CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY

PhD THESIS

ISSN 1213-2365

Large-Scale Content-Based Sub-Image Search

PhD Thesis

Andrej Mikulík

mikulik@cmp.felk.cvut.cz

CTU–CMP–2014–08

June, 2014

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/mikulik/mikulik-phdthesis.pdf

Thesis Advisor: Prof. Ing. Jiří Matas Dr.
Co-supervisor: Doc. Mgr. Ondřej Chum Ph.D.

Author was supported by Microsoft Research Cambridge Scholarship.

Research Reports of CMP, Czech Technical University in Prague, No. 8, 2014

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Large-Scale Content-Based Sub-Image Search

A Dissertation Presented to the Faculty of the Electrical Engineering of the Czech Technical University in Prague in Partial Fulfillment of the Requirements for the Ph.D. Degree in Study Programme No. P2612 - Electrotechnics and Informatics, branch No. 3902V035 - Artificial Intelligence and Biocybernetics, by

Andrej Mikulík

June 2014

Thesis Supervisor
Prof. Ing. Jiří Matas Dr.

Thesis Co-supervisor
Doc. Mgr. Ondřej Chum Ph.D.

Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
Karlovo náměstí 13, 121 35 Prague 2, Czech Republic
fax: +420 224 357 385, phone: +420 224 357 465
http://cmp.felk.cvut.cz

Abstract

In this work the problems of specific object and image retrieval including the more challenging sub-image are studied. Given a query image of a specific object a retrieval engine returns relevant images of the same object from a database. The thesis focuses on the bag-of-words approach which is one of the most effective content-based approach especially when the specific object covers only a part of the picture, can be occluded or only partially visible. The thesis improves a number of components of the standard bag-of-words retrieval approach.

A novel similarity measure for bag-of-words type large scale image retrieval is presented. The similarity function is learned in an unsupervised manner, requires no extra space over the standard bag-of-words method and is more discriminative than both L2-based soft assignment and Hamming embedding. The novel similarity function achieves mean average precision that is superior to any result published in the literature on the standard datasets and protocols.

We study the effect of a fine quantization and very large vocabularies (up to 64 million words) and show that the performance of specific object retrieval increases with the size of the vocabulary. This observation is in contradiction with previously published results. We further demonstrate that the large vocabularies increase the speed of the tf-idf scoring step.

All state-of-the-art image retrieval results in the literature have been achieved by methods that include a query expansion which brings a significant boost in performance. We introduce three modifications to automatic query expansion: (i) a method capable of preventing *query expansion failure* caused by the presence of confusers, (ii) an improved spatial verification and re-ranking step that incrementally builds a statistical model of the query object and (iii) we learn relevant spatial context to boost retrieval performance.

All three improvements of query expansion were evaluated on standard Paris and Oxford datasets and state-of-the-art results were achieved.

Finally, novel problems for image retrieval are formulated. It is shown that the classical ranking of images based on similarity addresses only one of possible user requirements. Instead of searching for the most similar images, the novel retrieval methods zoom-in and zoom-out answer the *"What is this?"* and *"Where is this?"* questions.

In addition, two other task are formulated: (i) given a query and a large image dataset, for every pixel location in the query, find an image with maximum resolution and (ii) return the frequency with which a pixel appears in the dataset.

The zoom-in and zoom-out required the development of two novel techniques: the hierarchical query expansion method and a geometric consistency verification step that is sufficiently robust to prevent a topic drift within a zooming search. Experiments show that the proposed methods find surprisingly fine details on the tested landmarks, even those that are hardly noticeable for humans.

Acknowledgments

I would like to express my thanks to my colleagues at Center for Machine Perception who I had the pleasure of working with. My special thanks go to Prof. Václav Hlaváč for maintaining such a high quality workplace with such a pleasant atmosphere.

I am thankful to my supervisor Prof. Jiří Matas for his patience, enthusiasm and his support and to my co-supervisor doc. Ondřej Chum for all the theoretical knowledge and critical view on research he shared with me.

I thank to my fellow student and friend Michal Perďoch who helped me to understand the CMP code-base and introduced me to the topic of image retrieval.

Finally, I would like to thank to my family and to my friends for their support that made it possible for me to finish this thesis.

I gratefully acknowledge the support by Microsoft Research Cambridge Scholarship.

Contents

Chapter 1

Introduction

In the early days of the Internet the vast majority of information available on the network was textual in the form of plain html web-pages. As the Internet became more popular, the amount of information stored on-line grew rapidly and the need of automated web search service became obvious. With over 11 billion web-sites currently on-line[1], querying with the search engines became a natural way of browsing the Internet. It is a common entry point.

Facilitated by the Web 2.0 concepts, the number of images stored on the Internet grew enormously in the last decade. Photo-sharing through various services like Panoramio [pan], Flickr [fli], Picasa [pic] and through social networks like Facebook [fac], Instagram [ins], Google+ [gooa] became very popular. The digital camera became a common equipment of every tourist, mobile phones without a camera have almost vanished. Posting a new image on a social network takes only few seconds. For instance, Facebook reported 350 millions of photos uploaded every day and the total number reached 250 billions in July 2013. And this alone is an order of magnitude bigger then the number of web-pages.

Besides social networks, other huge sets of digital images became publicly available due to commercial efforts like Google Street View [gooc], San Francisco Landmark [nok] or are being created for specific private reasons (surveillance, hospitals, etc.)

The datasets differ in several properties. Some sets covering views from a car evenly from whole cities, others contain densely sampled well known touristic landmarks, landscapes or various social events. The statistics of the image sets differ in resolutions and image quality – from low resolution images taken by cell phones and uploaded on social networks through mid-resolution images taken by specialized omnidirectional cameras for street views up to high-resolution artistic images taken with DSLR cameras that can be found in datasets like Panoramio.

It is a challenging task to process, index and enable the search in a huge amount of data. The conventional search using tags, annotations or surrounding texts from a webpage is limited. Many images do not have any kind of annotation and the information is present only in the content. This motivated Sivic and Zisserman [SZ03] to create an content-based analogy of text retrieval methods for the image and video domain, which attracted a significant attention of the computer vision community. Several commercial systems already exist, but the problem remains open.

[1] Estimated in May 2012 by [www] using indexed page totals from several major search engines.

The goals of this work are twofold: to improve the existing state-of-the-art retrieval methods and to define and add new functionality – type of queries, which are useful to the users despite of having no analogy in text retrieval.

1.1 Motivation

There are many applications based on exploration of large image datasets: creating maps and 3D models of landmarks or whole cities [ASS+09], localization based on the photos taken with mobile phone [SBS07], browsing collections [SSS06, CM10a], discovering canonical views [WL11], searching different image of the object or scene, searching for information about the object taken by camera [CBK+11, goob], Internet-based in-painting [OJA09], and many others.

Applications based on these collections have to be able to process, index, categorize or search the data. Due to the vast amount of images, methods suitable for querying large datasets must not only have sub-linear running time that grows slowly with the size of the collection, but must be very efficient in the use of memory. As soon as the representation of the complete collection fails to fit into operation memory, running time for a single search jumps by orders of magnitude – from a fraction of a second to 15-35 seconds per query on single computer as reported in [CPS+07]. For web-scale datasets parallelization of the whole process must be possible.

A rapid increase in the size and ubiquity of these photo collections has motivated significant developments in image and specific-object recognition and retrieval but still current state-of-the-art search engines have good results only in some domains. The best results are achieved on rigid and well structured objects with good texture. This includes man-made objects as buildings, streets, landmarks, paintings, etc... However many other domains remain a challenging open problem - searching for individual persons, animals, threes or flowers, objects with lack of texture as chairs or keys, wire-ish objects like fences, nets, or transparent or glossy objects.

Another type of challenges in image retrieval are visual properties like large differences in point of view, scale, colors, differences between day and night or presence of blur. The effort is put to widen the variety of conditions in which the system performs well.

The core problem addressed in this work is to retrieve as many relevant images as possible while avoiding false positives.

1.2 Large Sub-Scale Image Search – Classical Problem Definition

Large scale image search is the problem of *specific object retrieval* from a web-scale, unordered set of images. The user provides a *query* - an image or an image with a region of interest covering a specific object selected. The retrieval engine searches the database and ranks images according to *relevance*. The ranking expected by the user depends on the application. The majority of retrieval systems in the literature define *relevance* as a *similarity* with the query image. In this work (Chapter 6) we show that

different metrics can be used to better serve different type of user questions: *"What it is?"*, *"Where it is?"*, *"What is interested here?"*, ...

In this work we are interested in *sub-image* retrieval. This is a more difficult problem than a *whole-image* retrieval, since the searched object can appear only in a small part of the retrieved image (see Figure 1.1 - Nike logo). Moreover search engine should be able to retrieve partially occluded objects, taken from a different viewpoint or under different lighting conditions.

In the literature the term *near duplicate image* is often used in connection with image search. The definition of a near duplicate varies depending on which photometric and geometric variations are deemed acceptable. The application ranges from exact duplicate detection where no changes are allowed or only small change of scale, different compression or border removal to a more general definition that requires the images to be of the same scene, but with possibly slightly different viewpoints and illumination. The *specific object retrieval* problem is more general. The same scene is not required and different conditions of images acquiring are expected.

Large scale, in this work, means databases with about 10^7 images. While being two or three order of magnitude smaller than in web-scale databases or collections stored on social networks, our experimental retrieval system is running in real time on a single conventional machine. Moreover, all parts of the system are designed in a way that parallel computation is possible and time complexity of the online phase is sub-linear.

1.3 Performance of a Retrieval System

One of the goals of this work is to broaden the conditions under which the retrieval system performs well – to increase the *precision* and the *recall*. These two properties are measured to evaluate the performance of the system. Loosely defined, to increase the recall means to find and return more images from the database relevant to the query image without increasing the size of the output, while to increase the precision means

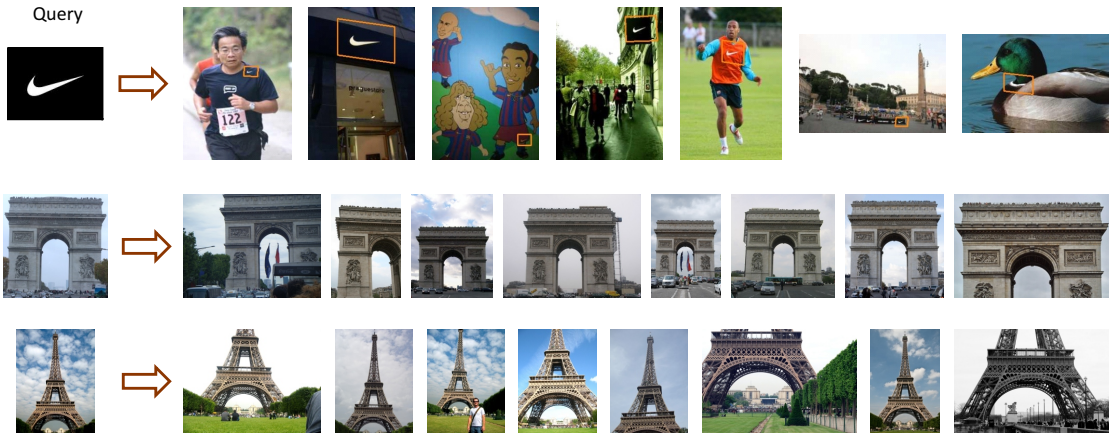


Figure 1.1: An examples of a query and ranked results from the image retrieval system.

