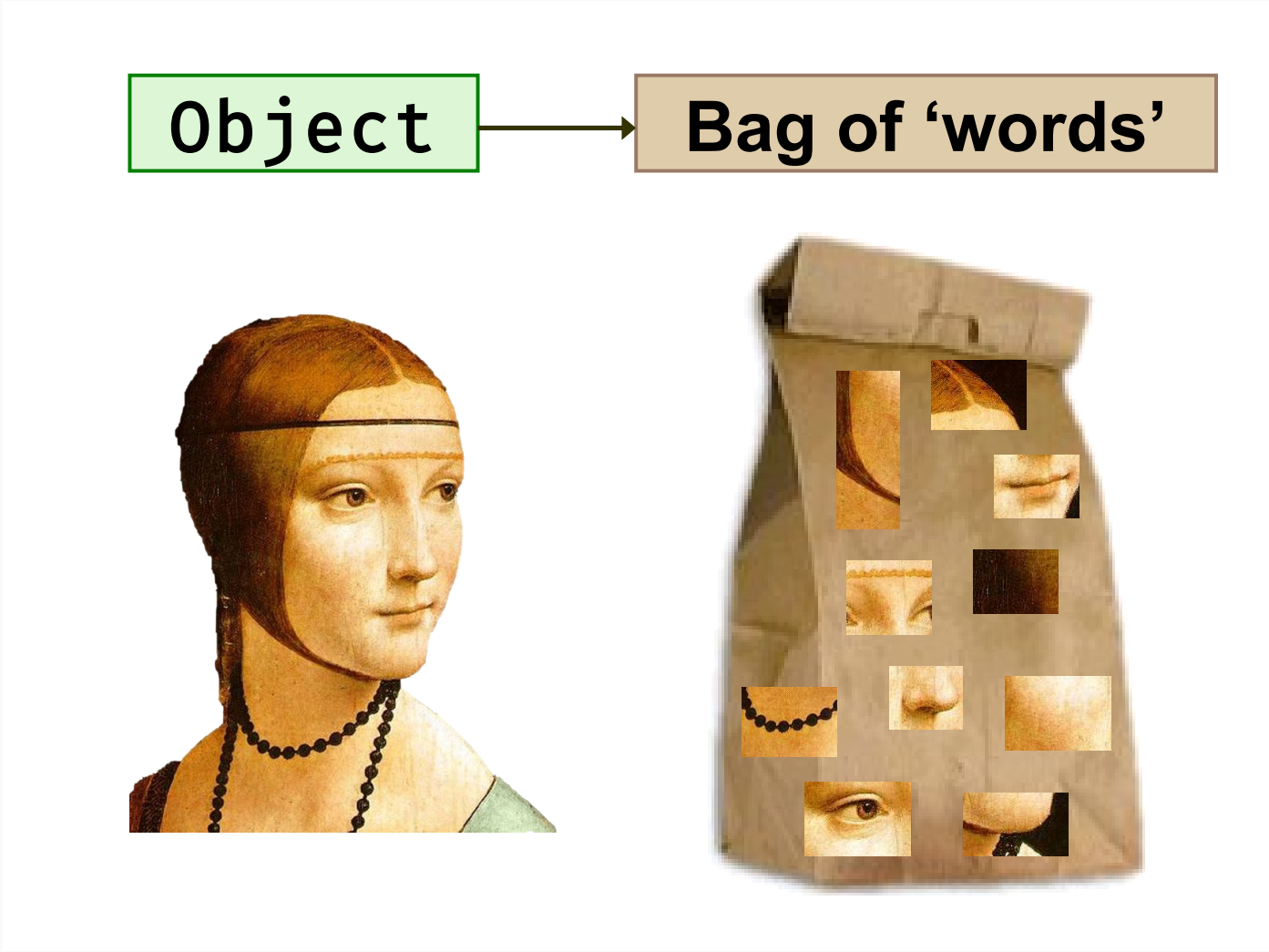


Figure: Bag-of-Words model[4]

[4] Slide credit: Rob Fergus; Classical Methods for Object Recognition.

