Brown University

Regions in Retrievals

Using memorability regions in image retrieval pipelines

Vibhu Ramani (vibhu_ramani@brown.edu)
5/15/2013

Table of Contents

Introduction

The growing volume of image collections presents the challenge of retrieving relevant matches with increased index sizes. Many modern image retrieval pipelines return a decent rate of matches for known structures and landmarks. Another aspect is to establish ranking of images based on defined criteria. This project specifically addresses these two challenges by conducting experiments to check if human memorability of images results in improved performance & reduced indexed sizes.

Background

With the ease of access to cameras capable of capturing high quality images and instant uploading for public viewing, efficiently searching a needle in the growing haystack becomes a priority.

Indexing Pipeline

Image retrieval has always been an active area of research not just from the computer vision perspective, but also from the database management view. Most of the image retrieval pipelines are classified as either text based or visual based. Both these methods involve extracting features from images and indexing them. This dataset is then used to conduct efficient retrieval based on near matching features. Potential features include annotated textual words, color histograms or a set of features like Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Compressed Histogram of Gradients (CHOG), Gradient Location and Oriented Histogram (GLOH) extract features on a point based on interest or chosen at random. Points of interest are those that may be easily located across multiple images for the same object. Corners are examples of such points.

The extracted features of an image are then mapped onto a reverse index. Additionally, the features may be clustered using a clustering technique such as K-means. An index may then be built on these cluster centroids, mapping to the images containing them. It is referred to as a reverse index due to the inverted structure of reference.

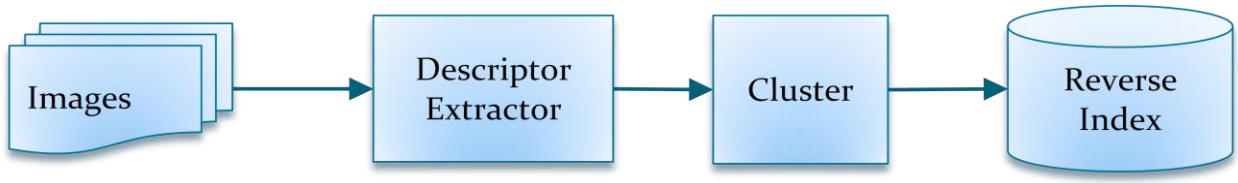


Figure 1 Indexing Pipeline

Retrieval Pipeline

Given an image as input for finding similar matches, the feature extraction process begins by extracting features from the input and proceeds by mapping each feature to a set of matching centroids. Each of these centroids returns an image list and all such lists are merged to yield a histogram. The histogram is then used to rank the images from the database. Additionally, the retrieval pipeline may include an additional step termed ‘query expansion’ which appends additional features to the extracted feature list.

The size of index is one of the major factors contributing to retrieval performance. Another optimization can be to use a smarter image ranking algorithm. The project focuses on these two aspects to bump up space efficiency while maintaining/improving search quality.

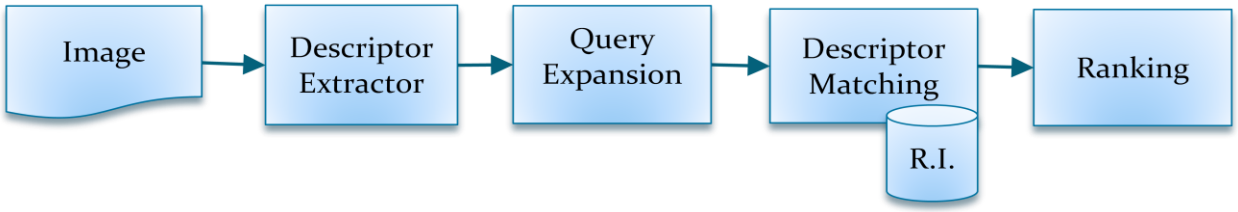


Figure 2 Retrieval Pipeline

Memorability of images using regions

Human brain has a tremendous capacity to store visual data, but its recall degrades over time. Some images stay longer in memory in comparison to others and there have been recent works to show us that memorability of an image is an intrinsic property. We can perform some estimations of memorability using state of the art features and machine learning models. One very recent work deeply explored if memorability of images could be determined based on the types of regions in an image. This work suggests that we might be able to determine regions of interest in an image based on memorability.