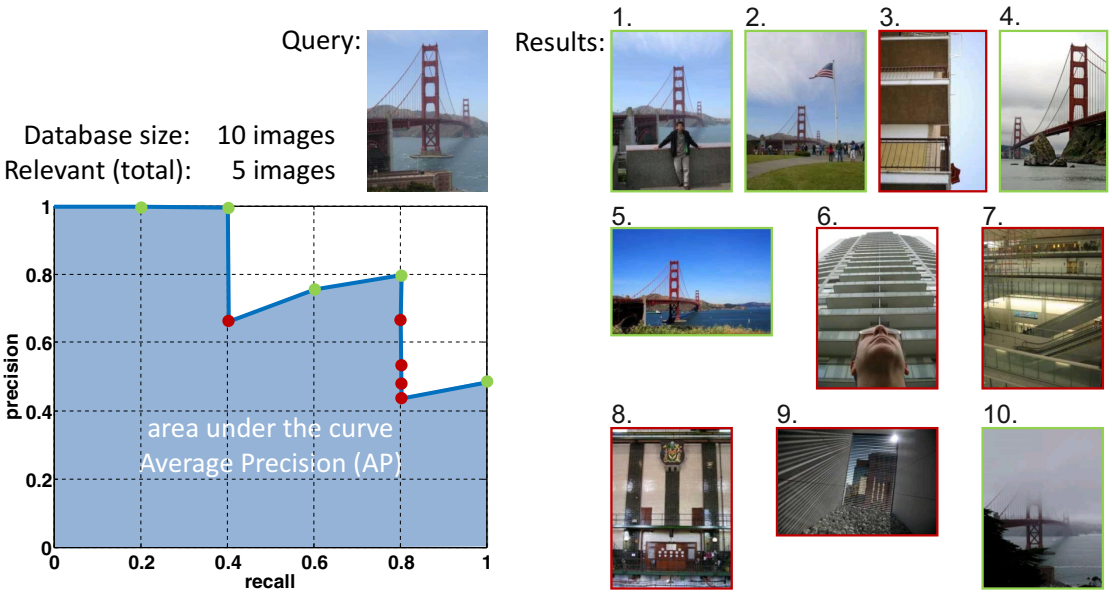


Figure 1.2: A toy example of average precision (AP) calculation for image a retrieval query. True positive results are highlighted with green color, false positives with red.

to avoid false positives among the relevant results. Defined accurately [Zhu04]:

$$precision = \frac{|\{\text{relevant images}\} \bigcap \{\text{retrieved images}\}|}{|\{\text{retrieved images}\}|}$$

$$recall = \frac{|\{\text{relevant images}\} \bigcap \{\text{retrieved images}\}|}{|\{\text{relevant images}\}|}$$

These two metrics are based on the whole result set returned by the system. The quality of the retrieval however, as perceived by the user, depends mainly on the ranking of the result set. For systems returning a ranked result set, the *Average Precision* (AP) is defined:

$$AP = \frac{1}{2} \sum_{k=1}^{n} (p(k) + p(k-1))(r(k) - r(k-1)),$$

where n is number of retrieved images, $p(k)$ resp. $r(k)$ is precision resp. recall of the set of images ranked from 1 to k. ($p(0) = 1$ by definition). A toy example of average precision is shown in Figure 1.2.

A common measure for evaluating performance of the information retrieval systems is *mean Average Precision* (mAP). The mAP is a mean of the average precision scores for each query defined by a protocol over a given dataset. To enable comparison between research groups several dataset became standard. We describe them in Section 1.4.

We are using measure and standard datasets with protocols for evaluations in Chapters 3, 4 and 5. However, in Chapter 6 different metrics than similarity are discussed for retrieving and ranking images. While different metrics can be more error prone or have

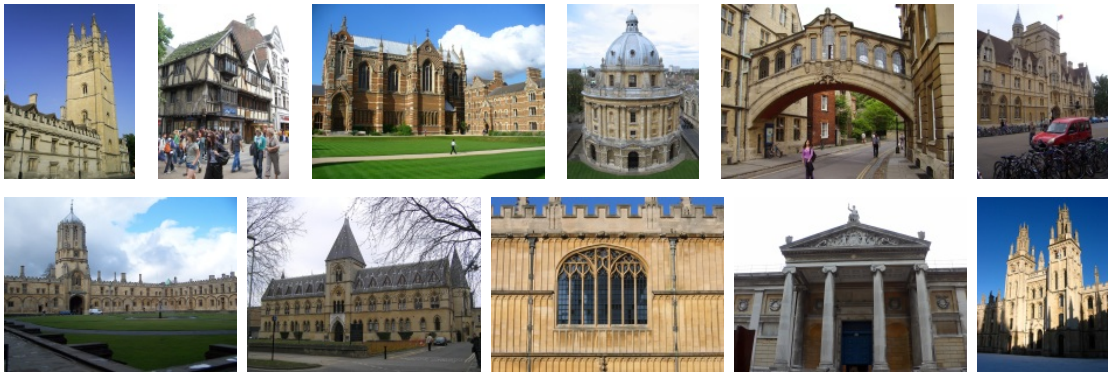


Figure 1.3: Example query images of eleven landmarks from the Oxford Buildings dataset.

lower recall, they still can be far more useful to the user. We show that the mAP is not a good estimator of the retrieval performance in this case and the standard datasets are not sufficient enough to demonstrate the quality of response to the novel query types. Some qualitative experiments are therefore carried out and results are shown on much bigger non-standard dataset as well.

1.4 Standard Datasets

During the last few years, several datasets and protocols become standard for comparison of the image retrieval systems. In this work the mAP of the developed image retrieval system is evaluated on the *Oxford Buildings dataset* [PCI$^+$07] (Oxford-5k), the *Paris Buildings dataset* [PCI$^+$08] (Paris-6k) and the *INRIA Holidays dataset* [JDS08, hol] and their extensions.

Oxford-5k is a dataset containing 5062 images downloaded from Flickr. Eleven different Oxford landmarks were identified and 5 queries (image and bounding box) were set for every landmark. Images were manually annotated and ground truth for all 55 queries was created. One query example of each landmark is shown in Figure 1.3.

Paris-6k is a dataset containing 6412 images collected in the same way as Oxford-5k. The same protocol is used and 11 landmarks with the ground truth were manually picked and annotated. Landmarks and their query examples are shown in Figure 1.4.

INRIA Holidays is a dataset containing 1491 images from personal holidays photos of INRIA researchers with a few images added to exploit robustness to rotations, viewpoints, illumination changes and blurring. Unlike the previous datasets, INRIA Holidays contains a large number of scenes difficult for retrieval approach chosen in this work – nature sceneries, water and fire effects and similar images lacking rigid and textured structures or spatial consistency. The images were manually grouped into 500 groups and the protocol with a ground truth created. The first image of each group is

Figure 1.4: Example query images of eleven landmarks from the Paris Buildings dataset.

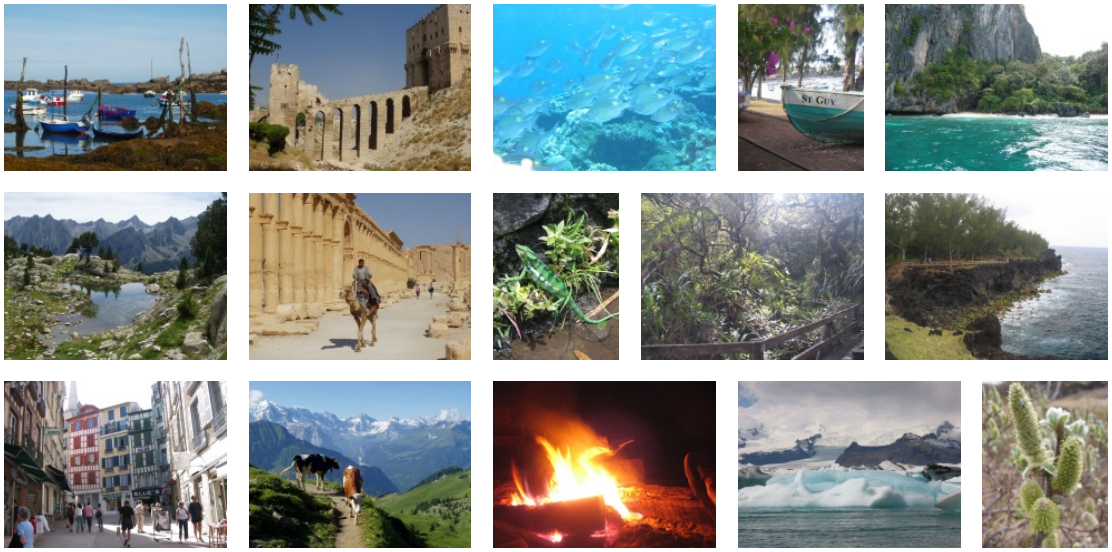


Figure 1.5: Example images from the INRIA Holidays dataset.

the query image and the other images of the group are the expected true positive results. Examples of images from the dataset are shown in Figure 1.5.

Datasets Oxford-5k and Paris-6k have been often augmented with 100.000 unrelated images downloaded from Flicker. The augmented sets are referred to as **Oxford-105k** and **Paris-106k** datasets. The results have been reported for many methods on these datasets which is important for performance comparison. They are available for download from creator's web pages.

1.5 Baseline Bag-of-Words Image Retrieval

In this work we are interested in large scale image search systems based on the state-of-the-art *bag-of-words* (BoW) approach. The approach was inspired by text search engines and applied to visual domain by Sivic *et al.* [SZ03]. This section briefly describes the retrieval pipeline based on bag-of-words. A more detailed description and other approaches with their modifications are presented in Chapter 2.

The process of the image search consist of two phases, which can be roughly divided into several parts common to most of the search systems.

First phase, which runs off-line, consists of parts for preparation of the system – building the database index of given images for the fast image search. (Alg. 1, Fig. 2.1):

Algorithm 1 Overview of the off-line phase – preparation of the bag-of-words retrieval system.

Input: a dataset of images
Output: a visual vocabulary, the index for fast image retrieval - *inverted file*
Main parameters: selection of feature detector, selection of descriptor, size of the vocabulary, vocabulary learning method

1. **Feature detection and description**, the process of reduction of image data that allows to speed-up the search by keeping only relevant and discriminative information from the image.

2. **Vocabulary learning** is a quantization of the descriptor space into clusters called visual words. The visual words are used instead of descriptors in the index file. The use of the visual words (their integer IDs) instead of the full descriptors in the index file leads to significant data reduction, which allows to index large collections of images and search them in real-time.

3. **Visual word assignment** is the process of assignment of one or more visual words to a feature descriptor. In the off-line phase all descriptors of all images in the dataset need to be assigned to visual words.

4. **An inverted file construction.** The inverted file is an index file used in the retrieval that stores for every visual word a list of documents containing the word.

Second phase, querying the system with given image, is running on-line (Alg. 2, Fig. 2.2):

Algorithm 2 Overview of the on-line querying phase.

Input: query image, visual Vocabulary, inverted file
Output: ranked result list of images from dataset
Main parameters: length of shortlist, ranking function, spatial verification thresholds

1. **Feature detection and description.** The features are detected and described in the query image using the same detector and descriptor as in the off-line phase.

2. **Word assignment.** Query image descriptors are assigned to visual words of the learned vocabulary. The assignment can be done in a different way than in the off-line phase.

3. **Index look-up.** Using the inverted file, images containing at least one common visual word are ordered according to score estimated according to bag-of-words weighted similarity.

4. **Spatial verification.** Images with the highest rank (short list) are spatially verified by looking for geometrical transformation between the retrieved image and the query.

5. **Query expansion [optional].** To achieve a higher recall, a new query can be formed from the visual words of top retrieved and verified images. This is an optional step. If executed, steps 2.–4. (an index look-up and spatial verification) are repeated for a new expanded query.

1.6 Contributions

The contributions of this thesis can be divided into three categories: vocabulary construction, online querying phase and a novel retrieval problems formulations. The technical contributions are related to large vocabularies, their construction and learning.

In the online querying phase, we propose a method for increasing precision by improving spatial verification, increasing recall by adding context to the query expansion and automatic recovery from failure caused by *confuser* features. All these improvements were evaluated on standard datasets and protocols and state-of-the-art results were achieved.

Finally, novel problems for image retrieval are formulated. The user is able to ask new types of queries which, despite of having no analogy in text-retrieval world, might be very useful in many situations.

Large vocabularies The size of the visual vocabulary is one of the main parameters of bag-of-word retrieval system, which influences the precision, the recall and the speed of the retrieval system. We build and test large vocabularies disproving the common assumption which is present in community that is not worth to build vocabularies larger than 1 million visual words. We present a method that enables to build large vocabularies efficiently and without disadvantages of deep tree structures. We propose to keep the vocabulary structure balanced by adapting branching factor and shallow tree structure [MPCM10].

Learning a fine vocabulary We propose a novel similarity measure that is learned in an unsupervised manner, requires no extra space (only O(1)) in comparison with the standard systems using a bag-of-words. It is more discriminative than both 0–∞ and L2-based soft assignment and increases precision of the system [MPCM10, MPCM13].

Incremental spatial verification After querying the index file, images in the shortlist are spatially verified. In this work we show that by extending the query model with every spatially verified image during the process, we improve precision of the system and decrease the number of false positives. This way, images in the shortlist are verified against the query and the already verified images which appeared in the shortlist with a higher rank. In the text we refer to this contribution as to *incremental spatial verification* or *iSP* [CMPM11].

