Image Retrieval for Online Browsing in Large Image Collections

Andrej Mikulik, Ondřej Chum, Jiří Matas

Center of Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

Abstract. Two new methods for large scale image retrieval are proposed, showing that the classical ranking of images based on similarity addresses only one of possible user requirements. The novel retrieval methods add zoom-in and zoom-out capabilities and answer the "What is this?" and "Where is this?" questions.

The functionality is obtained by modifying the scoring and ranking functions of a standard bag-of-words image retrieval pipeline. We show the importance of the DAAT scoring and query expansion for recall of zoomed images.

The proposed methods were tested on a standard large annotated image dataset together with images of Sagrada Familia and 100000 image confusers downloaded from Flickr. For completeness, we present in detail components of image retrieval pipelines in state-of-the-art systems. Finally, open problems related to zoom-in and zoom-out queries are discussed.

1 Introduction

A rapid increase in the size and ubiquity of shared image collections has motivated recent significant developments in image and specific-object retrieval. Most object-retrieval methods take into account the requirements for efficient content-based navigation and browsing of large-scale image collections.

Text search engines have provided the inspiration for the canonical approach to visual retrieval [30]. The user provides a query against which the retrieval engine ranks image relevance (or similarity). The performance of such an approach is typically assessed by a measure inherited from the text retrieval community: the average precision (AP). The state of the art and the standard components of the visual retrieval pipeline are reviewed in Section 2.

We show, however, that a similarity or relevance ranking of image-query results is not always suitable for browsing an image collection. This is demonstrated in the Fig. 1 rows denoted "nn", which depict the output of a query in a large-scale image-retrieval system. All the results are similar to the original image in scale and viewpoint, providing little additional information. The phenomenon is an inherent problem of ranking by approaches using similarity. The problem becomes more pronounced as the size of the collection increases, since more images from similar viewpoints and of similar scales are present in the dataset. On the other hand, the rows of Fig. 1 denoted “zoom in” show regions of interest in the highest detected resolution. We advocate that *“the most detailed view”* or, in, short “zoom-in”, is very probably the user intention after bounding-box selection.

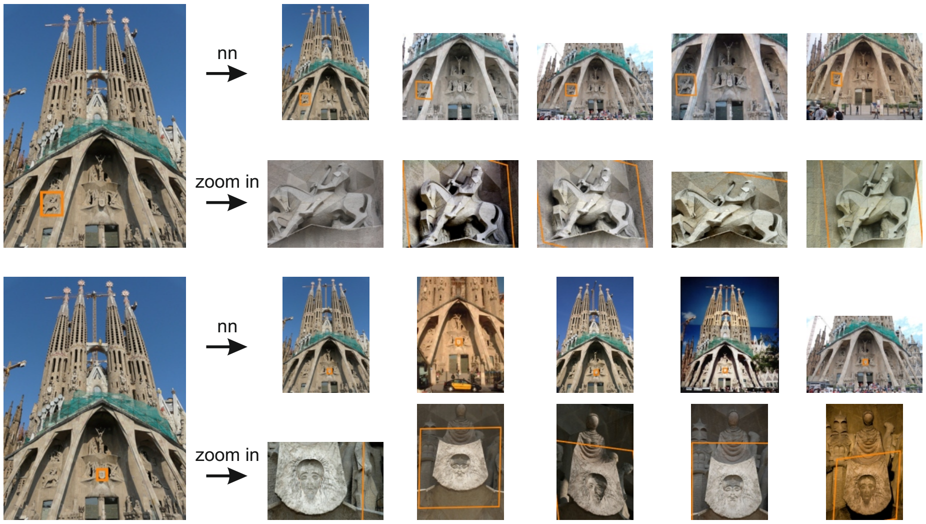


Fig. 1. Comparison of outputs of the standard and novel approaches. Two queries differing only by bounding-box were issued on the image in the leftmost column. The standard “most similar image” approach (nn, top rows) retrieves nearest neighbor matches, which provide no detailed images local to the bounding box and produce nearly identical results. The novel “most detailed view” approach or, *zoom-in*, maximizes the number of pixels inside the bounding box resulting in very different results (zoom in, bottom rows).