REVIEW OF TEXT RETRIEVAL

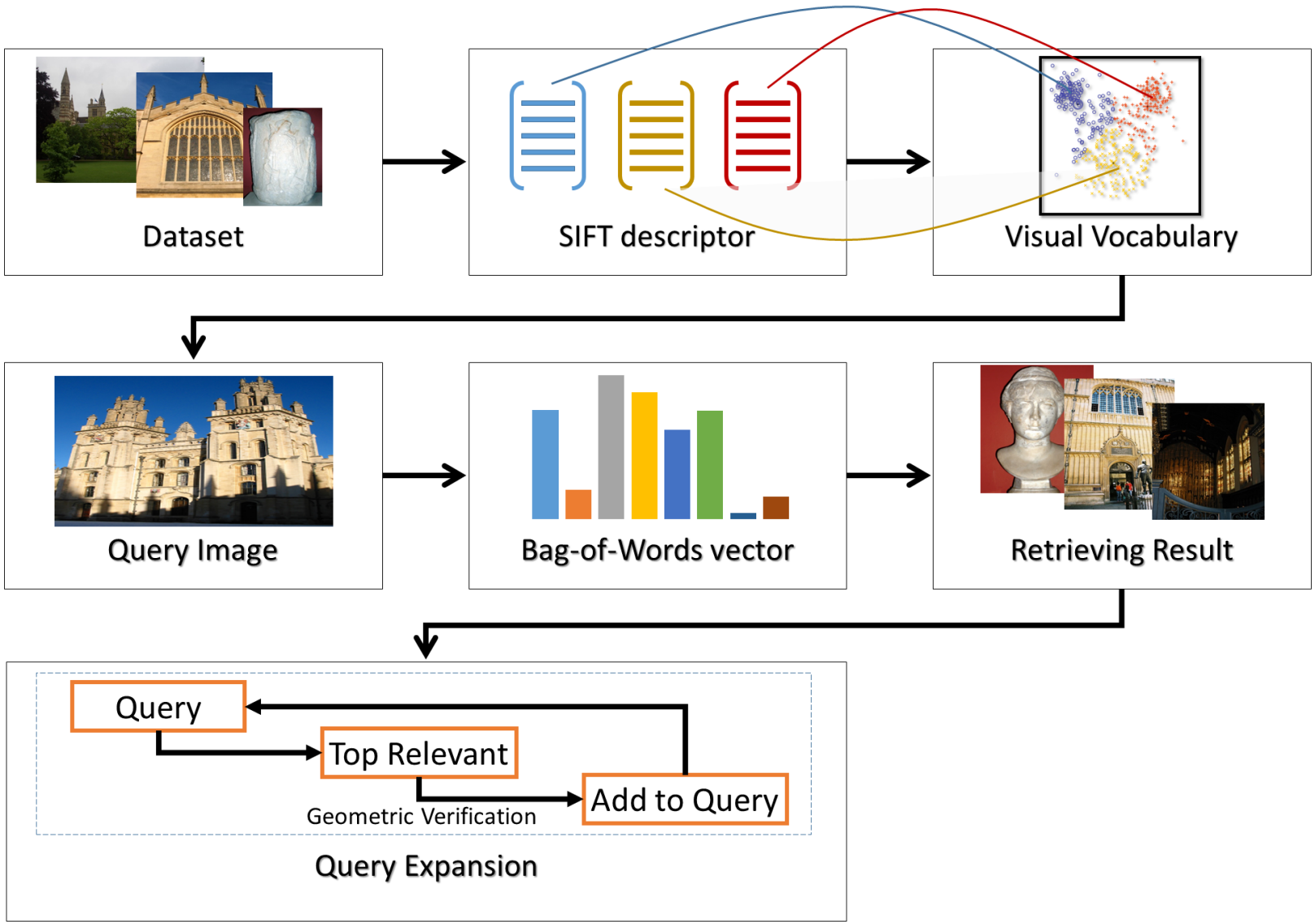
Common steps in text retrieval systems[5]:

- Parse documents into words
- Represent words by their sterms
- Reject common words by using stop list
- Represent each document by a vector of words' frequency (BoW)
- Retrieve a document by computing its BoW vector and returning documents with the closest vector

[5] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press; ISBN: 020139829; 1999.