Context-based query expansion This contribution is based on the incremental spatial verification approach. We show that during the spatial verification, the consistent context of the query image can be learned and used to greatly improve the recall of the system. In this work we refer to this approach as to context-based query expansion or in short *ctxQE* [CMPM11].

Automatic failure recovery Another contribution which is based on spatial verification. If the spatial verification fails to find the geometric transformation between the query and retrieved images despite of the high number of matched visual words it is

probably caused by a high number of *confuser* features (features breaking the assumption of occurring independently). In our approach, confusers are identified, extracted from the original query, and a new query is issued. These steps increase the chances of retrieving true positive images from the database [CMPM11].

New retrieval problems: zoom-in, zoom-out, focus of the crowds We formulated new retrieval problems and new types of queries. We show how to insert a new functionality by replacing the ranking function, adjustment of query expansion and changes in scoring mechanism. We demonstrate this by adding a zoom-in and zoom-out capability to the system which instead of the most similar images retrieves the most zoomed images of the selected scene. We show how this new functionality can be used to answer users questions like "What is this?", 'Where is this?" or using the information of the crowds even a question "What is interesting here?" [MCM13].

1.7 Outline of the Work

The outline of this document is as follows. In the next chapter, various state-of-the-art methods for image retrieval are described. Different image representations and their properties, different vocabulary learning techniques and word assignment, as well as details and different ways of query expansion – the last stage of image search.

A novel method for vocabulary learning together with achieved results is proposed in Chapter 3. Two extensions for query expansion are proposed in Chapters 4 and 5. Novel ranking function and their use are shown in Chapter 6. The last chapter concludes the work.

1.8 Authorship

I hereby certify that the results presented in this thesis were achieved during my own research in cooperation with my thesis supervisor Jiří Matas and co-supervisor Ondřej Chum.

1.9 Used Terms and Abbreviations

Abbreviation	Description
AP	Average Precision – the measure of the quality indicator of the ranked result set (defined in Section 1.3).
BoW	Bag-of-words – an image representation (Sec. 2.1.1).
confusers	Visual words which confuse the retrieval system and cause a query failure – too many false positives retrieved. Confusers are features breaking the assumptions of independent occurrence (Sec. 5).
ctxQE	context-based Query Expansion – a variant of query expansion based on incremental spatial verification proposed in Chapter 3.
DAAT	Document at a time – a method for processing the inverted file proposed by Stewenius *et al.* [SGP12].
inverted file	The index file of the image retrieval engine. Contains a posting list for every visual word.
iSP	incremental Spatial Verification – variant of spatial verification proposed in Chapter 5.
mAP	mean Average Precision – the metric of the image retrieval performance (defined in Section 1.3).
posting list	One line of inverted file – a list of all documents in the database containing a particular visual word.
QE	Query Expansion – a step in the query phase that exploits result images to create a new expanded query (see Alg, 2).
shortlist	Top ranked images in the result set. These images are further processed – spatially verified, re-ordered and returned to the user.
SIFT	Scale-Invariant Feature Transform – a widely used local feature descriptor proposed by Lowe [Low04].
SP	Spatial Verification – The verification process checking the spatial alignment of local features in two images. It is applied between the query and each retrieved image in a shortlist (see Alg. 2).
tf-idf	term frequency - inverse document frequency – a weighting factor used for document scoring. It is intended to reflect importance of the visual word based on its frequency statistics [BYRN99].

Chapter 2

State of the Art

Virtually all aspects of specific object BoW-type retrieval have been intensively studied: feature detectors and descriptors [Low04, BTVG06, WHB09, MS04, MTS+05], vocabulary construction [SZ03, NS06, PCI+07, JDS08], spatial verification and re-ranking [PCI+07, JDS08], document metric learning [JHS07, JDS09, CM10b] and dimensionality reduction [JDSP10, PLSP10].

At the beginning of this chapter we present components of image retrieval pipelines in state-of-the-art systems. Later in the chapter we review in a greater detail recent approaches to three sub-problems of content based image search. First, vocabulary learning, which is one of the main and also one of the most time consuming part of the preparation of the system running off-line, is reviewed in detail. Next, the spatial verification and query expansion is described. These are the parts of the search system, which we are improving in next chapters.

2.1 Standard components 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.

2.1.1 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. Histogram of visual words is called bag-of-words.

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 [MS02, SZ02], Hessian-affine [MS02] and MSER [MCUP02], 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.* [MTS+05], which confirms the high performance of the above listed detectors.

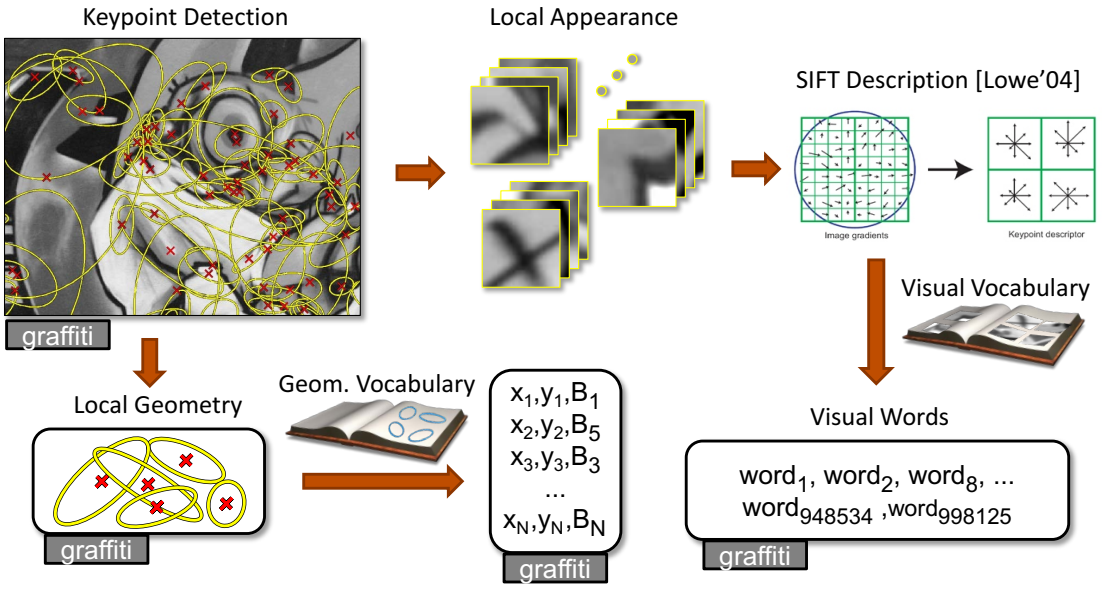


Figure 2.1: Visualization of the bag-of-words image representation computation with geometry compression. Courtesy of Michal Perďoch.

Detected interest regions are described by a feature descriptor. The SIFT descriptor [Low04], 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 [AZ12] and SURF [BTVG06].

Feature descriptors are vector quantized into visual words [SZ03] 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 tasks, up to $k \approx 10^7$ as we show in chapter 3. Different approaches and modifications of unsupervised k-means clustering is given in Section 2.2.

The process of image description is visualized in Figure 2.1.

2.1.2 Image representation with VLAD

The vector of locally aggregated descriptors (VLAD) [JDSP10] is another successful image representation method. It combines the advantages of the bag-of-words and the Fisher kernel [JH99]. 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.* [JDSP10] 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 [JDS11] is used to construct the final quantized descriptor creating a compact representation that fits into 20 bytes.

2.1.3 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 [OT06a]. 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.

2.1.4 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 [ALR03, BDH03] 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 [SZ03]. Inverted files map visual words to documents containing these 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 [BYRN99], 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.* [JDS11].

2.1.5 Spatial verification and query expansion

As shown in [PCI$^+$07, PCM09], 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

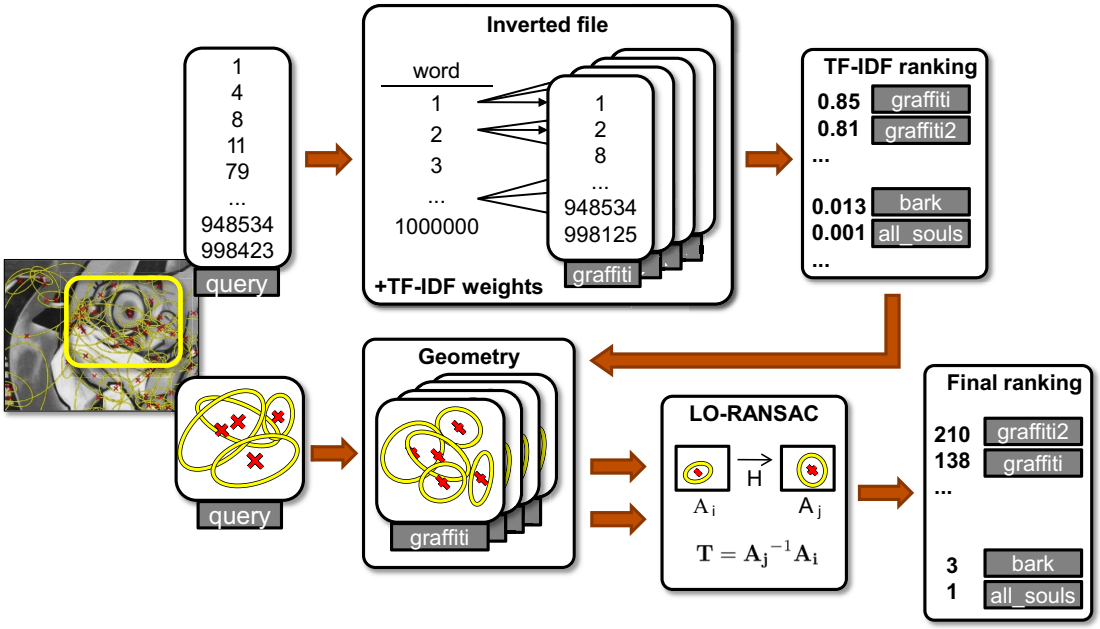


Figure 2.2: Visualization of image retrieval with spatial verification for the bag-of-words representation. Courtesy of Michal Perdoch.

with local optimization [CMK03] 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. Verified images in the shortlist are subsequently re-ranked. Chapter 4, shows 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.

Chum *et al.* [CPS+07] proposed a query expansion (QE) method – another technique inspired by text retrieval [BSAS95, SB97] – 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.

In Chapter 4 we 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 2.2.

2.2 Vocabulary Learning

As we mentioned in previous section bag-of-words was introduced to image retrieval by Sivic *et al.* [SZ03]. They represented the image by a histogram of 'visual words', *i.e.* discretized SIFT descriptors [Low04]. The BoW representation possesses many desirable properties required in large scale retrieval. If represented as an inverted file, it is compact and supports fast search. It is sufficiently discriminative and yet robust to acquisition 'nuisance parameters' like illumination and viewpoint change as well as occlusion. In this section we consider and compare methods that support queries that cover only a (small) part of the test image. Global methods like GIST [OT06b] or VLAD [JDSP10] achieve a much smaller memory footprint at the cost of allowing whole image queries only.

The discretization of the SIFT features is necessary in large scale problems as it is neither possible to compute distances on descriptors efficiently nor feasible to store all the descriptors. Instead, only (the identifier of) the vector quantized prototype for visual word is kept. After quantization, Euclidean distance in a high (128) dimensional space is approximated by a 0–∞ metric - features represented by the same visual word are deemed identical, else they are treated as 'totally different'. The computational convenience of such a crude approximation of the SIFT distance has a detrimental impact on discriminative power of the representation. Recent methods like soft assignment and in particular the Hamming embedding aim at obtaining a better space-speed-accuracy trade off.

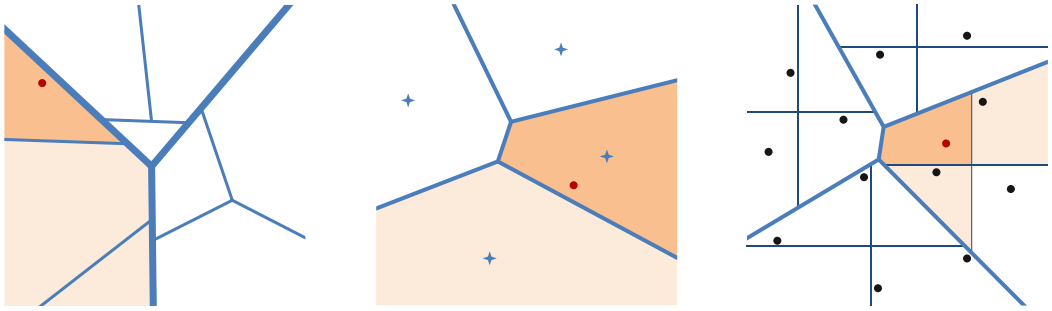
In this section, approaches to vocabulary construction and soft assignment suitable for large-scale image search are reviewed and compared.