In the paper we address two tasks the user might be interested in: *“What is this?”* and *“Where is this?”*. The user expresses the first by selecting a bounding box from an image or simply moving a pointer over an image and forward-scrolling the mouse wheel. The expected result is a detailed image of the scene selected by the bounding box or of the local region centered around the pointer. The second expresses a desire for a broader contextual query.

In principle there are many tasks that might be of user interest: “What is to the left or right of this?”; “On which backgrounds can this object be seen?”; “Which objects can be seen on this background?”; “How does this object look like in the dark?”. To demonstrate the concept we focus exclusively on the zoom-in and zoom-out tasks 2.

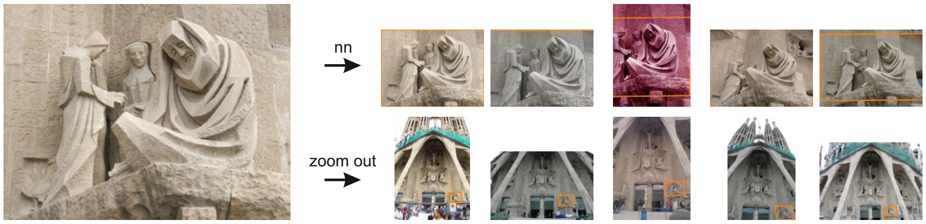


Fig. 2. Comparison of outputs of the standard and the proposed approach. The standard "most similar image" approach (nn, top rows) retrieves nearest neighbour matches, while the "context view" approach answers the question "Where is this?" by maximizing the scene content surrounding the bounding box, in this case, the whole query image (zoom out, bottom rows).

2 Standard components and state-of-the-art methods in large scale image retrieval

In this section we review three popular approaches that each use vector representations for images. Additionally, we present image retrieval approaches derived from techniques used in text search as well as standard methods for increasing precision and recall after scoring in the index file.

2.1 The bag of words image representation

One of the most popular image representations is the *bag of words* (BoW). Images are represented as collections of local features. A local feature has its visual appearance represented by a visual word and its spatial extent defined by a point and an ellipse.

Features, typically affine covariant regions, are detected for each image in the dataset. The most frequently used detectors in image retrieval engines are the Harris-affine [19, 29], Hessian-affine [19] and MSER [18], which have different detection characteristics, but collectively represent the state-of the art. A comprehensive performance survey of features detectors is given by Mikolajczyk *et al.* [20], which confirms the high performance of the above listed detectors.

Detected interest regions are described by a feature descriptor. The SIFT descriptor [17], which describes an interest region by a point in a 128-dimensional space, is ubiquitous in state-of-the-art systems. Many modifications have been proposed in the literature, including two effective and popular variants: rootSIFT [2] and SURF [5].

Feature descriptors are vector quantized into visual words [30] creating a visual vocabulary. Many approaches have been studied in the literature, with modifications addressing different goals and constraints.

The canonical vocabulary construction method is the unsupervised k-means clustering. The parameter k denotes the number of visual words in the vocabulary. The choice of k varies: from $k \approx 10^3$, usually suitable for classification