METHOD

OVERVIEW OF IMAGE RETRIEVAL SYSTEM



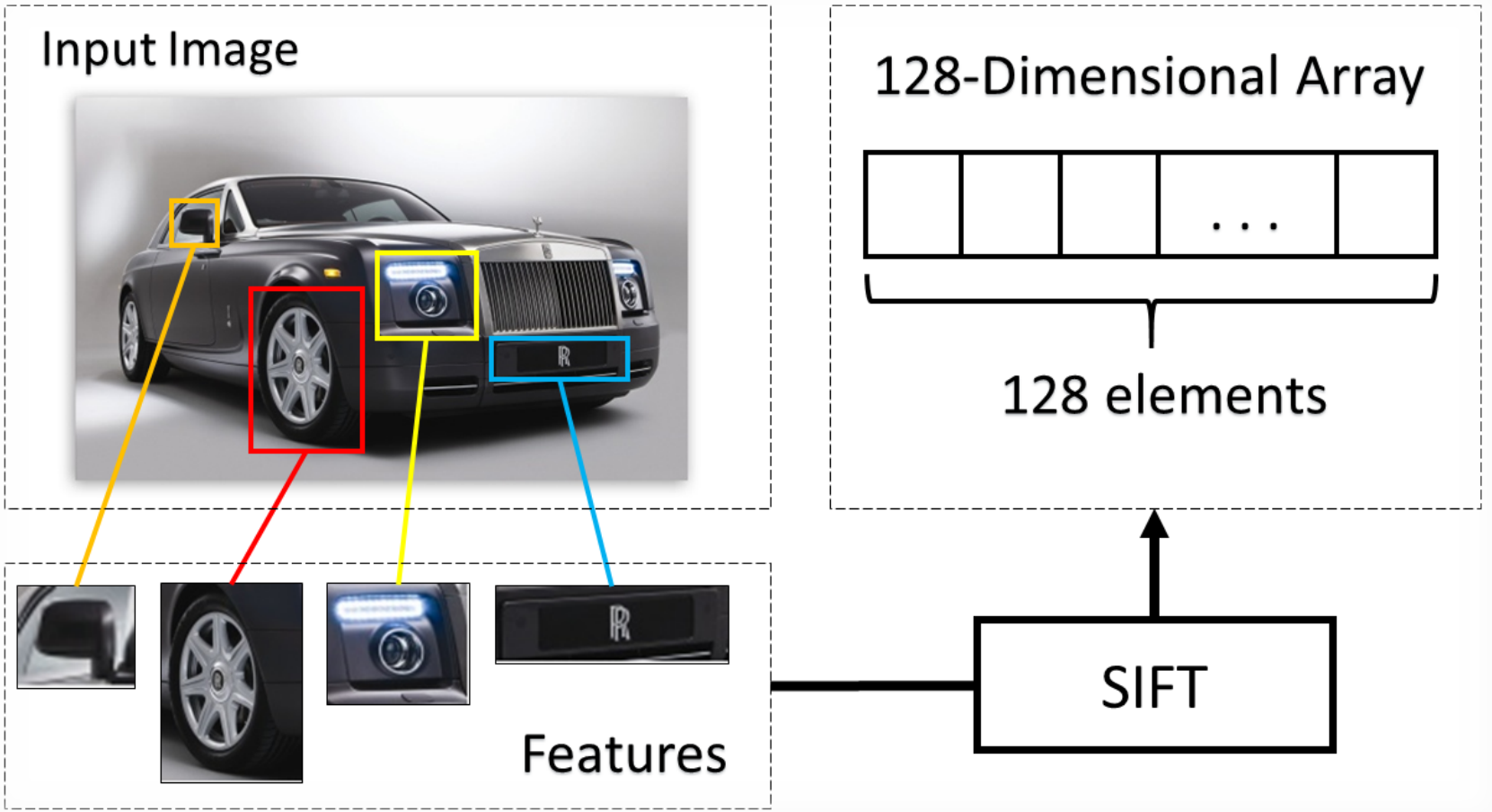
PREPARATION: FEATURE EXTRACTION

Apply ”compute descriptors” tool from University of Surrey with the following parameters:

-hesaff -sift -noangle

hesaff: Scale and Affine invariant interest point detector
sift: use Scale Invariant Feature Transform (SIFT) descriptor
noangle: no angle estimation

PREPARATION: FEATURE EXTRACTION



PREPARATION: CODEBOOK BUILDING

Use **approximate k-Means** to cluster all features to **1M** clusters

- Use FASTANN and FASTCLUSTER libraries[6][7]
- Run with **50** iterations

Each image is presented by a **1M-dimensional** vector

---

[6] Muja M. and Lowe D.;
Fast approximate nearest neighbours with automatic algorithm configuration;
Proceedings VISAPP 2009.

[7] Philbin J. Chum O. Isard M. Sivic J. and Zisserman A.;
Object retrieval with large vocabularies and fast spatial matching;
Proceedings CVPR 2007.