In [SZ03], the vocabulary (the number of visual words $\approx 10^4$) was constructed using a standard k-means algorithm. Adopting methodology from text retrieval applications, the image score is efficiently computed by traversing inverted files related to visual words present in the query. The inverted file related to a visual word W is a list of image ids that contain the visual word W. It follows that the time required for scoring the documents is proportional to the number different visual words in a query and the average length of an inverted file.

Hierarchical clustering

The hierarchical k-means and scoring of Nistér and Stewenius [NS06] is the first image retrieval approach that scales up. The vocabulary has a hierarchical structure which allows efficient construction of large and discriminative vocabularies. The quantization effect are alleviated by the so called hierarchical scoring. In such a type of scoring, the scoring visual words are not only stored in the leafs of the vocabulary tree. The non-leaf nodes can be thought of as virtual or generic visual words. These virtual words naturally score with lower *idf* weights as more features are assigned to them (all features in their sub-tree).

The advantage of the hierarchical scoring approach is that the soft assignment is given by the structure of the tree and no additional information needs to be stored for each feature. On the downside, experiments in [PCI$^+$08] show that the quantization artefacts of the hierarchical k-means are not fully removed by hierarchical scoring, the



(a) (b) (c)

Figure 2.3: Different approaches to the soft assignment (saturation encodes the relevance): (a) hierarchical scoring [NS06] – the soft assignment is given by the hierarchical structure of the search tree; (b) soft clustering [PCI$^+$08] assigns features to the r nearest cluster centers; (c) Hamming embedding [JDS10] – each cell is divided into orthants by a number of axis-aligned hyperplanes, the distance of the orthants is measured by the number of separating hyperplanes.

problems are only shifted up a few levels in the hierarchy. An illustrative example of the soft assignment performed by the hierarchical clustering is shown in Fig. 2.3(a).

Lost in quantization

In [PCI$^+$08], an approximate soft assignment is exploited. Each feature is assigned to $n = 3$ (approximately) nearest visual words. Each assignment is weighted by $e^{-\frac{d^2}{2\sigma^2}}$ where d is the distance of the feature descriptor to the cluster center.

The soft assignment is performed on features in the database as well as the query features. This results in n times higher memory requirements and n^2 times longer running time – the average length of the inverted file is n times longer and there are up to n times more visual words associated with the query features. For an illustration of the soft assignment, see Fig. 2.3(b).

Hamming embedding

Jégou *et al.* [JDS10] have proposed to combine k-means quantization and binary vector signatures. First, the feature space is divided into relatively small number of Voronoi cells (20K) using k-means. Each cell is then divided by n independent hyper-planes into 2^n subcells. Each subcell is described by a binary vector of length n. Results reported in [JDS10] suggest that the Hamming embedding provides good quantization. The good results are traded off with higher running time requirements and high memory requirements.

The higher running time requirements are caused by the use of coarse quantization in the first step. The average length of an inverted file for vocabulary of 20K words is approximately 50 times longer than the one of 1M words. Recall that the time required to traverse the inverted files is given by the length of the inverted file. Hence 50 times smaller vocabulary results in 50 times longer scoring time on average. Even if two

query features are assigned to the same visual word, the relevant inverted file has to be processed for each of the features separately as they will have different binary signature.

While the reported bits per feature required in the search index ranges from 11 bits [PCM09] to 18 bits [PCI$^+$08], Hamming embedding adds another 64 bits. The additional information reduces the number of features that can be stored in the memory by a factor of 6.8.

2.3 Spatial Verification

As shown in [PCI$^+$07, PCM09], the results can be significantly improved using the feature layout to verify the consistency of the retrieved images with the query region. The initially returned result list is re-ranked by estimating an affine transformation between the query image and result image. However, the spatial verification is significantly more time consuming than the BoW scoring, and is performed only on a shortlist of top scoring images. The shortlist is subsequently re-ranked based on the number of spatially verified inliers.

2.4 Query Expansion

In the text retrieval literature, one of the standard methods for improving performance is query expansion. A number of the highly ranked documents from the original query are re-issued as a new query. In this way, additional relevant terms can be added to the query.

In [CPS$^+$07], the authors brought query expansion into the visual domain. A strong spatial constraint between the query image and each result enables an accurate verification for each return, resulting in a suppression of false positives that typically ruin text-based query expansion. These verified images can be used to learn a latent feature model to enable controlled construction of expanded queries.

In [CPS$^+$07], the authors proposed a number of query expansion strategies. All of them follow a similar pattern: images in a shortlist are spatially verified against the query features, images with sufficient numbers of matches (inliers) are back-projected by the estimated affine transformation into the query region, and, finally, a new query is issued. The differences in the proposed strategies are either in the number of repeated applications of the process, or in the method of feature selection.

The simplest well performing query expansion method is called average query expansion. A new query is constructed by averaging a number of document descriptors. This approach is the quickest from all the suggested strategies, and has been adopted in a number of publications [PCI$^+$08, PCM09, JDS09]. We use the average query expansion as the baseline method.

2.5 Commercial solutions

In the last few years commercial image search engines became available to the public on the Internet. The best known are Google's Web image search and Google Goggles

(for android phones) started in 2010, Bing image search from Microsoft launched for public in 2012 and the TinEye started in 2008.

There are no recent publication to our knowledge describing the back-ends of these systems or evaluating their performance on standard datasets. From the user experience our guess is that the bag-of-words approach was used in combination with textual information which often accompany the images – image name, anchor-text, tags or the text in the body of the page around image. Some engines are able to detect bag-of-words failure for a particular search and use different methods such as color histograms or a text search if text was detected on the image.

Figures 2.4 to 2.7 display mentioned the interfaces with results of the above-mentioned search engines as well as our demo engine called CMP::G2 (fig. 2.8) querying the Gothic tympanum on St. Vitus Cathedral in Prague.

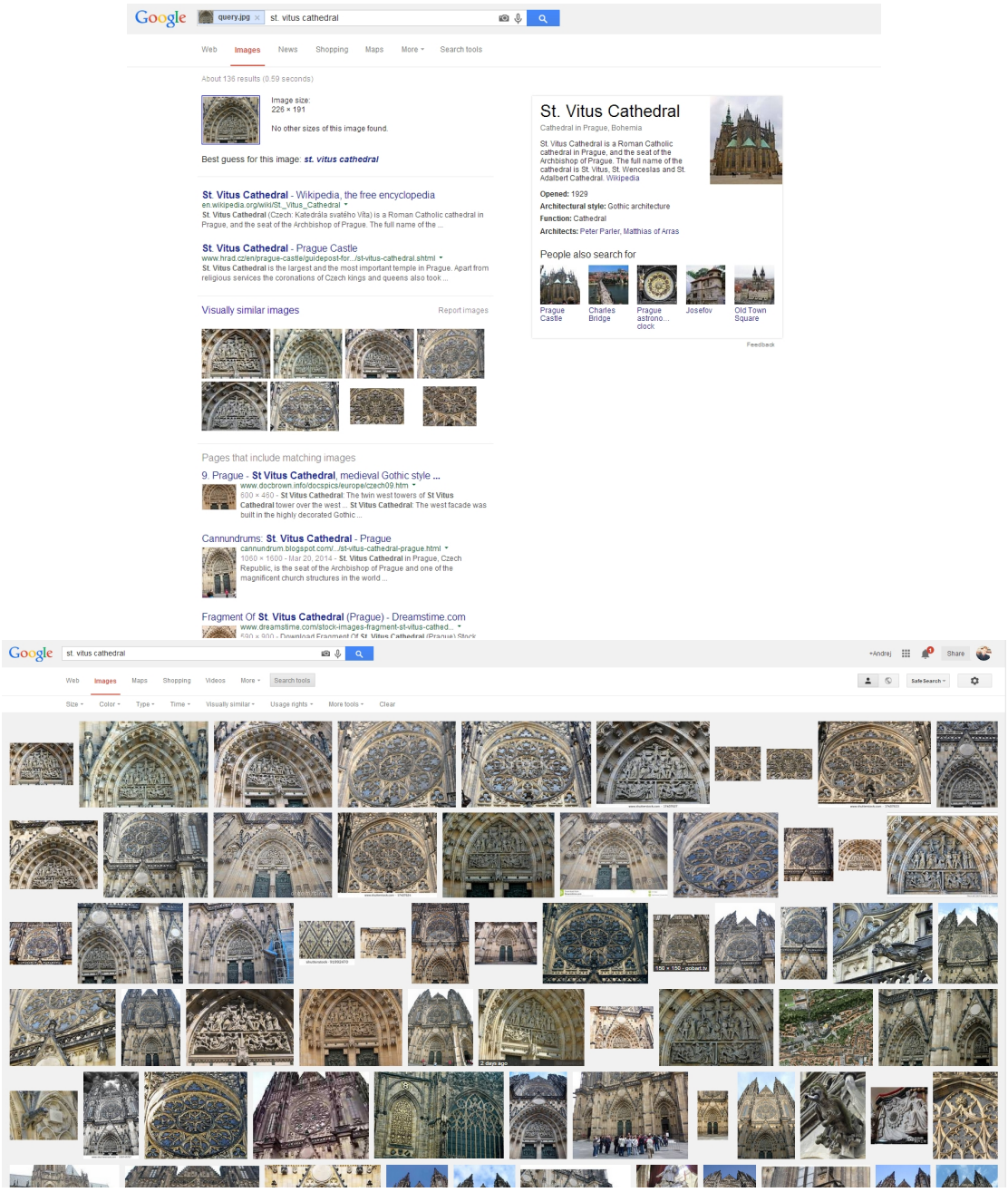
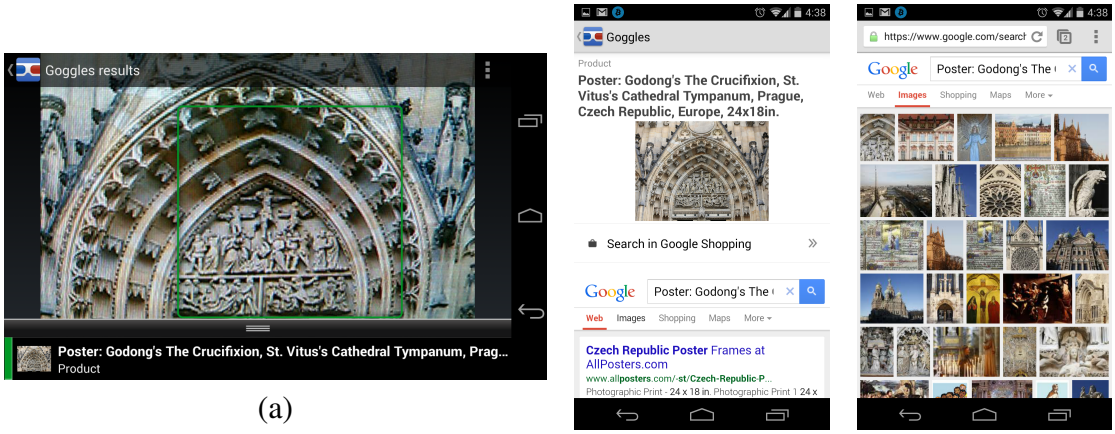


Figure 2.4: **The Google Images** application successfully identifies the building and offers webpages about St. Vitus Cathedral, a summary taken from Wikipedia as well as similar images after choosing ”Visually similar images” link.



(a) (b) (c)

Figure 2.5: **The Google Goggles** application identifies the image as a poster "Godong's The Crucifixion, St. Vitus's Cathedral Tympanum" (a), which is more or less correct but then navigates the user to webpage selling posters (b), which would not be expected in this case. Choosing the image search issues probably a textual query according to the name of the poster not giving the best results either (c).