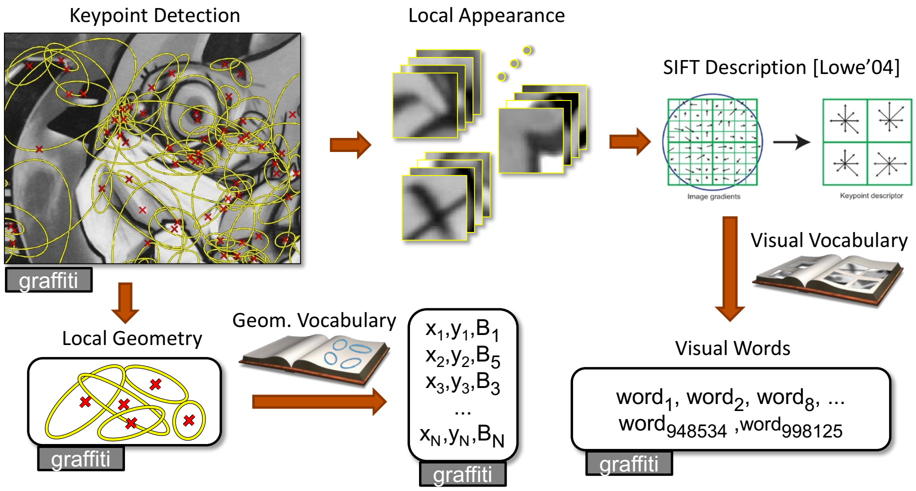


Fig. 3. Visualization of the bag of word image representation computation with geometry compression. Courtesy of Michal Perdˇoch.

tasks, up to $k \approx 10^7$ [21]. To efficiently construct large vocabularies, Nister *et al.* [23] proposed the use of a hierarchical vocabulary tree and Philbin *et al.* [26] use approximate nearest neighbour. Following the approach of Perdoch *et al.* [25], spatial information can be also compressed using unsupervised clustering without significant loss of precision. The process of image description is visualized in Figure 3.

2.2 Image representation with VLAD

The vector of locally aggregated descriptors (VLAD) [15] is another successful image representation method. It combines the advantages of the bag of words and the Fisher kernel [12]. As in the BoW representation, local features are detected and described. The vocabulary is created with k-means, but, unlike the BoW method, only a small number of visual words k are used. Jegou *et al.* [15] show that good results are achieved for $k \in [16, 256]$ visual words. Visual words are constructed by finding k cluster centers as before, but the descriptor assigned to a cluster center is computed as a sum of signed differences between the cluster center and its nearest feature descriptors, resulting in a $k \times d$ dimensional vector (d is the dimension of the local descriptor, *e.g.* 128 for SIFT). Product quantization [14] is used to construct the final quantized descriptor creating a compact representation that fits into 20 bytes.

2.3 GIST descriptor

A different approach to image representation is to create a global descriptor that captures the spatial layout and spatial relationships between regions or blobs of similar size, and the arrangement of basic geometric forms. One example is GIST, proposed by Oliva and Torralba [24]. A single GIST descriptor is used to represent an image, which results in a small memory footprint. The representation prevents partial matching of the image, it is sensitive to occlusion and there are no keypoints that can be used for spatial verification.

2.4 Image retrieval

The nearest neighbor (NN) search for similar images is slow for large datasets, even if it uses sophisticated data structures avoiding exhaustively examination of the image database. Approximate NN search offers a big improvement.

Text search engines [1, 4] face similar scalability problems for document retrieval, and the computer vision community has looked there for inspiration. In particular, image database indexing by the inverted file data structure leads to a dramatic speedup over the nearest neighbor search [30]. Inverted files map visual words to documents containing the words. The inverted file serves as in index into the database: upon a query, a subset of matching documents is returned, *i.e.*, those that contain the visual words of the query. The document ranking proceeds by calculating the similarity between the query vector and the matching document vectors. For sparse queries, the use of an inverted file ensures that only documents that contain query words are examined, which leads to a substantial speedup over the alternative of examining every document vector.

Efficient computation of the relevance of an image to a query is achieved by traversing the inverted file and reading the posting lists associated with the visual words of the query. The posting list (one row of the inverted file) associated with a visual word W is the list of image identifiers that contain visual word W. The standard tf-idf weighting scheme [3], also adopted from the document search community, is used to weight the document's relevance by de-emphasizing commonly occurring, less discriminative words.