PREPARATION: CODEBOOK BUILDING

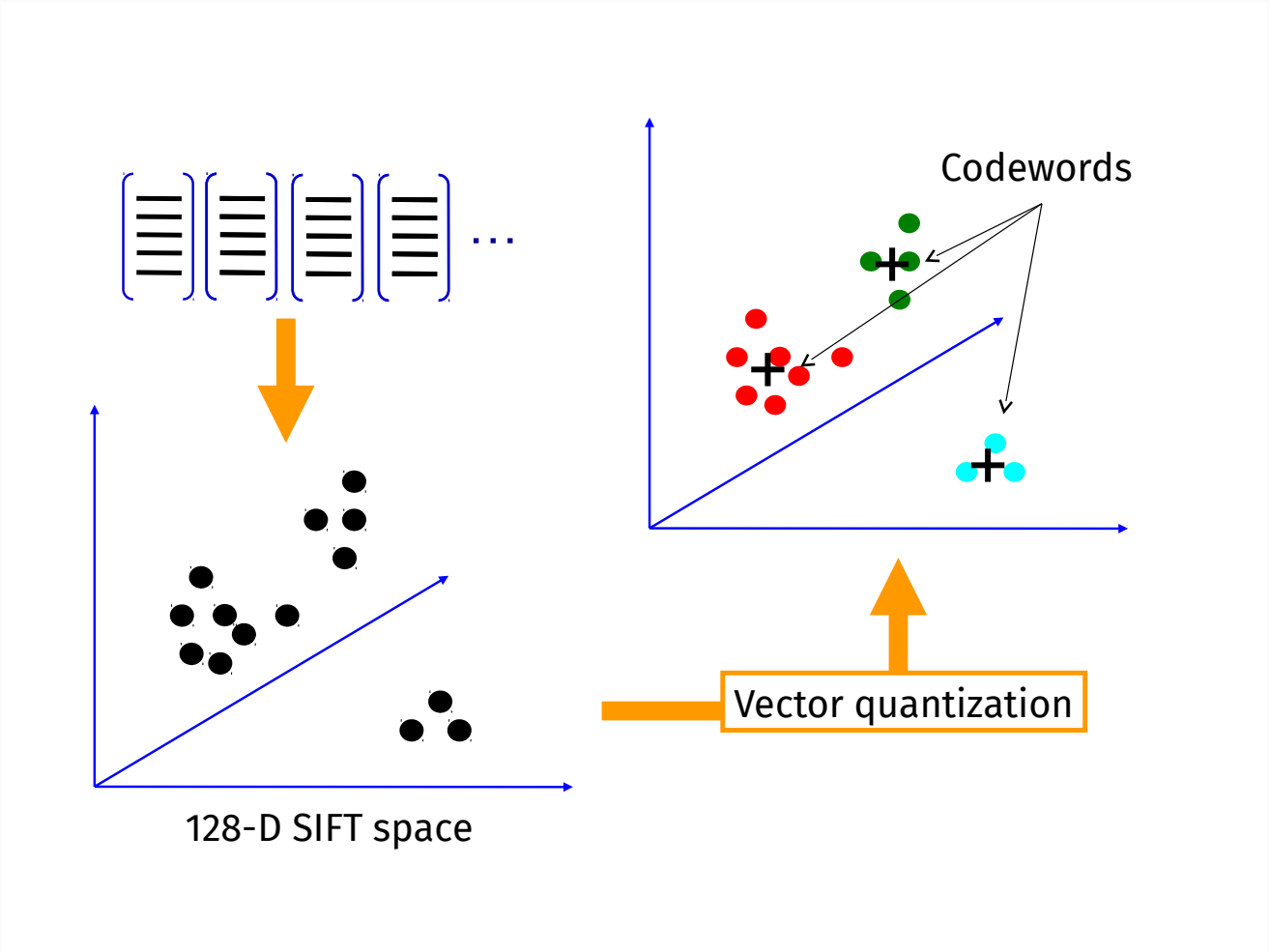


Figure: Clustering[8]

[8] Slide credit: Josef Sivic.

PREPARATION: QUANTIZATION

Soft assignment: Each **128-dimensional** feature vector is reduced to **3-dimensional** by looking for its **3 nearest visual words**
Each of the **nearest visual words** is assigned with weight:

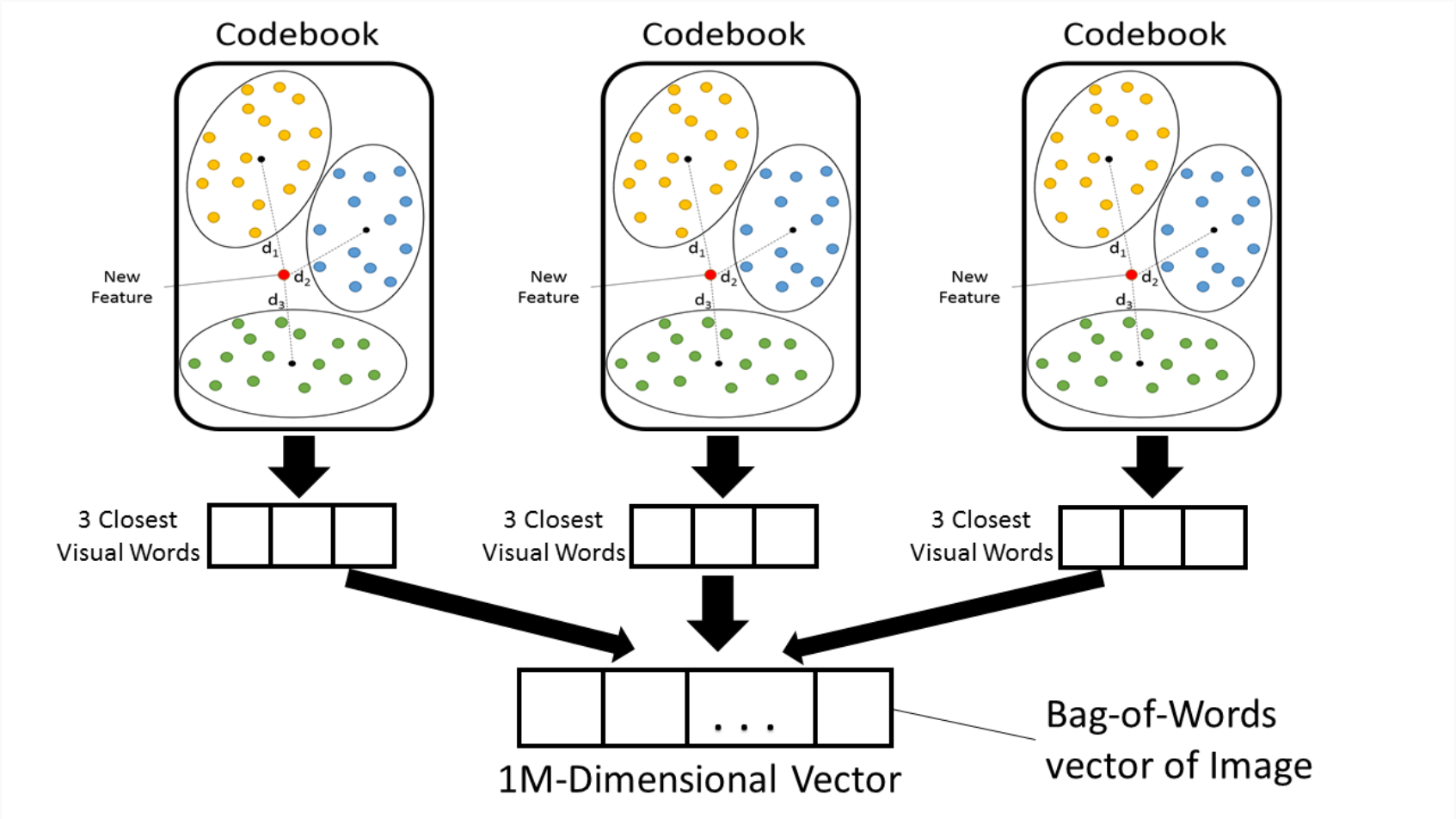
$$\text{weight} = \exp(-\frac{d^2}{2\delta^2})$$