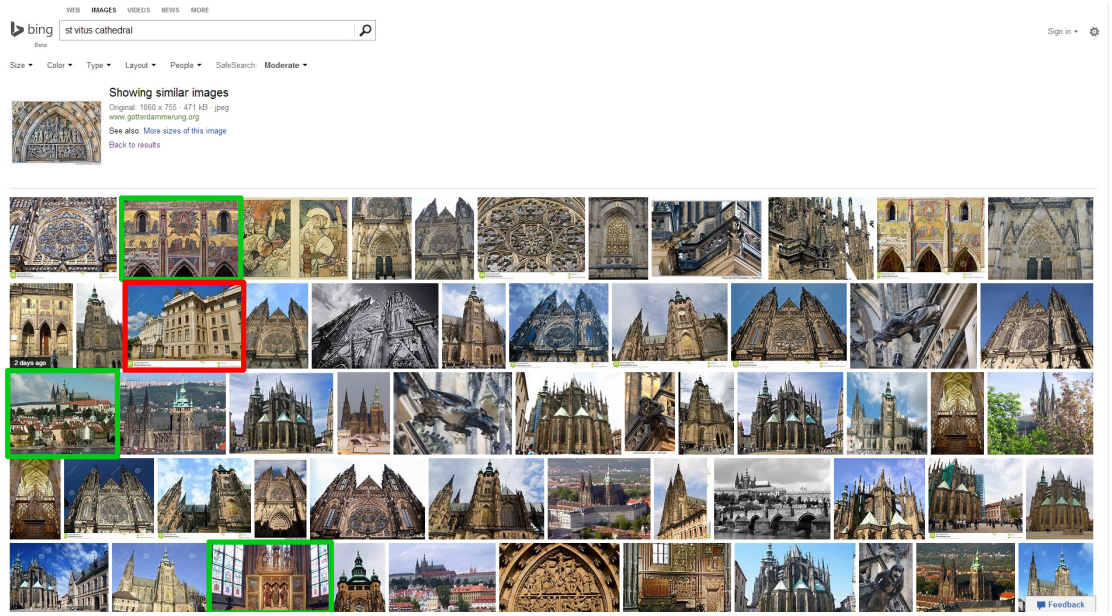


Figure 2.6: **The Bing Images** application does not allow the user to select his own image *i.e.* an image not contained in Bing's dataset. We found a similar query image in the indexed database and run a query. The most of the results are correct (one false positive is highlighted red). It is obvious that images are linked not only by the visual content but various textual information is used as well – consider examples highlighted green.

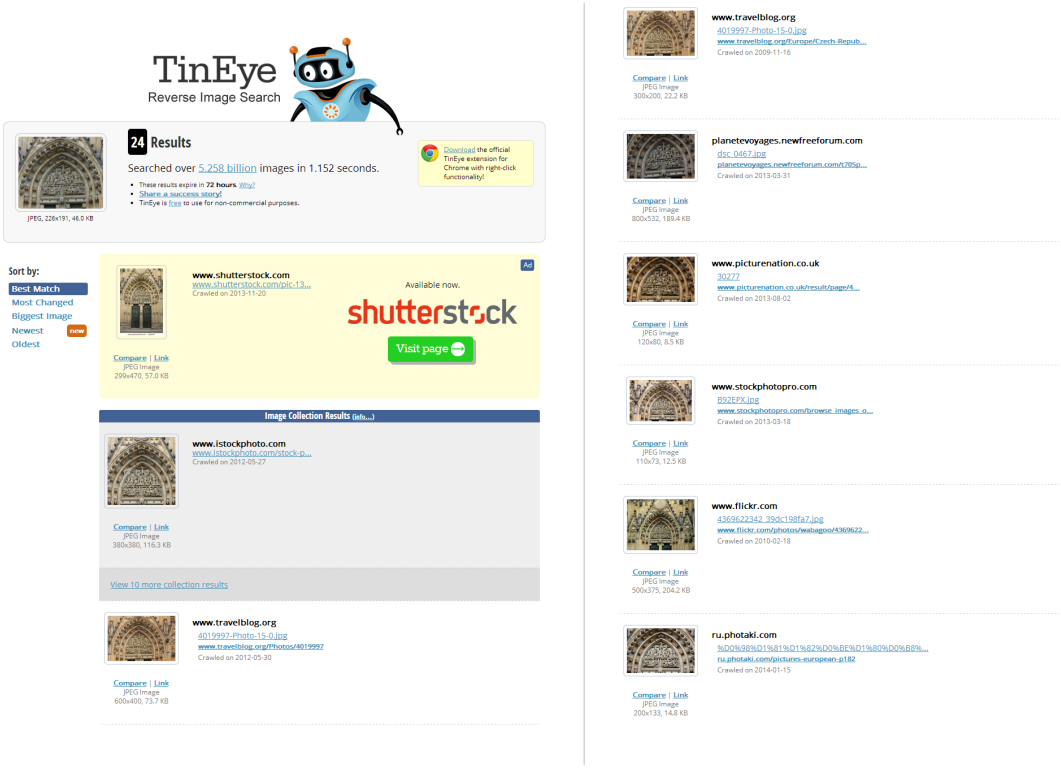


Figure 2.7: **TinEye** is the first publicly available content based search engine with over 5.283 billion images indexed. Probably due to the effort of avoiding false positives only 24 images were retrieved. All of them are correct. The results support the argument of Chapter 6 that displaying the most similar images is not always useful.

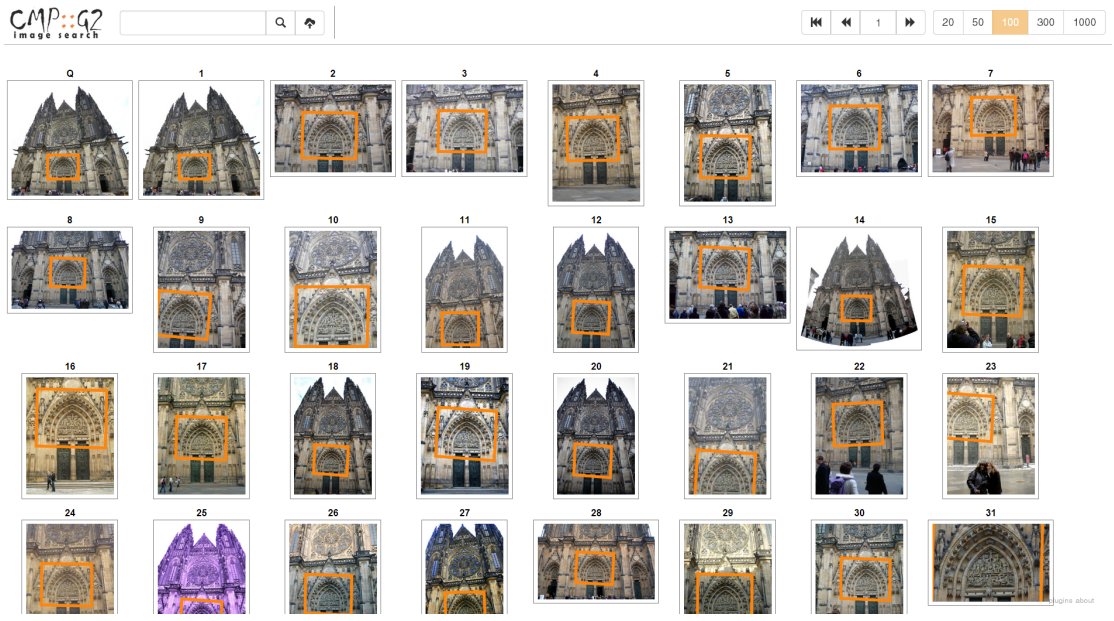


Figure 2.8: **CMP::G2 Image Search** is the demo application implementing the methods and algorithms described in the thesis. Displayed here for comparison.

Chapter 3

Learning a Fine Vocabulary

Quantization of the descriptor space (construction of vocabulary) brings essential retrieval speed-up and lower memory footprint for the image description. Parameters of a visual vocabulary have major impact on the performance of the retrieval engine, and are thoughtfully studied and addressed in literature. One of the main parameter is the size of the vocabulary – number of quantization cells (visual words). A larger vocabulary yields more false negatives (lower recall), but higher precision and faster retrieval. The smaller the vocabulary the slower and more prone to false positives (lower precision).

Features from the two different physical pre-images have lower probability of being assigned into the same quantization cell (visual word) in a larger vocabulary. On the other hand larger vocabulary have bigger problem with *quantization effects* – the higher probability of two similar features from the same pre-image assigned to different visual word. If hard-assignment (0–∞ metric) is used, such features are considered to be completely distant. If this probability is not too high, it is not a big problem assuming that each image is described by hundreds of visual words. Otherwise, different types of soft-assignment are used to deal with these effects.

Other important property of the visual vocabulary influencing query speed is *imbalance factor* [JDS10]. The balanced inverted file (*i.e.* posting lists of about the same length) is essential for small variations from expected retrieval time and good user experience.

In this chapter we propose a novel vocabulary construction method for very large vocabularies, which according to experiments achieves the state-of-the-art mAP results while keeping imbalance factor low.

3.1 Motivation

All approaches to soft-assignment mentioned in Chapter 2 are based on the distance (or its approximation) in the descriptor (SIFT) space. It has been observed that the Euclidean distance is not the best performing measure. Learning a global Mahalanobis distance [HBW07, MM07] showed that the matching is improved and/or the dimensionality of the descriptor is reduced. However, even in the original work on SIFT descriptor matching [Low04] it is shown that the similarity of the descriptors is not only dependent on the distance of the descriptors, but also on the location of the features in the feature space. Therefore, learning a global Mahalanobis metric is sub-optimal and a local simi-

larity measure is required. For examples of corresponding patches where SIFT distance does not predict well the similarity see Figures 3.1, 3.6, and 3.7. This is mainly due to the fact that SIFT distribution of the features detected on the same physical pre-image is not ellipsoid.

In this chapter, unsupervised learning on a large set of images is exploited to improve on the hartd-assignment – the 0–∞ metric. First, an efficient clustering process with spatial verification establishes correspondences within a large (>5M) image collection. Next, a fine-grained vocabulary is obtained by 2-level hierarchical approximate nearest neighbour clustering. The automatically established correspondences are then used to define a similarity measure on the basis of a probabilistic relationships of visual words; we call it the *PR visual word similarity*.

When combined with a large vocabulary, several millions of words (one or two orders of magnitude larger than commonly used), the PR similarity has the following desirable properties:

(i) it is more accurate (discriminative), than both standard 0–∞ metric and Hamming embedding.

(ii) the memory footprint of the image representation for PR similarity calculation is roughly identical to the standard method and smaller than that of Hamming embedding.

(iii) search with the PR similarity is faster than the standard bag-of-words.

A novel similarity measure presented in this chapter is learned in an unsupervised manner, requires no extra space regard to size of the database (only O(1)) in comparison with the bag-of-words.

Further, we experimentally disprove the common assumption which is present in community that is not worth to build vocabularies larger than 1M. To construct a well performing large vocabulary, we propose to build shallow hierarchical – tree based – vocabularies with adaptive branching to speed up the process but not to bring the disadvantage of large imbalance factor of deeper ones.

Figure 3.1: An example of corresponding patches. A 2D PCA projection of the SIFT descriptors (left); two most distant patches in the SIFT space and the images where they were detected (right); a set of sample patches (bottom). The average SIFT distance within the cluster is 278, the maximal distance is 591.

3.2 The Probabilistic Relation Similarity Measure

Consider a feature in the query image with descriptor $D \in \mathcal{D} \subset R^d$. For most accurate matching, the query feature should be compared to all features in the database. The contribution of the query feature to the matching score should be proportional to the probability of matching the database feature. It is far too slow, *i.e.*, practically not feasible, to directly match a query feature to all features in a (large) database. Also, the contribution of features with low probability of matching is negligible.

The success of fast retrieval approaches is based on efficient separation of (potentially) matching features from those that are highly unlikely to match. The elimination is based on a simple idea – the descriptors of matching patches will be close in some appropriate metric (L2 is often used). With appropriate data structures, enumeration of descriptors in proximity is possible in time sub-linear in the size of the database. All bag-of-words based methods use partition $\{w_i\}$ of the descriptor space $\mathcal{D}$: $\bigcup w_i = \mathcal{D}, \quad w_i \cap w_{j \neq i} = \emptyset$. The cells are then used to separate features that are close (potentially matching) from those that are far (non-matching).

In the case of hard assignment, features are associated with the visual words defined by the closest cluster center. In the scoring that evaluates query and database image match, only features with the same visual word as the query feature are considered.

We argue that the descriptor distance is a good indicator of patch similarity only up to a limited distance, where the variation in the descriptors is caused mostly by the imaging and detector noise. We abandon the assumption that the descriptor distance provides a good similarity measure of patches observed under different viewing angles or under different illumination conditions. Instead, we propose to estimate the probability between a feature observed in the query image and a database feature. Since our aim is to address retrieval in web-scale databases where store requirements are critical, we constrained our attention to solution that have a minimal overhead in comparison with the standard inverted file representation.