Application of this approach is straightforward for sparse BoW vectors. For VLAD, similar speedup is achieved by combining the inverted file with asymmetric distance computation (IVFADC) proposed by Jegou *et al.* [14].

2.5 Spatial verification and query expansion

As shown in [26, 25], retrieval results are significantly improved by using the locations of features to verify their spatial consistency with the query region. This is achieved by a fast and robust hypothesize-and-test procedure that estimates an affine transformation between the query region and the target image. The RANSAC algorithm with local optimization [8] is widely used for spatial verification in state-of-the-art retrieval systems.

A caveat is that spatial verification is significantly more time consuming than BoW scoring. Thus it is performed only on the shortlist consisting of top scoring images. Furthermore, Chum *et al.* [9] show that if the model of the query (bag of words with feature geometries) is updated with newly spatially verified images by adding their visual words and geometries during the spatial verification, the probability of verifying other related images increases. Verified images in the shortlist are subsequently re-ranked.

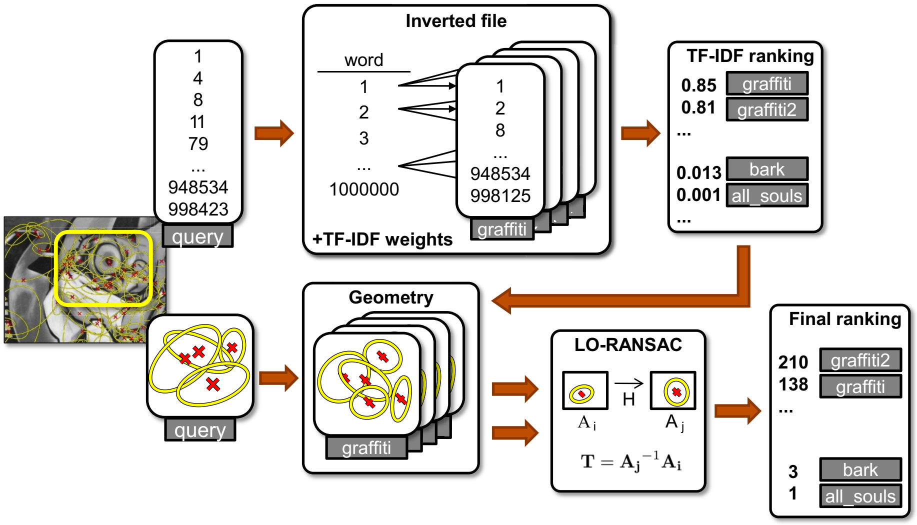


Fig. 4. Visualization of image retrieval with spatial verification for the bag of words representation. Courtesy of Michal Perdǒch.

Chum *et al.* [10] proposed a query expansion (QE) method – another technique inspired by text retrieval [6, 28] – to image retrieval and demonstrated impressive gains to recall. In QE, visual words from highly ranked images are composed in a new, expanded query. Unlike in text retrieval, features come with spatial information, typically keypoints, so geometric constraints and can be checked with spatial verification to ensure that the expanded query does not include visual words from a false positive image.

Chum *et al.* [9] added spatial context to queries by incorporating matching features that locally surround the initial query boundary into the query expansion. A latent model of the context of the query object is constructed by exploiting features surrounding the bounding-boxes of images verified by incremental spatial verification. A consistent context is learned and features belonging to the context can aid the expanded query, thus further improving recall. The process of image retrieval for BoW representation is summarized in Figure 4.

3 Overview of the zooming algorithm

The zooming algorithm, which implements the novel *"What is this?"* and *"Where is this?"* functionalities, is based on the standard bag of words image retrieval method. The distinction is in the choice of ranking function. Instead of ordering images according to similarity, it is designed to address new goals: maximizing detail or maximizing content.