d = distance from feature vector to cluster centroid

$\delta^2 = 6250$

All weights are added to their corresponding visual word in the **1M-dimensional** representation of the image

PREPARATION: QUANTIZATION



QUERY: TF-IDF WEIGHTING

Similar with text retrieval

- **raw term frequency:** raw $tf_{i,j}$ = weight of visual word i in image j
- **document frequency:** df_i = # of images that visual word i appears
- **raw inverse document frequency:** raw $idf_i = |D|/df_i$

QUERY: TF-IDF WEIGHTING

Observation

- The **more time** a visual word occurs, the **less important** it is
- A visual word is **more discriminate** if it occurs in **fewer** images

Hence, it is necessary to **normalize** the **values of TF-IDF**

$$\mathrm{tf}_{i,j} = \frac{\text{raw } \mathrm{tf}_{i,j}}{\sum_k \text{raw } \mathrm{tf}_{k,j}} \text{ (for all visual words k in image j)}$$

$$\mathrm{idf}_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

QUERY: TF-IDF WEIGHTING

Weight of visual word i in image j is therefore: $\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$
The tf-idf weight is used to **compute similarity** between an image d_i and a query q

$$s_{d_i,q} = \vec{\text{tfidf}}_i \cdot \vec{\text{tfidf}}_q = \sum_{j=1}^{|T|} \text{tfidf}_{i,j} \times \text{tfidf}_{q,j}$$

By **sorting** list of images based on their **similarity score** with a query, we achieve the **raw ranked list** which is used for the **Query Expansion** step

QUERY: QUERY EXPANSION

Apply **geometric verification** between the query image **Q** and each top-ranked image **A**:

1. (x, y) is a **matched pair** of features if $x \in Q$, $y \in A$, x and y are assigned to the same visual word
2. Randomly choose 4 pairs of features to build the **homography matrix**. A matched pair (x, y) is called **inliner** if apply the computed homography matrix on feature x produces feature y. Repeat 100 times to find the matrix that produces the **largest** number of inliners. These inliers are the **verified** visual words
3. **TF-IDF weight** of the **verified** visual words are added to query

Run this process for all top-ranked images. The added TF-IDF weight are averaged before running the query again

EXPERIMENTAL RESULTS

DATASET

To prove our hypotheses, the authors test both the BoW systems with and without spatial rerank on the Oxford Building 5K Dataset [9] which consists of 5,062 images belong to 11 different classes and 55 queries along with their ground-truth.