The Proposed Approach

We propose to use a fine partition of the descriptor space, to minimize a probability of false match inside a single cell. Even though the fine partition is learned in a data dependent fashion (as in the other approaches), the fine partition unavoidable separates matching features into a number of cells.

For each cell (visual word) we learn which other cells (called *alternative visual words*) are likely to contain descriptors of matching patches with the same pre-images (Fig 3.2). This step consist of estimating the probability of observing visual word w_j in a matching database image when visual word w_q was observed in the query image

$$P(w_j|w_q). \tag{3.1}$$

The probability (Eqn. 3.1) is estimated from a large number of matching patches.

A simple generative model, independent for each feature, is adopted. In the model, image features are assumed to be (locally affine) projections of a (locally close to planar) 3D surface patches z_i. Hence, matching features among different images are those

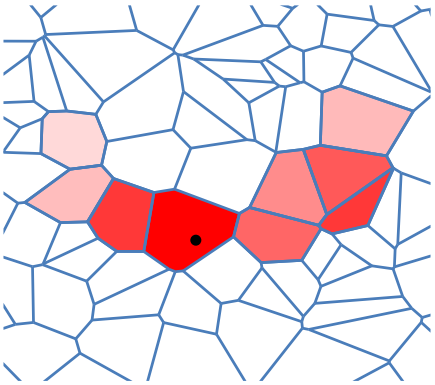


Figure 3.2: The set of alternative words in the proposed PR similarity measure.

that have the same pre-image z_i. To estimate the probability $P(w_j|w_q)$ we start with (a large number of) sets of matching features, each set being different projections of a patch z_i. Using the fine vocabulary (partition) the sets of matching features are converted to sets of matching visual words. We estimate the probability $P(w_j|w_q)$ as

$$P(w_j|w_q) \approx \sum_z P(w_j|z_i)P(z_i|w_q). \tag{3.2}$$

For each visual word w_q, a fixed number of alternative visual words that have the highest conditional probability (Eqn. 3.2) is recorded.

3.3 Learning a PR similarity

The first step of our approach is to obtain a large number of matching image patches. The links between matching patches are consequently used to infer relationship, between quantized descriptors of those patches, *i.e.*, between visual words. As a first step towards unsupervised collection of matching image patches, called "feature tracks", clusters of matching images are discovered. Within each cluster, feature tracks are found by a wide-baseline matching method. This approach is similar to [ASS+09], where the feature tracks are used to produce 3D reconstruction. In our case, it is important to find larger variety of patch appearances than precise point locations. Therefore, we adopt a slightly different approach to the choice of image pairs investigated.

3.3.1 Image Clusters

The algorithm starts with analyzing connected components of the image matching graph (graph with images as vertices, edges connect images that can be matched) produced by a large-scale clustering method [CM10a, LWZ+08]. Any matching technique is suitable provided it can find clusters of matching images in a very large database. In our case, an image retrieval system was used to produce the clusters of spatially related images. The following structure of image clusters is created. Each cluster of spatially related images is represented as an oriented tree structure (the skeleton of the cluster). The children of each parental node were obtained as results of an image retrieval using

the parent image as a query image. Retrieved images, which are already in the cluster, are ignored. Together with the tree structure, an affine transformation (approximately) mapping child image to its parent are recorded. These mappings are later used to guide (speed-up) the matching.

3.3.2 Feature Tracks

To avoid any kind of bias (by quantization errors, for example), instead of using vector quantized form of the descriptors, the conventional image matching (based on the full SIFT [Low04]) has to be used. In principle, one can go back even to the pixel level [FTVG04, CMP08], however such an approach seems to be impractical for large volumes of data.

It is not feasible to match all pairs of images in the image clusters, especially not of clusters with large number of images (say more than 1000). It is also not possible to simply follow the tree structure of image clusters because not all features are detected in all images (in fact, only a relatively small portion of features is actually repeated). The following procedure, that is linear in the number of images in the cluster, is adopted for detection of feature tracks that would exhibit as large variety of patch appearances as possible. For each parental node, a sub-tree of height two is selected. On images in the sub-tree, a $2k$-connected graph called circulant graph [GR01] is constructed. Vertices of a graph are ordered and connected with K steps of the length random chosen between 1 and $\lfloor (N-1)/2 \rfloor$ but always including step 1, to force connectivity. (*i.e.*. for chosen step 4, the edges are created between vertices $v_i, v_j \in V$, where $i - j \mod N = 4$). The algorithm for construction of minimal $2k$-connected graph is summarized in Algorithm 1.

Images connected by an edge in such a graph are then matched using standard wide-baseline matching. Since each image in the image cluster participates in at most 3 sub-trees (as father, son and grand-son), the number of edges is limited to $6kN$, where N is the size of the cluster. Instead of using epipolar geometry as a global model, a number of close-to-planar (geometrically consistent) structures is estimated (using affine homography). Unlike the epipolar constraint, such a one-to-one mapping enables to verify the shape of the feature patch. Connected components of matching and geometrically consistent features are called *feature tracks*.

Tracks that contain two different features from a single image are called inconsistent [ASS+09]. These features clearly cannot have a single pre-image under perspective projection and hence cannot be used in the process of 3D reconstruction. Such inconsistent tracks are often caused by repeated patterns. Inconsistent feature tracks are (unlike in [ASS+09]) kept as they provide further examples of patch appearance.

3.3.3 Computing the conditional probability.

To compute the conditional probability (eqn. 3.2) from the feature tracks, an inverted file structure is used. The tracks are represented as forward files (named z_i), *i.e.*, lists of matching SIFT descriptors. The descriptors are assigned to their visual word from the large vocabulary. Then, for each visual word w_k, a list of patches z_i so that $P(z_i|w_k) >$

Algorithm 3 Construction of the 2K connected graph with a minimal number of edges as a union of circulants.

Input: K - requested connectivity, N - number of vertices
Output: V a set of vertices, $E \subset V \times V$ a set of edges of $2K$ connected graph (V,E).

1. **if** $2K \geq N - 1$ **then**
   return fully connected graph with N vertices.
   end
2. $S := \{1\} \bigcup$ a random subset of $\{2, \ldots, \lfloor \frac{N-1}{2} \rfloor\}$,$|S| = K$
3. $V := \{v_0, \ldots, v_{N-1}\}$
4. $E := \{(v_i, v_j) \,|\, v_i, v_j \in V, i - j \bmod N \in S\}$

0 (the inverted file) is constructed. The sum (eqn. 3.2) is evaluated by traversing the relevant inverted file.

3.3.4 Statistics

Over 5 million images were processed using geometric min-hash technique [CPM09]. Almost 20,000 clusters containing 750,000 images were found. Out of those 733,000 were successfully matched in the wide-baseline matching stage. Over 111 million of feature tracks were established, out of which 12.3 millions are composed of more than 5 features. In total, 564 million features participated in the tracks, 319.5 million features belong to tracks of more than 5 features. Some examples of feature tracks are shown in Figures 3.8 and 3.9. Only negligible portion of visual words were not present in any feature track. There was 2 such words in the 1M vocabulary and 74005 (0.4%) in 16M vocabulary. The distribution of visual words over tracks in these vocabularies are shown in Figure 3.3.

3.3.5 Memory and time efficiency

For the alternative words storage, only constant space is required, equal to the size of the vocabulary times the number of alternative words. The pre-processing consists of image clustering ([CM10a] reports near linear time in the size of the database), intra-cluster matching (linearity enforced by the $2k$-connected circulant matching graph), and of the evaluation of expression eqn. (3.2) for all visual words. The worst case complexity of the last step is equal to the number of tracks (correspondences) times size of the vocabulary squared. In practice, due to the sparsity of the representation, the process took less than an hour in our settings for over 5 million images.

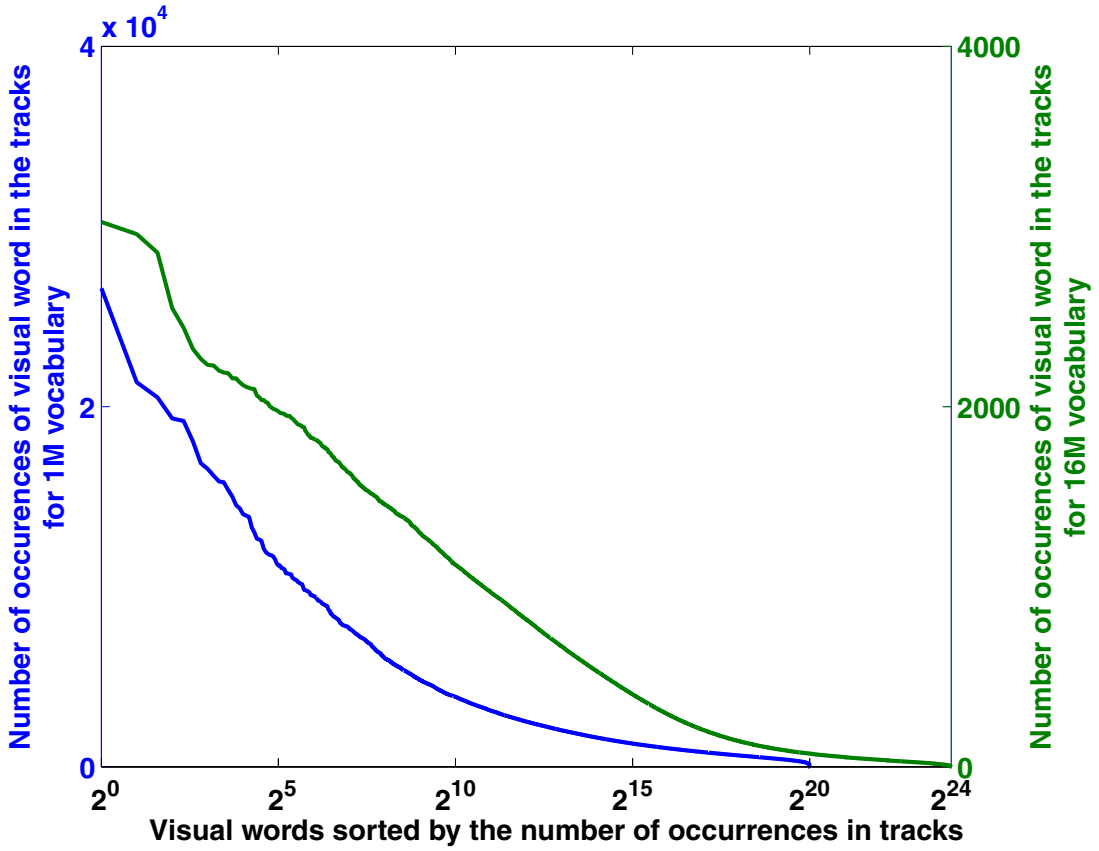


Figure 3.3: The distribution of visual words in tracks in the 1M and 16M vocabularies. Only a negligible part of the visual words were not present in any feature track.

3.4 Large Vocabulary Generation

To efficiently generate a large visual vocabulary we employ a hybrid approach – approximate hierarchical k-means. A hierarchy tree of two levels is constructed. For instance, for vocabulary of 16M words, each level has 4K nodes on average. In the assignment stage of k-means, an approximate nearest neighbour, FLANN [ML09], is used for efficiency reasons.

First, a level one approximate k-means is applied to a random sub-sample of 5 million SIFT descriptors. Then, a two pass procedure on ≈ 11 billion SIFTs (from almost 6 million images) is performed. In the first pass, each SIFT descriptor is assigned to a word in the level one of a vocabulary. For each visual word in the first level a list of descriptors assigned to it is recorded. In the second pass, approximate k-means on each list of the descriptors is applied. The whole procedure takes about one day on a cluster of 20 computers.

3.4.1 Balancing the Tree Structure

For the average speed of the retrieval, it is important that the vocabulary is balanced, *i.e.*, there are approximately the same number of instances of each visual word in the database.