To encourage a scale change, the ranking function requires knowledge of the geometric transformation between the query and the shortlisted images. The transformation is estimated by the RANSAC algorithm. The ranking function reorders only verified images, *i.e.*, the images for which a geometric transformation was found, preferring zoomed-in or zoomed-out images.

To increase recall, scoring with the inverted file is weighted to account for scale change. To achieve this, compressed geometric information of the features is stored with their visual words and the *document at a time* (DAAT) scoring [31] is used to process the posting lists. Using DAAT, the geometry of the features is examined concurrently with computation of image scores, and the standard tf-idf score is re-weighted according to the scale change of features and user intention.

Query expansion plays an important role in the method, and the incremental spatial verification and context learning as proposed in [9] is used. In our experiments, good results were achieved when images selected for query expansion were chosen with the same ranking function as used for final ranking. Optionally, the query expansion step can be repeatedly issued until the requested zoom is found or the system fails to retrieve new, zoomed-in images. The method is summarized in Algorithm 1.

Algorithm 1 Overview of the zooming algorithm. Note that step 5 represents a trade-off between the query time and output quality.

```
Input: Bag-of-words of the query image
Output: Ranked list of images

1. Fetch posting lists for query visual words and score in DAAT order for each scale band separately.
2. Re-weight scores in scale bands to prefer desired change in scale and create a shortlist.
3. Spatially verify images in the shortlist, incrementally building an expanded query.
4. Rank images according to the desired goal (zoom-in/zoom-out)
5. Return the result or form the expanded query with context learning and goto 1
```

3.1 Ranking functions

Many different tasks might be addressed with specific ranking functions. There are several options for zooming which can be useful for different tasks.

Zoom in. The simple option of ordering images according to the determinant of the geometric transformation between the query and the database image returns maximally zoomed images first. However, the top ranked images often cover only a small part of the scene selected by the bounding box. This ranking can be still useful if the images are going to be further processed, *i.e.*, compiled to a super-resolution image, used in a new expanded query, *etc.*.

We suggest that a user who browses the database expects to see the whole scene in the retrieved image. However, simply restricting the results to images that contain the whole bounding box often rejects significantly zoomed images with only a small fraction of the scene missing. Such images might be easily accepted by the user who usually does not want to be very precise while specifying the query bounding box.

A good trade-off between the zoom-in and a bounding box coverage was observed for the following ranking function:

$$z_{in} = \sqrt{\frac{A_r}{A_q}},$$

where A_r is the area inside the bounding box within the retrieved image and A_q is the area inside the query bounding box. The square root plays no role in the raking. It allows interpreting z_{in} as an estimate of the scaling of lengths (not the areas), which is consistent with zoom factor specification.

Zoom out. In this case, the naive "determinant of transformation" solution retrieves just images with similar scene content at lower resolution, providing no additional information.

To achieve the "*where is this*" or zoom-out goal, the user intuitively expects to see a large context of the query image. For this purpose, we propose the ranking function

$$z_{out} = \sqrt{\frac{A_r}{A_w}},$$

where A_r is the area inside the bounding box and A_w is the area of the whole retrieved image. In this case, we add the constraint that the whole bounding box must be visible in the result.

4 Experiments

A search engine was built for an expanded Oxford dataset (5063 images of Oxford landmarks) [26], which was augmented with 100000 confuser images and 15000 landmark images. The Oxford dataset, as well as other standard datasets, is not very suitable for demonstrating the zoom capabilities since it does not contain significantly zoomed-in or zoomed-out images. For this reason we added 15000 images downloaded from flickr containing the tag *Sagrada Familia*. This favorite landmark is very well covered with photos from the distance up to the greatest details on the Sagrada Familia facade.

4.1 Design choices.

Following most of the recent work on image retrieval, multi-scale Hessian-affine features were used for feature detection. As we are interested in zooming, a global descriptor cannot be used, because it does not allow parts-based search of images.