From a given image, patches of image are randomly sampled. From each of these patches, features are extracted using independent models of gradient, color, texture, saliency, shape and semantic. A dictionary of 1024 patch types is created by K-means clustering. Each image is described as a set of boolean values indicating whether a region type exists or not. Memorability is calculated as the ratio of number of times humans were able to recall a previously shown image to the total number of times the image was shown to them. An SVM-Rank algorithm is used to build a model which helps to calculate the memorability score of an image. The parameters learned in this model are used to define the estimated memorability score of regions and an overall memorability score of the image.

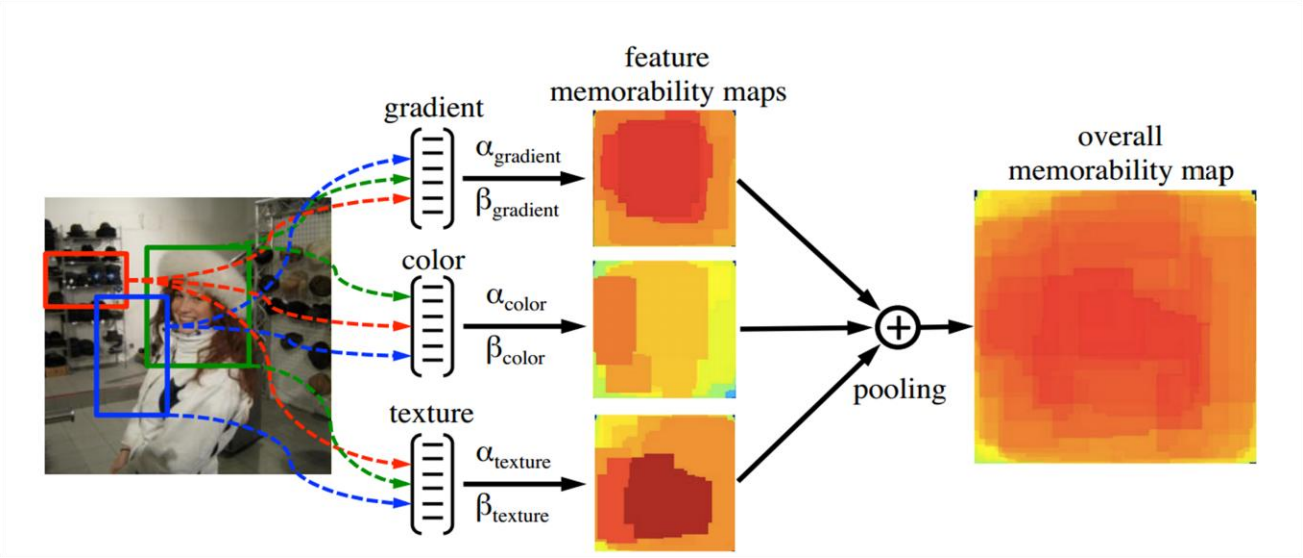


Figure 3 Memorability region detection

Project Idea

The current methods of selecting points of interest are very local and do not take into account the human perspective of the image, and maybe be picking non-informative regions. Since the idea of identifying memorability regions in an image is relatively new, this project aims to check if this concept can be incorporated here or in the ranking step of Indexing & Retrieval pipeline.

Experiments

Datasets

The image retrieval pipeline was set on the following two datasets:

Paris Dataset

The dataset consists of 6412 images collected from Flickr by searching for particular Paris landmarks.

Oxford Buildings Dataset

The dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks.
These collections have been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated.