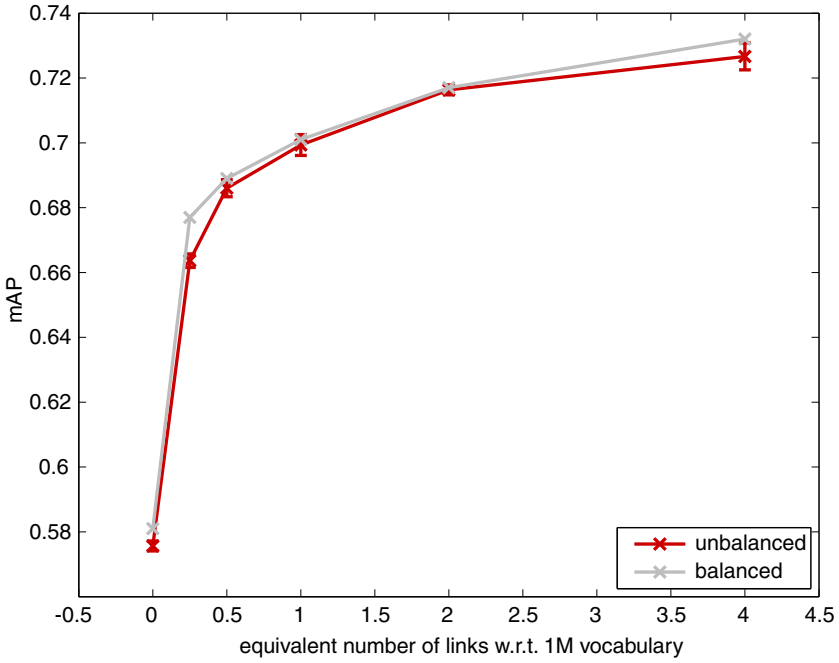
We compared unbalanced and balanced vocabulary constructions (Figure 3.4). In the balanced construction, the second level of the vocabulary uses an adaptive branching factor, which is proportional to the weight of the branch (*i.e.* cluster A with 2 times more features than cluster B will be split into two times more clusters in the second level of hierarchy than cluster B). We also explored the balancing on the first level by constraining the length of the mean vectors (this stems from the fact that SIFT features live approximately on a hyper-sphere), which is similar to the method [TAJ10]. As the latter method has not brought better results while implied higher computation costs, it was not explored further.

In our experiments, a balanced vocabulary with adaptive branching factor at the second level is used. With such a construction we reached an imbalance factor [JDS10] of 1.09 for the training image set (>5M images) (compared to 1.21 in [JDS10]) and 1.26 for the testing set – Oxford 105k. Fraundorfer *et al.* [FSN07] report estimate of imbalance factor 5 for hierarchical trees introduced in [NS06]. The experiment shows that the balancing does not significantly affect mAP. The advantage is the gain in query speed.

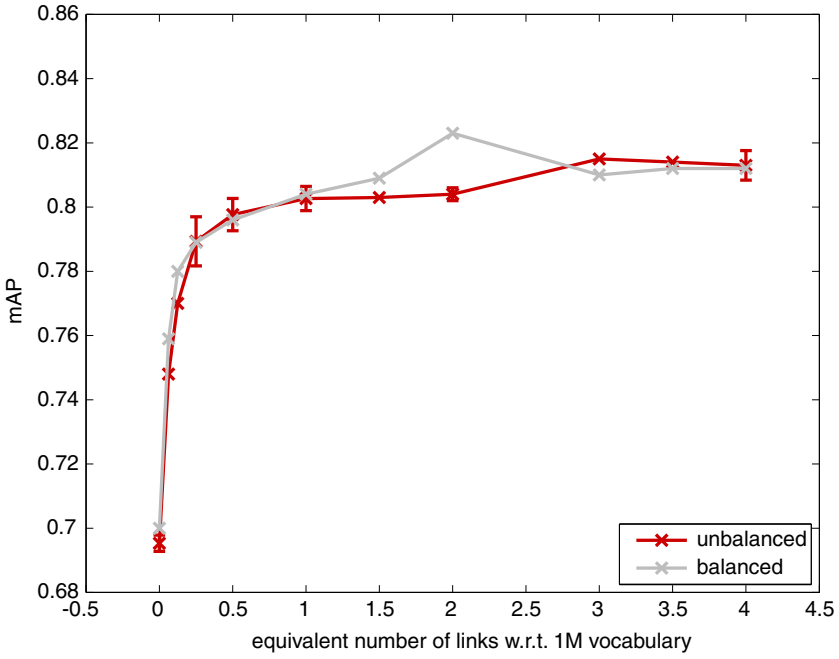
Comparison of the imbalance factors of our balanced and unbalanced vocabulary is show in Table 3.1.

3.4.2 Size of the Vocabulary

There are different opinions about the number of visual words in the vocabulary for image retrieval. Philbin *et al.* in [PCI+07] achieved the best mAP for object recognition with a vocabulary of 1M visual words and predict a performance drop for larger vocabularies. We attribute the result in [PCI+07] to a too small training dataset (16.7M descriptors). In our case the vocabularies with up to 64M words is built using 11G



(a)



(b)

Figure 3.4: A comparison of the mean average precision (mAP) for an unbalanced and balanced 16M vocabulary (a) with and (b) without the query expansion. The experiment shows that the balancing does not significantly affect mAP (the advantage is the gain in query speed). The error bars are shown where three vocabularies with different random initialization were evaluated.

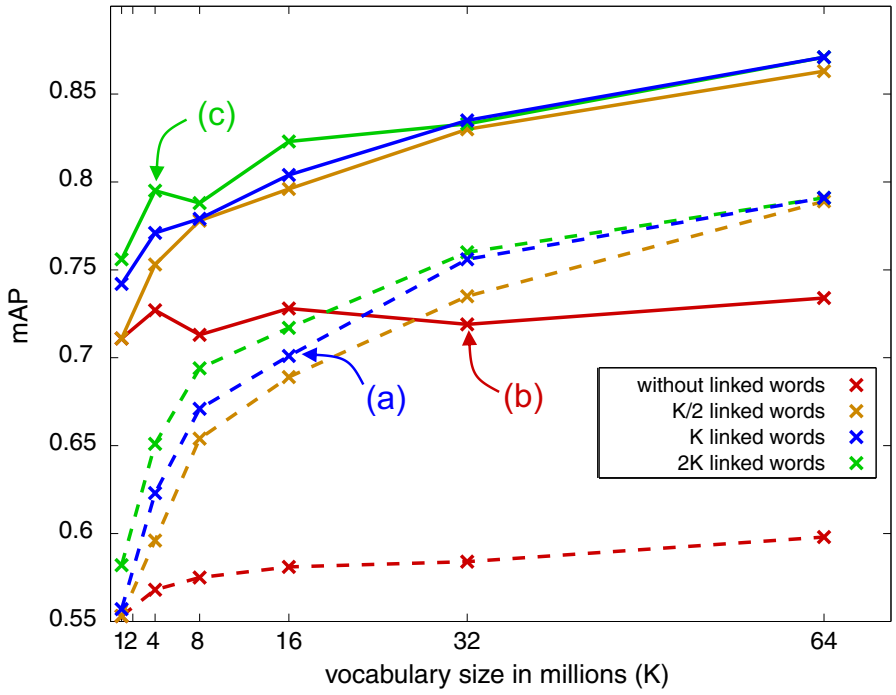


Figure 3.5: Comparison of mAP for the balanced vocabularies with 1 to 64 millions visual words. Solid lines show results after the query expansion (QE), dashed lines without QE. Red lines show results using plain bag-of words (no alternative words). The number of alternative words is proportional to the vocabulary size to compare results of equal time complexity. In this way, approximately the same number of entries of the inverted file is traversed, since the average length of a list of an inverted file for 16M vocabulary is 16 times smaller than for 1M vocabulary, 16 lists with alternative words can be crawled within the same time. To clarify the plot: (a) the result of 16M vocabulary with (L16) 16 linked words (1 original and 15 alternatives) and without QE. (b) 32M vocabulary (L1) without alternative words with QE, and finally (c) 4M vocabulary (L8) 8 linked words with QE.

training descriptors. Experiments show that the larger the vocabulary is, the better performance is achieved, even for plain bag-of-words retrieval.

Introducing the alternative words, the situation is changed even more rapidly and, as expected, they are more useful for larger vocabularies (Figure 3.5). We have not built vocabularies larger than 64M because the memory footprint of the assignment tree started to be impractical and the performance has almost converged.

3.5 Experiments

The implementation of the retrieval stage is fairly standard, using inverted files [SZ03] for candidate image selection which is followed by fast spatial verification and query expansion [CPS+07]. The modifications listed below are the major differences implemented in our retrieval stage.

method	imbalance factor level 1		imbalance factor level 2	
	training set	testing set	training set	testing set
unbalanced	1.028	1.097	1.122	1.311
balanced	1.028	1.097	1.093	1.259

Table 3.1: Comparison of the imbalance factor [JDS10] of the unbalanced and balanced versions of the two-level hierarchical vocabulary. An adaptive branching factor was used at the second level of the tree hierarchy to balance the vocabulary.

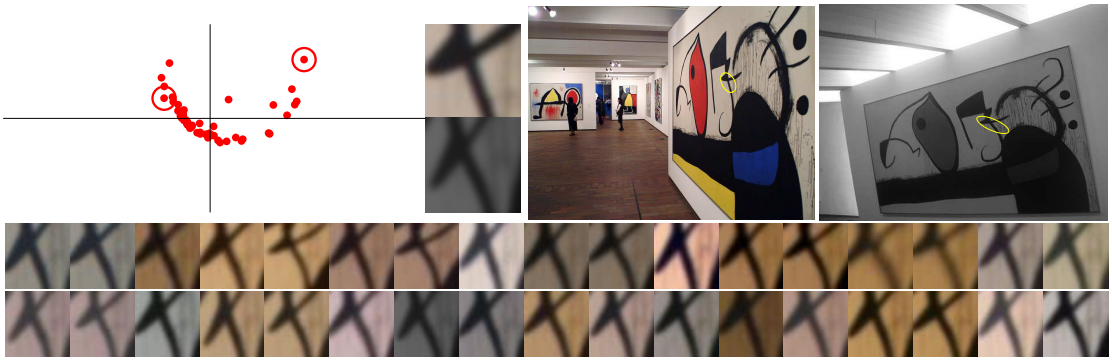


Figure 3.6: A 2D PCA projection of a feature track of SIFT descriptors (left); the most distant patches and their images (right); sample of feature patches from the track. The distance of the most distant SIFT descriptors is 542 and is caused by an enormous change in the viewpoint.

Unique Matching

Despite being assigned to more than one visual word, each query feature is a projection of a single physical patch. Thus it can match only at most one feature in each image in the database. We find that applying this uniqueness constraint adds negligible computational cost and improves the results by approximately 1%. The order in which are the alternative words traversed and matched in an inverted file is given by their probability of being an alternative word (3.2).

Weights of Alternative Words

Contribution of each visual word is weighted by the *idf* weight [BYRN99]. A number of re-weighting schemes for alternative words have been tried, none of them affecting significantly the results of the retrieval.

Datasets

We have extensively evaluated the performance of the PR similarity on a standard retrieval datasets Oxford buildings, INRIA Holidays and Paris buildings. The experiments focus on retrieval accuracy and the retrieval speed. Since our training set of 6 million images were downloaded from FLICKR in a similar way as the testing datasets Oxford and PARIS, we have explicitly removed all testing images (or their scaled duplicates) from the training set.

Figure 3.7: A 2D PCA projection of a feature track of SIFT descriptors (left); the most distant patches and their images (right); sample of feature patches from the track. The distance of the most distant SIFT descriptors is 593 and is caused by the viewpoint and scale change.

3.5.1 Retrieval Quality

We follow the protocols of testing datasets defined in [PCI+07] and use the mean average precision as a measure of retrieval performance. We start by studying the properties of the PR similarity for a visual vocabularies of 1, 4, 8, 16, 32 and 64 million words.

In the first experiment, the quality of the retrieval as a function of the number of alternative words is measured, see Figure 3.10. The plots show that performance improves for visual vocabularies of all tested sizes monotonically for plain retrieval without query expansion and almost monotonically when query expansion is used.

The second experiment studies the effects of the vocabulary size (Figure 3.5), and compares the alternative words in the PR similarity with the euclidean nearest neighbours in soft assignment. The left-hand part of Table 3.2 shows results obtained with the 16M vocabulary with three different settings 'L1' – standard tf-idf retrieval with hard assignment of visual words; 'L5' and 'L16' – retrieval using alternative words (4 and 15 respectively). The righthand part presents results of reference state-of-the-art results [PCM09] obtain with a vocabulary of 1M visual words learned on the PARIS dataset. Two version of the reference algorithm are tested, without ('L1') and with the query soft assignment to 3 nearest neighbours ('SA 3NN').
The experiments support the following observations:

(i) PR similarity calculation with using the learned alternative words increases significantly the accuracy of the retrieval, both with and without query expansion.