Features were described by the 128-dimensional SIFT descriptor. The standard K-means algorithm with approximate nearest neighbor [22] is used to learn a vocabulary with one million visual words. The vocabulary is learned on the independent standard Paris dataset [27] (6412 images).

As in [25], feature geometries are compressed. Four bits are allocated for scale and 12 bits for shape compression. The compressed geometries are stored in the inverted file along with the visual words for fast access during DAAT scoring.

After scoring using the inverted file, a shortlist of the top 100 images is created. Incremental spatial verification is used and images are reordered with a chosen rank function. Optionally, the context of the query is learned and an expanded query is issued.

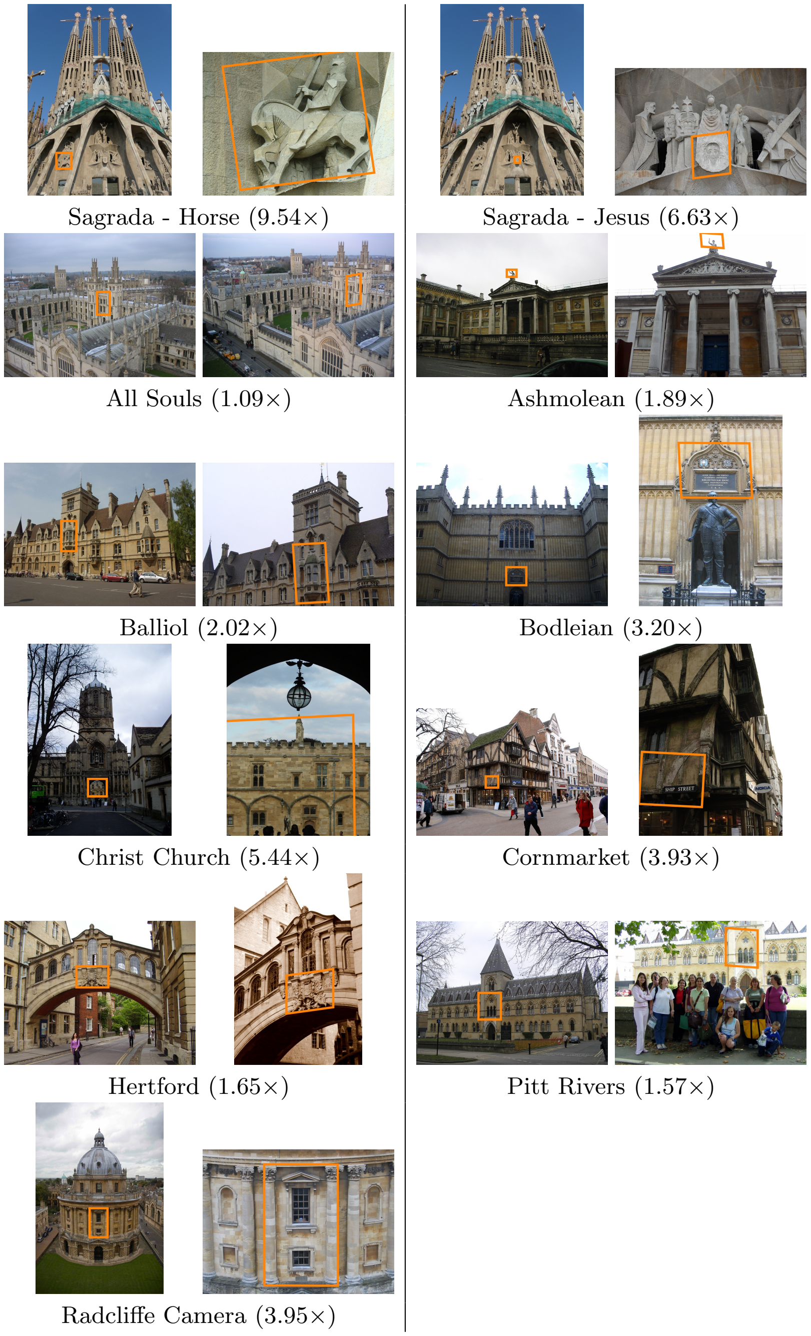
4.2 Evaluation protocol

To our knowledge there is no standard dataset with an evaluation protocol suitable for testing zooming capabilities. To demonstrate the method, we chose 2 queries from Sagrada Familia and 9 queries from the Oxford dataset. The queries and the top results retrieved with the zoom-in method are shown in Figure 5. Note that even if the Oxford dataset is not well covered with detailed views of landmarks, the user can, for instance, use the zoom-in to view architectural detail (Sagrada), read street names (Cornmarket), boards (Bodleian) or virtually navigate through the scene (going through the archway at Christ Church).

Table 1 shows, for 11 selected queries, the zoom-in result in top ranked image and an average zoom in top 5 retrieved images. The baseline nearest neighbour (nn) search with context based query expansion (QE) is compared with three zoom-in methods. First includes only ranking function (rank), second utilizes DAAT scoring in inverted file (DAAT), and the last adds query expansion (DAAT+QE).

5 Conclusions

We have presented two new methods for large scale image retrieval demonstrating that the classical ranking of the images based on similarity is only one of