For creating the memorability regions model the following dataset was used:

Scene Understanding (SUN) Database subset

The dataset consists of 2222 images from the SUN dataset. The images are fully annotated with segmented object regions and randomly sampled from different scene categories. The images are cropped and resized to 256*256 and a memorability score corresponding to each image is provided. The memorability score is defined as the percentage of correct detections by participants in their study.

Experiment 1

Establish a relation between average number of features per image and retrieval accuracy

Expected Result

Using memorability scores of regions instead of random points of interest should increase the retrieval accuracy while reducing the average number of descriptors per image

Baseline establishment

To proceed with experimentation, a baseline has to be established using a base model. An indexing and retrieval pipeline was setup using the PARIS dataset. The stepwise procedure is as follows:

- Fetch a set of SIFT features from every image in PARIS dataset
- Create a reverse index on these features

- For a query image, retrieve all relevant images using the reverse index
- Compare this result set against the ground truth which is a list of all good images that match this query
- Determine the accuracy of the result set
- Average the accuracy across all 50 input queries

We reduce average descriptors per image by:

- randomly selecting points of interest to compute SIFT features
- increasing edge threshold for determining points of interest to compute SIFT features

Experiment run

After the baseline has been established, the points of interest are ranked based on memorability score regions. The SIFT features are calculated on these points in the order of their ranking.