[9] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

DATASET

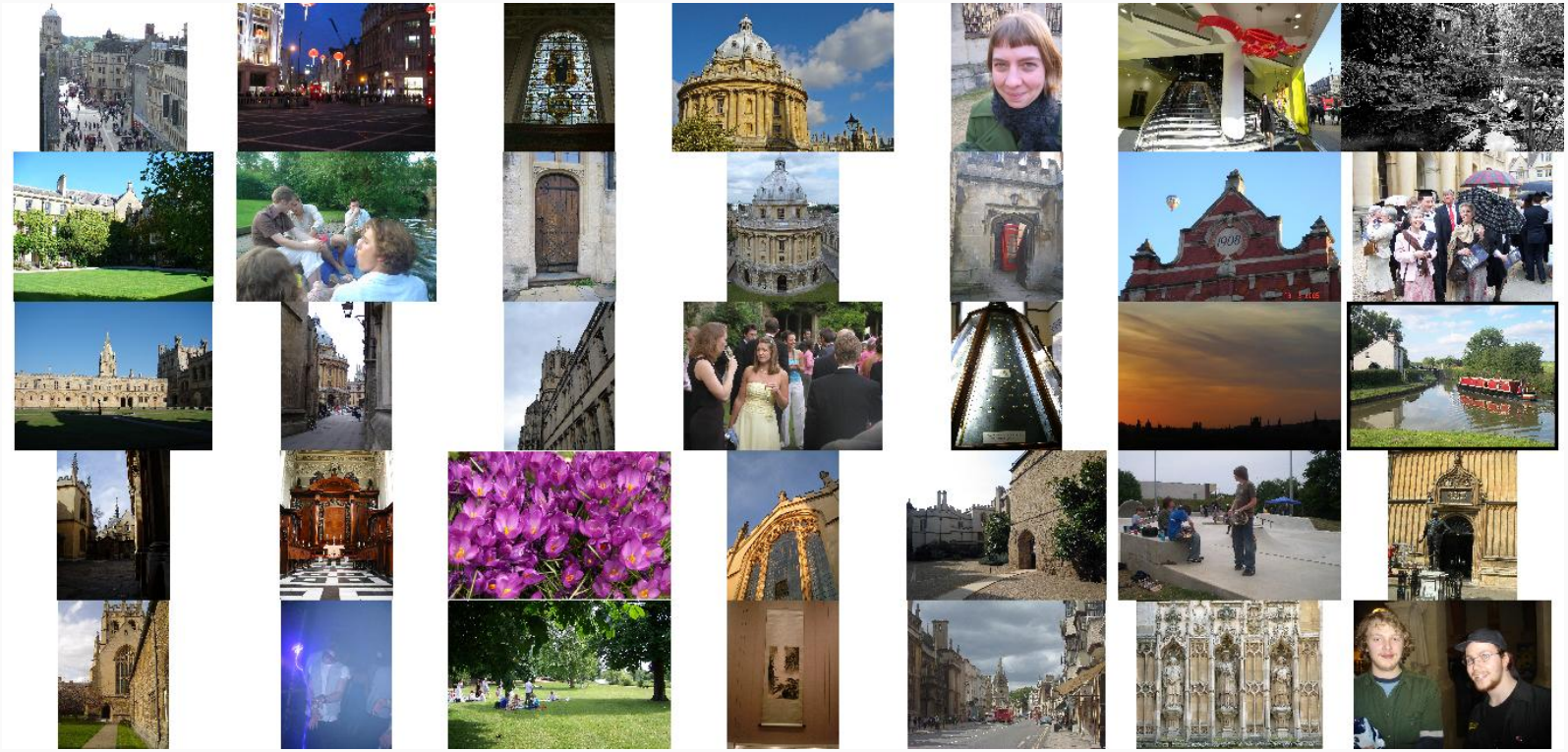


Figure: Oxford Building 5K Dataset

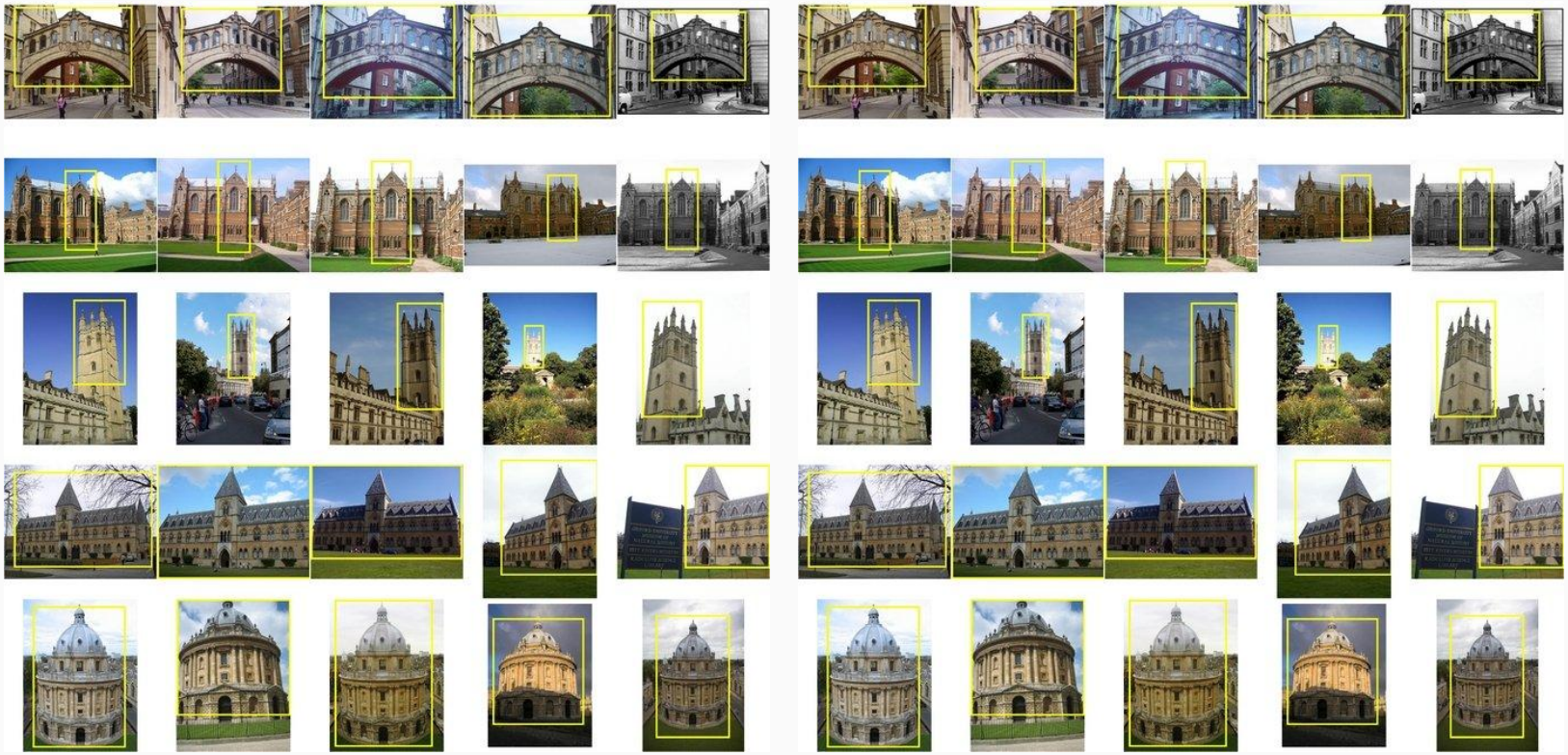


Figure: Oxford Building 5K Dataset - Queries

EXPERIMENTAL RESULT

The performance of our system is evaluated by mean average precision (mAP).

Our experiment shows that spatial rerank has significant impact on the retrieval quality of BoW model, an increase from 0.676 to 0.741 in term of mAP.