Sagrada - Horse (9.54×)

All Souls (1.09×)

Balliol (2.02×)

Christ Church (5.44×)

Hertford (1.65×)

Radcliffe Camera (3.95×)

Sagrada - Jesus (6.63×)

Ashmolean (1.89×)

Bodleian (3.20×)

Cornmarket (3.93×)

Pitt Rivers (1.57×)

Fig. 5. Query images (on the left in each column) and the top results using the zoom-in method with DAAT scoring and query expansion. The effective zoom is in parentheses.

	top 1				top 5 average			
	nn	zoom-in			nn	zoom-in		
query	QE	rank	DAAT	DAAT+QE	QE	rank	DAAT	DAAT+QE
Sagrada - Horse	0.98	1.82	4.09	9.54	1.16	1.41	2.04	8.03
Sagrada - Jesus	0.86	2.75	2.75	6.63	1.22	1.22	1.87	6.00
All Souls	1.03	2.31	2.31	1.09	1.03	1.41	1.50	1.08
Ashmolean	1.43	1.43	1.43	1.89	1.28	1.28	0.77	1.45
Balliol	0.95	2.02	2.02	2.02	1.00	1.00	0.61	0.81
Bodleian	0.92	1.82	2.85	3.20	1.10	1.08	1.20	2.11
Christ Church	1.77	1.77	5.44	5.44	1.52	1.52	2.57	1.77
Cornmarket	1.57	3.93	3.93	3.93	1.39	1.97	1.97	1.97
Hertford	1.28	1.65	1.65	1.65	1.02	1.35	1.35	1.35
Pitt Rivers	1.30	1.36	1.57	1.57	1.30	1.22	1.10	1.10
Radcliffe Camera	1.29	3.95	3.95	3.95	1.23	2.03	2.04	2.35

Table 1. Comparison of the standard method (nn) and zoom-in. We report the zoom of the first ranked image (top 1), and the average zoom of the top five images (top 5 average). Four methods were compared: 1. the baseline nearest neighbor search with query expansion (nn, QE), 2. Zoom-in only by shortlist re-ranking (rank), 3. DAAT scoring and re-rank (DAAT), 4. DAAT scoring, ranking function and query expansion (DAAT+QE). In all four cases, incremental spatial verification was used.

many retrieval problems. In very large databases, the standard retrieval of the most similar images is unlikely to be useful as in many cases it returns just near duplicates.

The newly proposed retrieval methods add zooming capabilities and answer the “What is this?” and “Where is this?” questions. The functionality has been achieved by modifying two steps of the standard bag-of-words retrieval pipeline, namely the scoring and ranking functions.