Observations

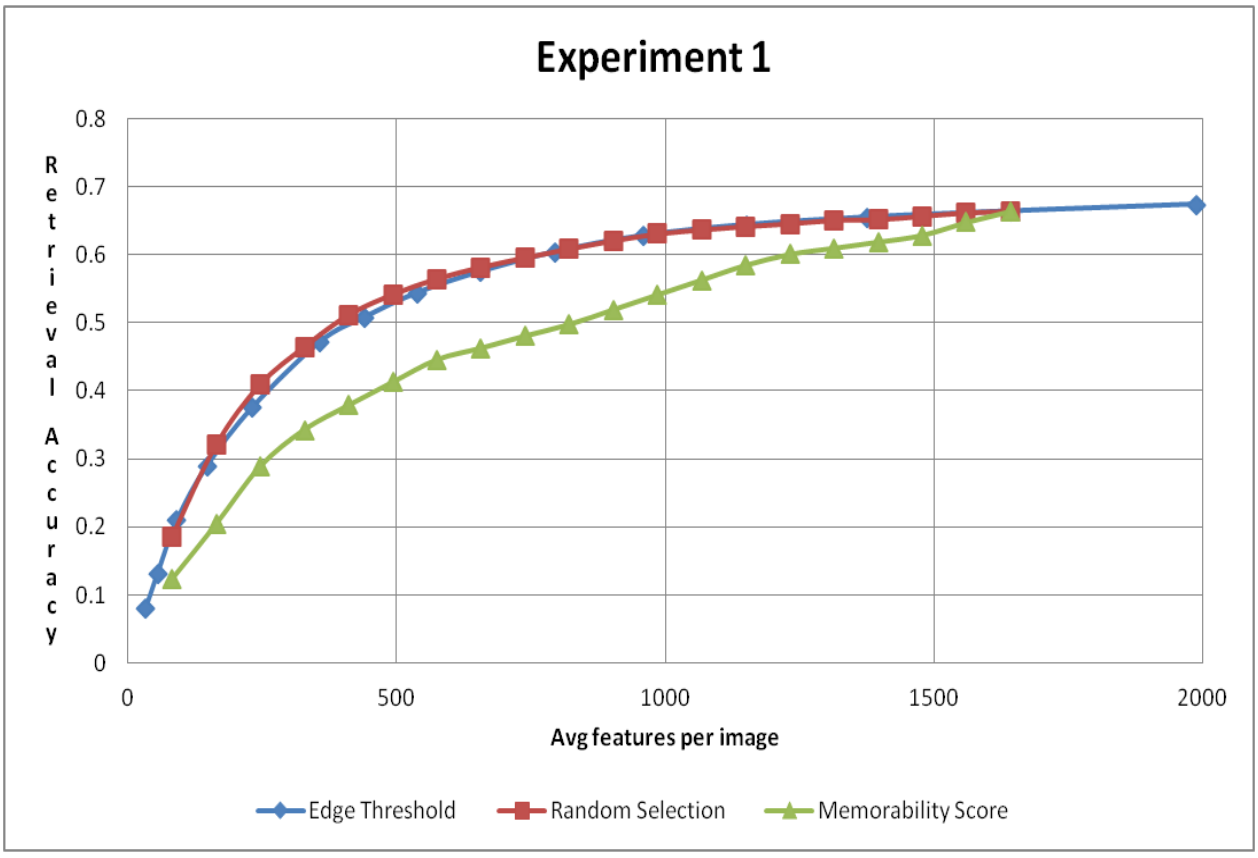


Figure 4 Experiment 1 results for PARIS dataset

Table 1 Retrieval accuracy and avg features per image for various edge thresholds for PARIS dataset

Edge Threshold	Avg features per Image	Retrieval Accuracy
1	1988.456	0.674
2	1641.336	0.664
3	1374.669	0.655
4	1150.517	0.644
5	958.35	0.628
6	794.228	0.604
7	655.326	0.575
8	538.416	0.544
9	440.12	0.508
10	357.688	0.472
12	232.241	0.376
14	147.286	0.29
16	91.348	0.21
18	55.461	0.132
20	32.931	0.08

Table 2 Retrieval accuracy and avg features per image for various % of selected features w.r.t. random selection/memorability score for PARIS dataset

% of selected features	Avg features per image	By random selection	By memorability score
100	1641.336	0.664	0.664
95	1559.747	0.662	0.648
90	1477.656	0.657	0.629
85	1395.611	0.652	0.619
80	1313.465	0.651	0.61
75	1231.377	0.646	0.601
70	1149.389	0.642	0.585
65	1067.346	0.637	0.5632
60	985.201	0.631	0.5414
55	903.208	0.621	0.5196
50	820.92	0.609	0.4978
45	739.079	0.596	0.481
40	656.938	0.582	0.463
35	574.941	0.565	0.446
30	492.85	0.542	0.413
25	410.709	0.512	0.379
20	328.673	0.465	0.343
15	246.676	0.41	0.289
10	164.582	0.321	0.204
5	82.54	0.186	0.123