COMPARISON CHART

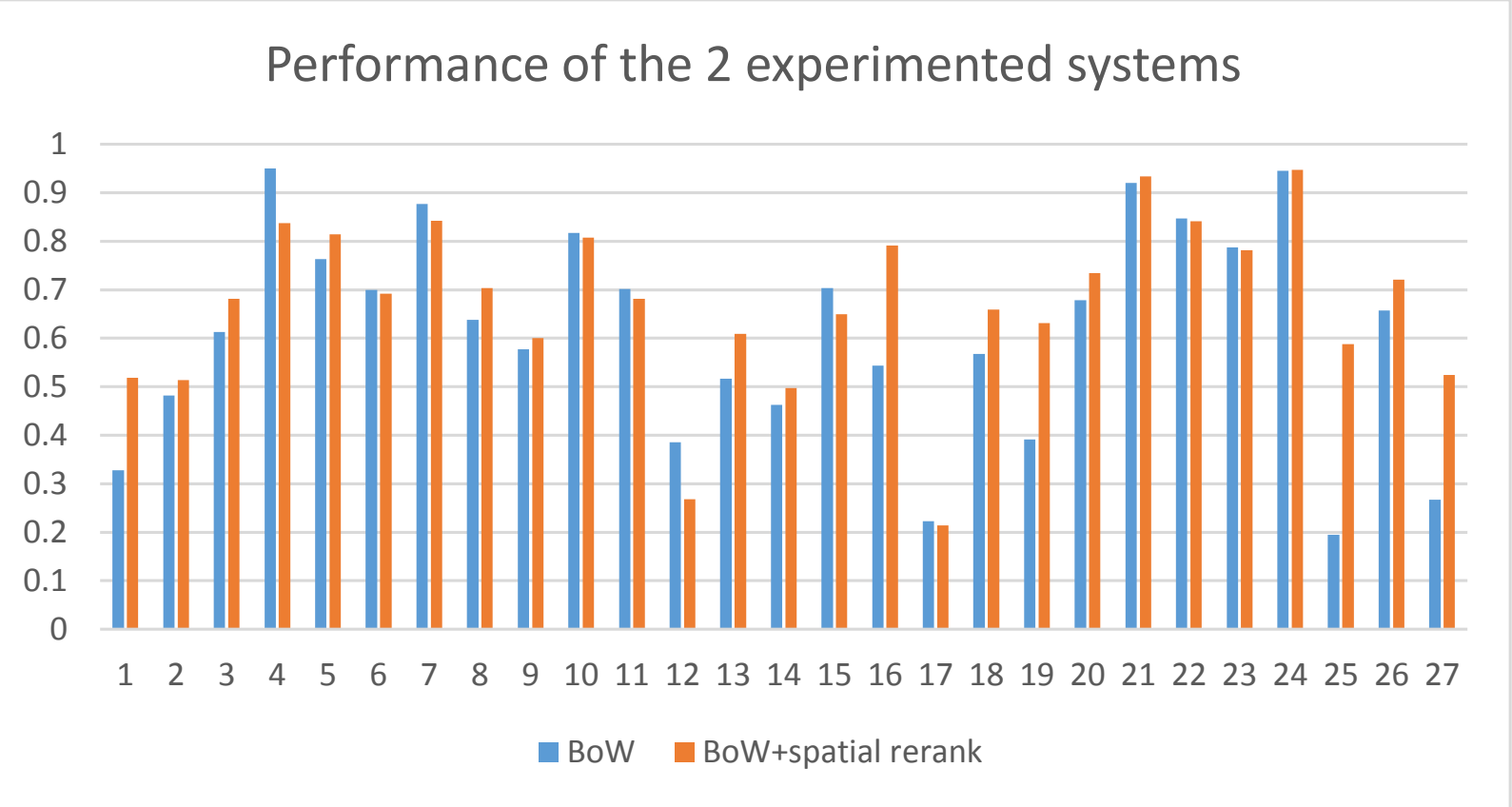


Figure: APs of queries 1 - 27

COMPARISON CHART

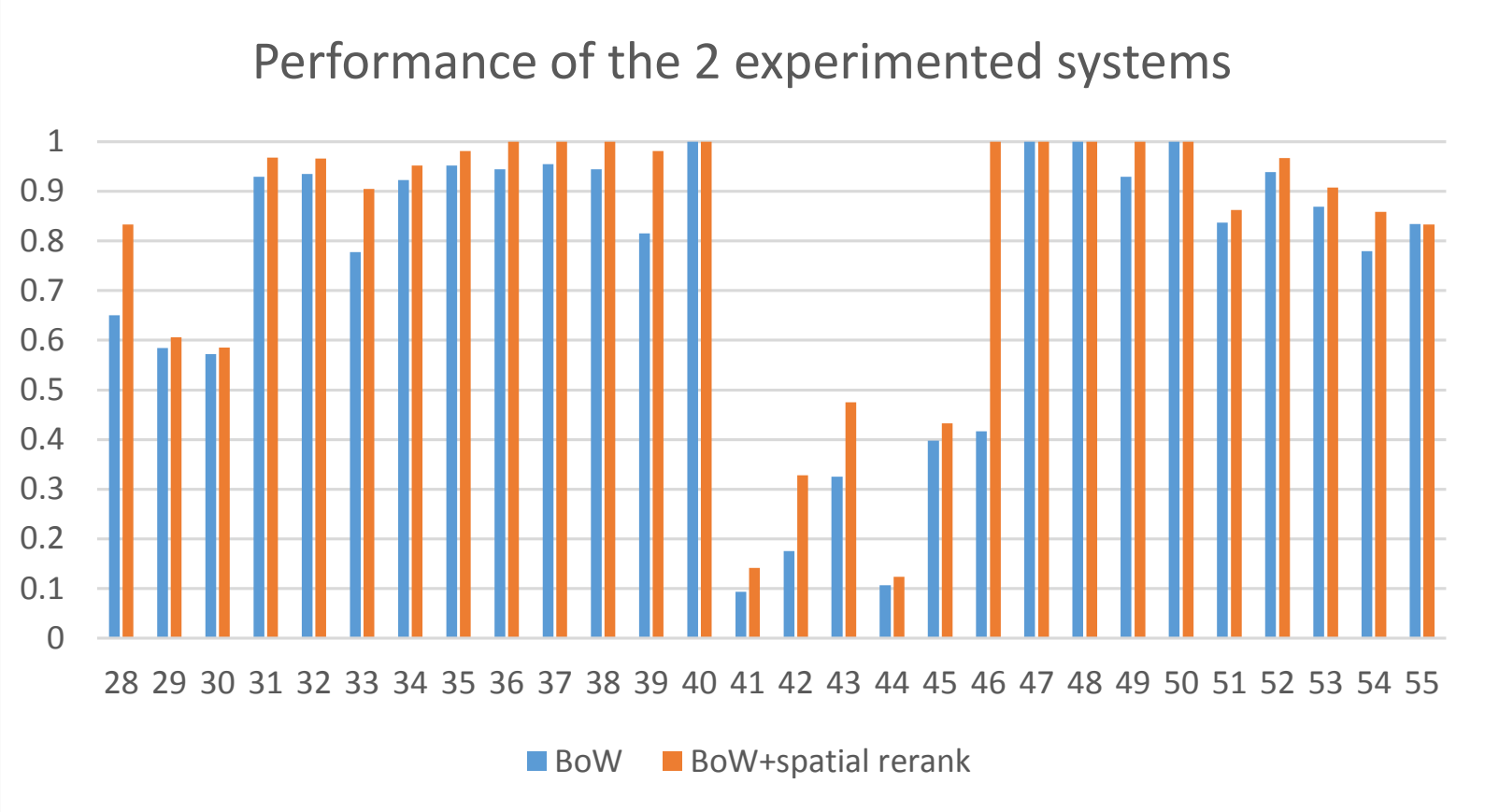


Figure: APs of queries 28 - 55

CONCLUSION

CONCLUSION

Through our experiments, it is proved that spatial rerank significantly boosts the performance of BoW model. This is a very potential result to further improve the performance of many image retrieval systems.

In the future, we plan to keep upgrading our system to operate on other datasets which are larger in size and also more variant in term of content. Our final goal is to deploy our system for real-time usage with limited computer resources.

Questions?