We show the importance of the DAAT scoring and query expansion for recall of zoomed images. The proposed methods were tested on a standard large annotated image dataset together with images of Sagrada Familia and 100000 image confusers downloaded from Flickr.

Open problems. Ordering images according to criteria other than similarity aggravates the problem of false positives. In a standard retrieval system, images are spatially verified and re-ordered according to the number of verified correspondences – the inliers of a geometric transformation obtained by the RANSAC algorithm. Irrelevant but highly similar (in the bag of words sense) images usually have only a small number of incidentally geometrically verified correspondences and are ranked after the true positive images. However, in the case of zooming, an incorrectly verified image often score as the zoomed version of the query. Such false positives are immediately recognized by the user. Moreover, the false positives are used in the subsequent query expansion which may lead to a complete failure of the system, *i.e.* to an irrelevant response to the query.

Retrieval with zooming is also sensitive to the presence of repetitive structures [16, 11, 32]. Zooming in on man-made objects very often reveals small repetitive patterns – textures on facades of building, bricks, fences, bars *etc.*, which can often cause failure of spatial verification and consequently of query expansion. Attention to burstiness [13], co-occurring features [7] and automatic failure recovery [9] alleviates the problem.

Acknowledgments

The research was supported by ERC-CZ-LL1303, the Czech Science Foundation Project GACR P103/12/G084 and Microsoft scholarship.

References

1. Y. Aasheim, M. Lidal, and K.M. Risvik. Multi-tier architecture for web search engines. *Web Congress, 2003. Proceedings. First Latin American*, 2003.
2. R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, pages 2911–2918. IEEE, 2012.
3. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, ISBN: 020139829, 1999.
4. L.A. Barroso, J. Dean, and U. Holzle. Web search for a planet: The google cluster architecture. *Micro, IEEE*, 23, 2003.
5. H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006.
6. Ch. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. *NIST SPECIAL PUBLICATION SP*, pages 69–69, 1995.
7. O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. CVPR*, 2010.
8. O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. In *Pattern Recognition*, pages 236–243. Springer, 2003.
9. O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *Proc. CVPR*, pages 889–896, Los Alamitos, USA, June 2011. IEEE Computer Society, IEEE Computer Society. CD-ROM.
10. O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.
11. Petr Doubek, Jiri Matas, Michal Perdoch, and Ondrej Chum. Image matching and retrieval by repetitive patterns. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3195–3198. IEEE, 2010.
12. T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.
13. H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Proc. CVPR*, 2009.
14. H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE PAMI*, 33(1):117–128, 2011.
15. H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.

16. T. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *Proc. ECCV*, pages 546–555. Springer, 1996.
17. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Proc. ICCV*, 60(2):91–110, 2004.
18. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and David Marshall, editors, *Proc. BMVC.*, volume 1, pages 384–393, London, UK, September 2002. BMVA.
19. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, pages 128–142. Springer, 2002.
20. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1/2):43–72, 2005.
21. A. Mikulik, M. Perd̆och, O. Chum, and J. Matas. Learning vocabularies over a fine quantization. *IJCV*, pages 1–13, 2012.
22. M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009.
23. D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006.
24. A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155, 2006.
25. M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR*, 2009.
26. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.
27. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in largescale image databases. In *Proc. CVPR*, 2008.
28. G. Salton and Ch. Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, pages 355–364, 1997.
29. F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *Proc. ECCV*, pages 414–431. Springer, 2002.
30. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470 – 1477, 2003.
31. H. Stewénius, S. H. Gunderson, and J. Pilet. Size matters: exhaustive geometric verification for image retrieval accepted for eccv 2012. In *Proc. ECCV*, pages 674–687. Springer, 2012.
32. A. Torii, J. Sivic, T. Pajdla, and Okutomi M. Visual place recognition with repetitive structures. In *Proc. CVPR*, 2013.