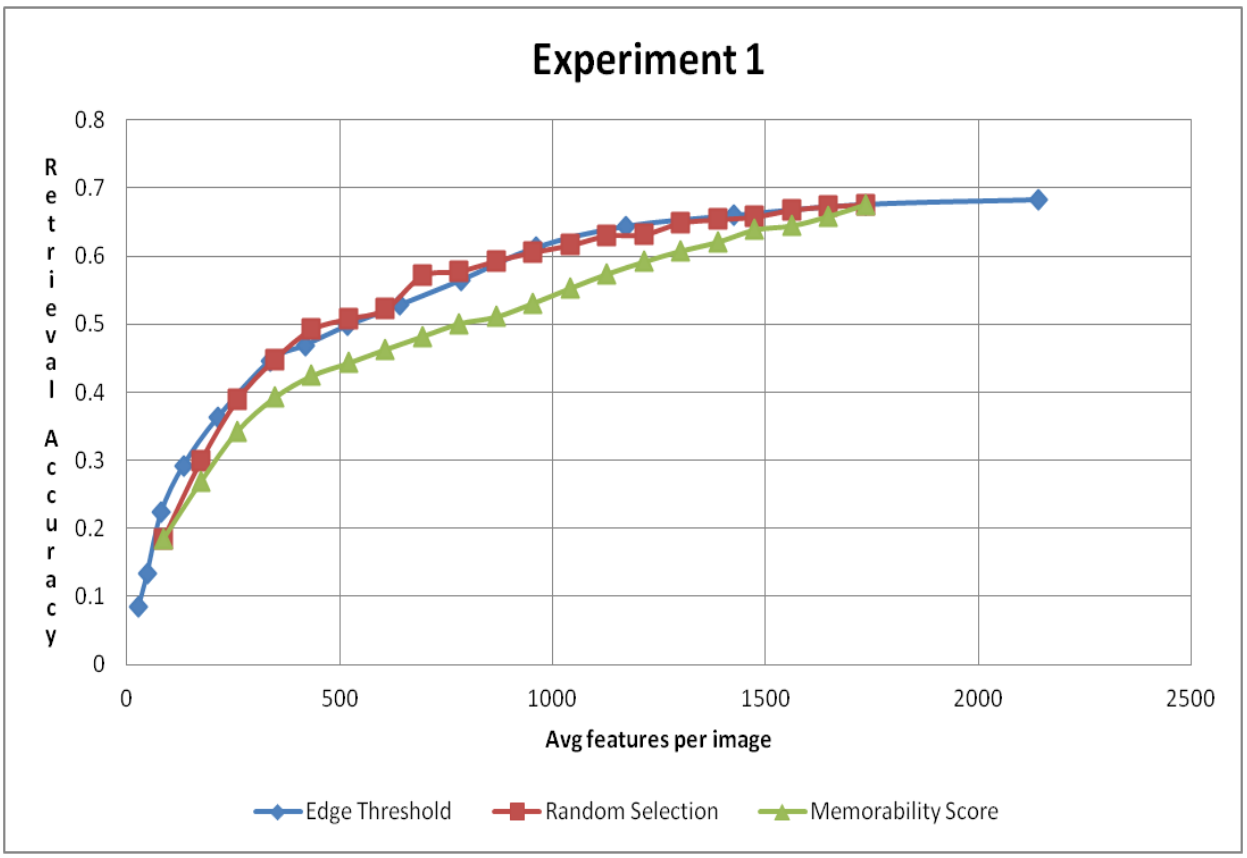


Figure 5 Experiment 1 results for OXFORD dataset

Table 3 Retrieval accuracy and avg features per image for various edge thresholds for OXFORD dataset

Edge Threshold	Avg features per image	Retrieval Accuracy
1	2140.552	0.683
2	1734.401	0.676
3	1425.142	0.66
4	1172.23	0.644
5	961.968	0.613
6	786.097	0.565
7	640.172	0.528
8	518.766	0.498
9	418.538	0.469
10	336.193	0.446
12	214.038	0.363
14	133.237	0.292
16	81.026	0.223
18	48.088	0.133
20	27.926	0.084

Table 4 Retrieval accuracy and avg features per image for various % of selected features w.r.t. random selection/memorability score for OXFORD dataset

% of selected features	Avg features per image	By random selection	By memorability score
100	1734.401	0.676	0.676
95	1648.155	0.673	0.659
90	1561.408	0.668	0.645
85	1474.719	0.658	0.639
80	1387.917	0.654	0.621
75	1301.172	0.649	0.608
70	1214.53	0.632	0.592
65	1127.837	0.63	0.574
60	1041.042	0.616	0.553
55	954.397	0.606	0.531
50	867.447	0.593	0.511
45	780.954	0.578	0.501
40	694.16	0.571	0.482
35	607.514	0.522	0.463
30	520.769	0.508	0.443
25	433.969	0.492	0.424
20	347.284	0.448	0.392
15	260.632	0.389	0.343
10	173.891	0.298	0.268
5	87.196	0.185	0.185

Results

A drop in retrieval accuracy was observed on dropping points of interest based on the memorability scores of the regions.

Experiment 2

Comparison of ranking of retrieved images based on a baseline setup and an ordering using memorability scores

Expected Result

Using memorability scores of regions the ranking of retrieved images based on memorability scores should appeal more to the human eye.

Baseline establishment

A baseline image retrieval pipeline was setup on the PARIS dataset which ranks matched images based on the distance of the matched feature. The stepwise procedure is same as for Experiment 1.

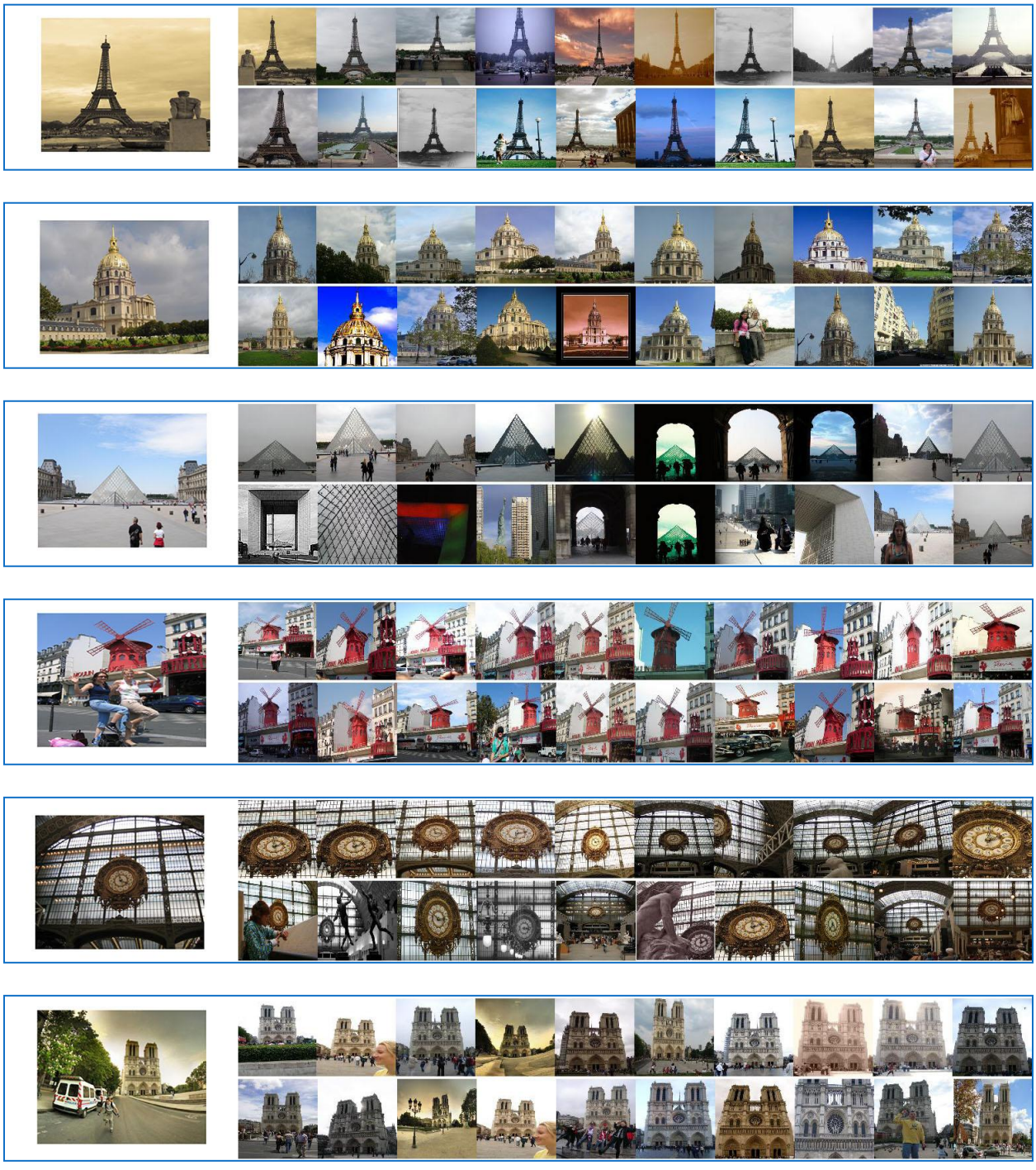
Experiment run

The match results in the baseline were reordered based on their memorability scores.

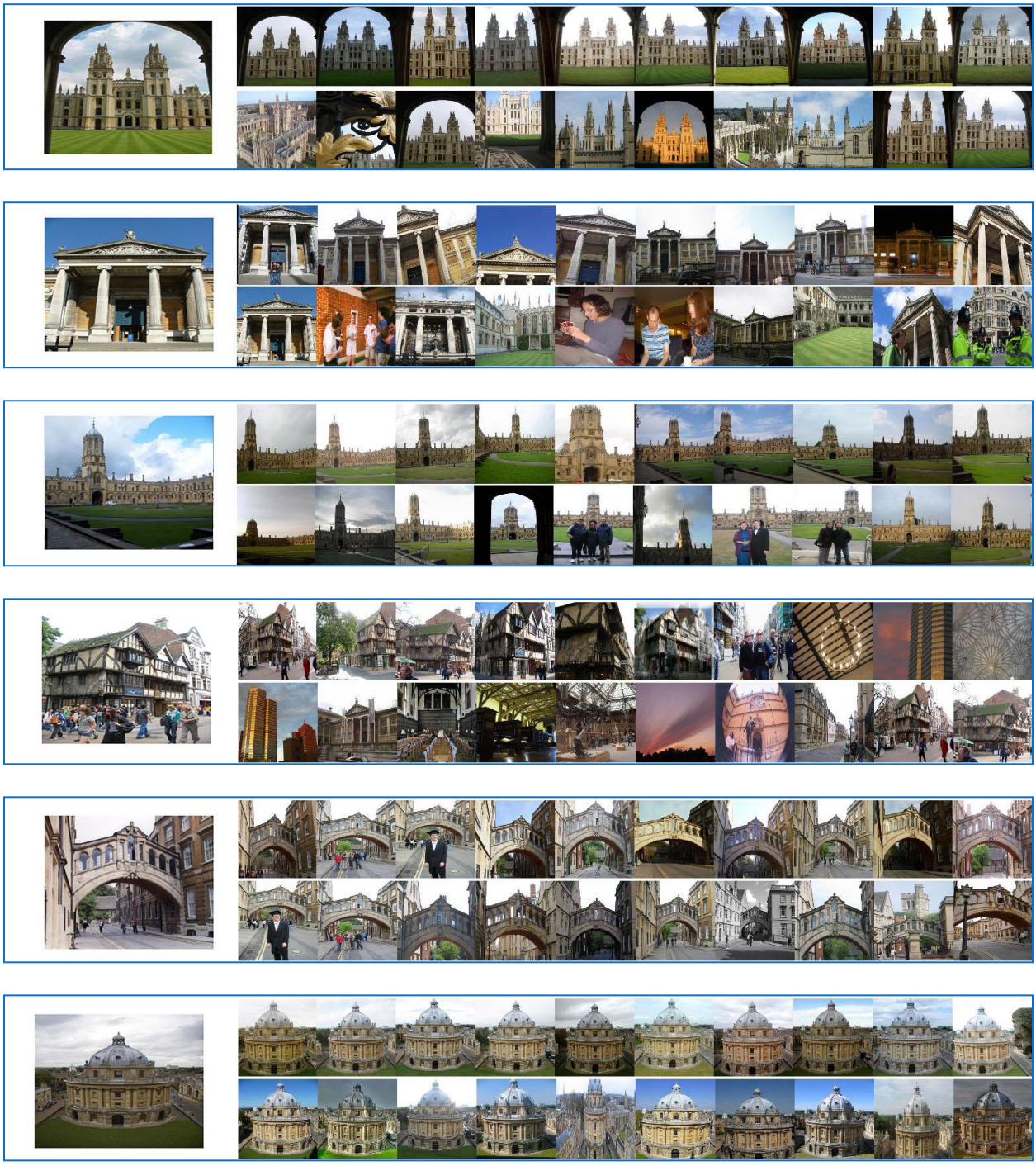
Observations

The image on the left is the query image and the top row represents the top matched images from the baseline setup while the lower represents the top 10 of the first 50 results from the baseline reordered based on their memorability scores.

Paris Dataset



Oxford Dataset



Result

No significant improvements could be observed by comparing the top 10% results.

Conclusion

Based on the results of the two experiments, it may be concluded that using memorability scores led to drop in the overall retrieval accuracy and impact on ranking could not be ascertained. Possible reasons for such results may be:

- Memorability scores are biased towards humans in photographs
- Majority of retrieval dataset contained landscapes and buildings
- Reordering of images based on memorability score is subjective
- Experiment setup and code – an attempt was made to simulate the paper but may have some flaws

Future Work

Even though this idea didn't turn out as expected, there may be variants to look into for future investigation. The same experiments could be run using different datasets, in which there are more images containing humans and the query images then could be set to have humans. The retrieved images ranking can be done by multiple people to see if the order has more significance. A hybrid ranking algorithm can be made so that the good matching images at least come above the insignificant ones. Maybe not all the components of the image memorability pipeline provide a positive feedback for detecting points of interest.

References

[1] K. Lenc, V. Gulshan, and A. Vedaldi, VLBenchmarks, *http://www.vlfeat.org/benchmarks/*, 2012

[2] Khosla, Aditya, et al. "Memorability of image regions." *Advances in Neural Information Processing Systems 25*. 2012.

[3] Isola, Phillip, et al. "What makes an image memorable?." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

[4] Khosla, Aditya, et al. "Image memorability and visual inception." *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012.

[5] Parikh, Devi, et al. "Understanding the intrinsic memorability of images."*Journal of Vision* 12.9 (2012): 1082-1082.

[6] Lowe, David G. "Object recognition from local scale-invariant features."*Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, 1999.

[7] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[8] Chandrasekhar, Vijay, et al. "CHoG: Compressed histogram of gradients a low bit-rate feature descriptor." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[9] Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos." *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

[10] Philbin, James, et al. "Object retrieval with large vocabularies and fast spatial matching." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007.

[11] Philbin, James, et al. "Lost in quantization: Improving particular object retrieval in large scale image databases." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[12] Philbin, James, et al. "Descriptor learning for efficient retrieval." *Computer Vision–ECCV 2010*. Springer Berlin Heidelberg, 2010. 677-691.

[13] Xiao, Jianxiong, et al. "Sun database: Large-scale scene recognition from abbey to zoo." *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010.

[14] *http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/*

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

[16] Naikal, Nikhil, Allen Y. Yang, and S. Shankar Sastry. "Informative feature selection for object recognition via sparse PCA." *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011.

[17] Knopp, Jan, Josef Sivic, and Tomas Pajdla. "Avoiding confusing features in place recognition." *Computer Vision–ECCV 2010*. Springer Berlin Heidelberg, 2010. 748-761.

[18] Everts, Ivo, Jan C. van Gemert, and Theo Gevers. "Per-patch descriptor selection using surface and scene properties." *Computer Vision–ECCV 2012*. Springer Berlin Heidelberg, 2012. 172-186.

[19] Ledwich, Luke, and Stefan Williams. "Reduced SIFT features for image retrieval and indoor localisation." *Australian conference on robotics and automation*. Vol. 322. 2004.

[20] Jegou, Herve, et al. "Aggregating local image descriptors into compact codes. "*Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.9 (2012): 1704-1716.

[21] Deselaers, Thomas, Daniel Keysers, and Hermann Ney. "Features for image retrieval: an experimental comparison." *Information Retrieval* 11.2 (2008): 77-107.

[22] Datta, Ritendra, et al. "Image retrieval: Ideas, influences, and trends of the new age. " *ACM Computing Surveys (CSUR)* 40.2 (2008): 5.