(ii) Alternative words are more useful for larger vocabularies

(iii) The PR similarity outperforms soft SA in term of precision, yet does not share the drawbacks of SA.

(iv) The PR similarity outperforms the Hamming embedding approach combined with query expansion, Jegou et al. [JDS09, JDS10] report the mAP of 0.692 on this dataset.

(v) The mAP result for 16M L16 is superior to any result published in the literature on the Oxford 105k dataset.

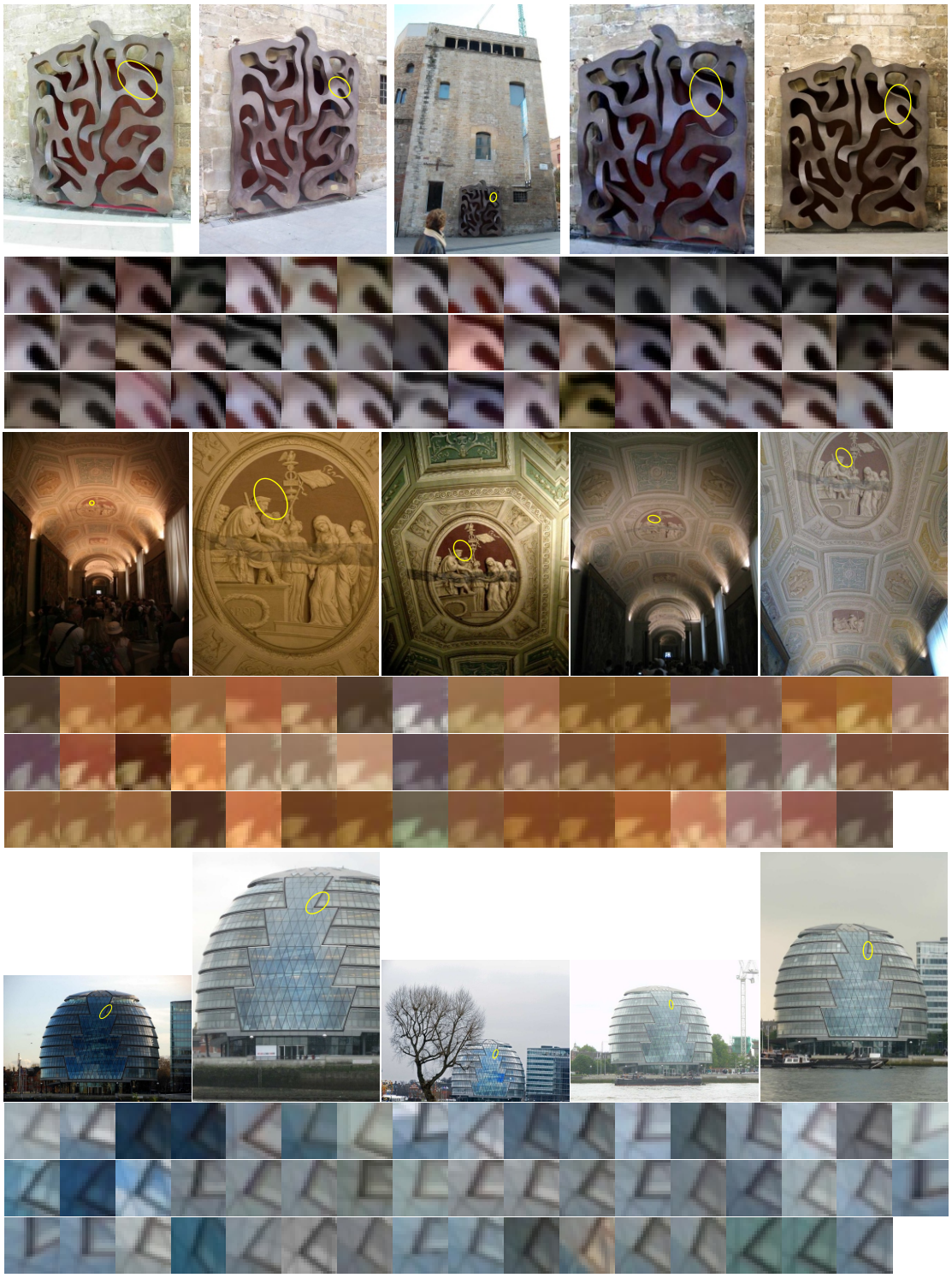


Figure 3.8: Three examples of feature tracks of size 50. Five selected images (top row) and all 50 patches of the track. Even though the patches are similar, the SIFT distance of some pairs is over 500.

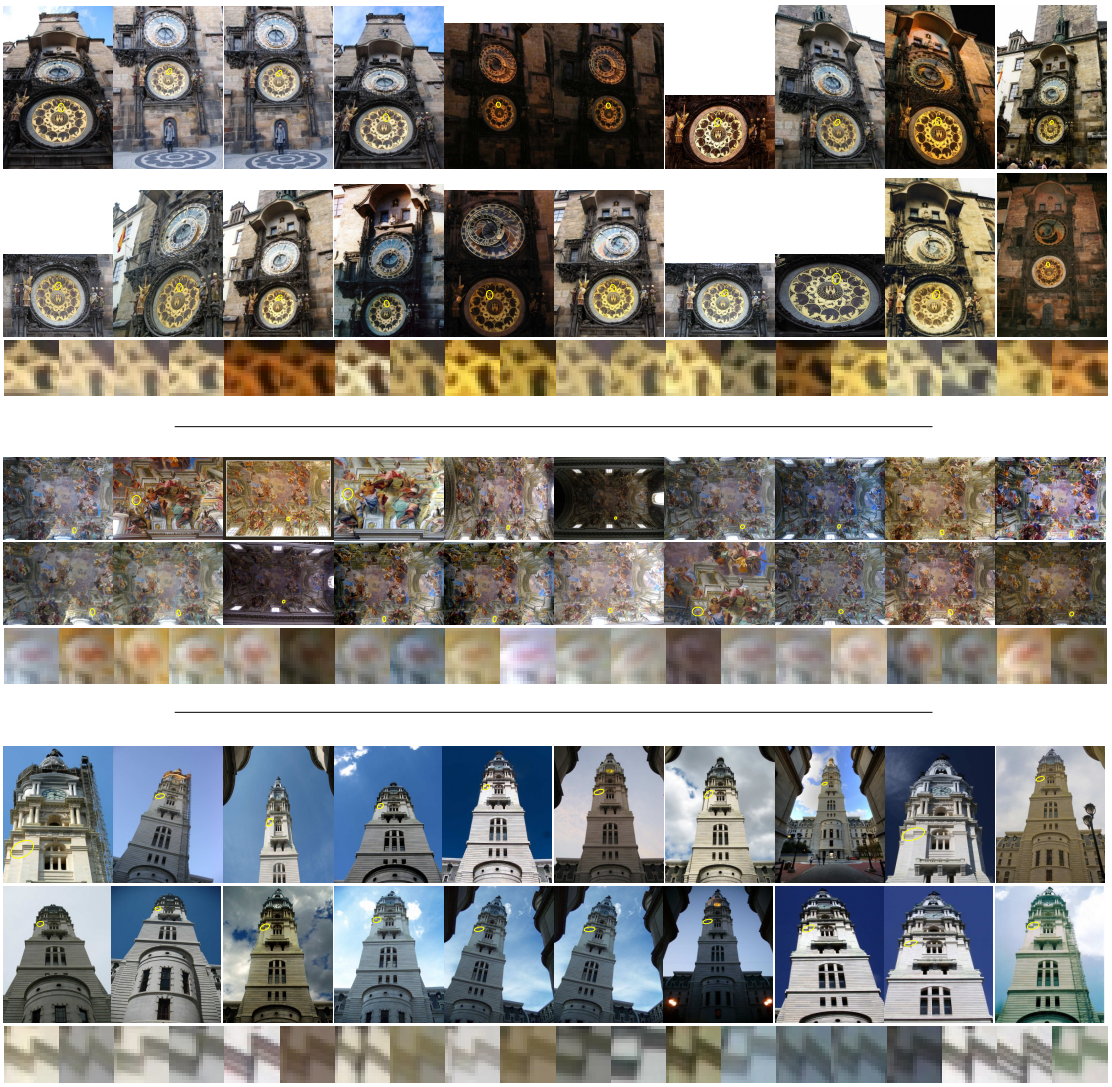


Figure 3.9: Three examples of feature tracks of size 20. Images (first two rows) and corresponding patches (third row). Note the variation in the appearance of the patches.

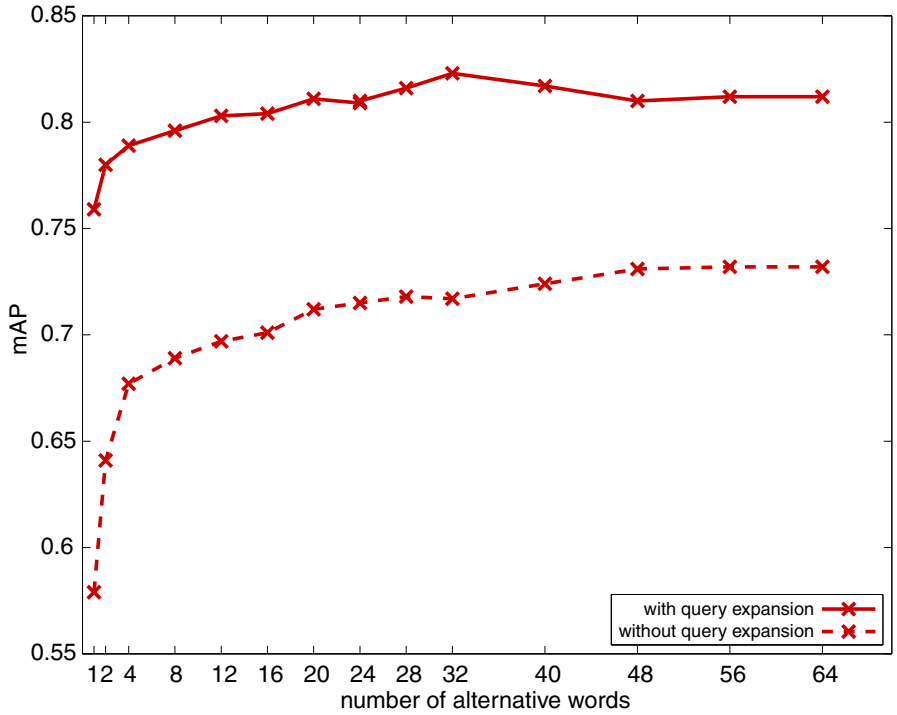


Figure 3.10: The quality of the retrieval, expressed as the mean average precision (mAP), increases with the number of alternative words. The mAP after (upper curve) and before (lower curve) query expansion is shown.

	16M L1	16M L5	16M L16	PARIS 1M L1	PARIS 1M SA 3NN
plain	0.554	0.650	**0.674**	0.574	0.652
QE	0.695	0.786	**0.795**	0.728	0.772

Table 3.2: The mean average precision for the 16M vocabulary on the Oxford 105k dataset is compared with the previous stat-of-the-art 1M vocabulary learned on Paris dataset [PCM09]. Setups with hard assignment (L1), 4 alternative words (L5), 15 alternative words (L16) and soft-assignment with 3 nearest neighbours (SA 3NN) were considered. Results without (plain) and with query expansion (QE) are shown.

(vi) Balancing by uneven splitting of the second layer discard drawbacks of growing imbalance factor for hierarchical vocabularies. We predict that this approach will be even more significant for deeper vocabularies.

3.5.2 Query Times

To compare the speed of the retrieval, an average query time over the 55 queries defined on the Oxford 105K data set was measured. Running times recorded for the same methods and parameter settings as above are shown in Table 3.3.

The plot showing dependency of the query time on the number of alternative words is depicted in Figure 3.11. The time for the reference PARIS 1M std method and the 16M L16 are of the same order. This is expected since the average length of inverted files is of the same order for both methods. The proposed method is about 20% faster,

	16M L1	16M L5	16M L16	PARIS 1M L1
Oxford 105K	0.071	0.114	0.195	0.247

Table 3.3: Average execution time per query in sec for selected vocabularies on Oxford 105k dataset. Query is executed on single machine and the time is measured excluding feature detection and description. Spatial verification was run in parallel (four concurrent threads) for 5000 images in shortlist. Query expansion step was not executed. The proposed 16M vocabulary is compared with the state-of-the-art method [PCM09].

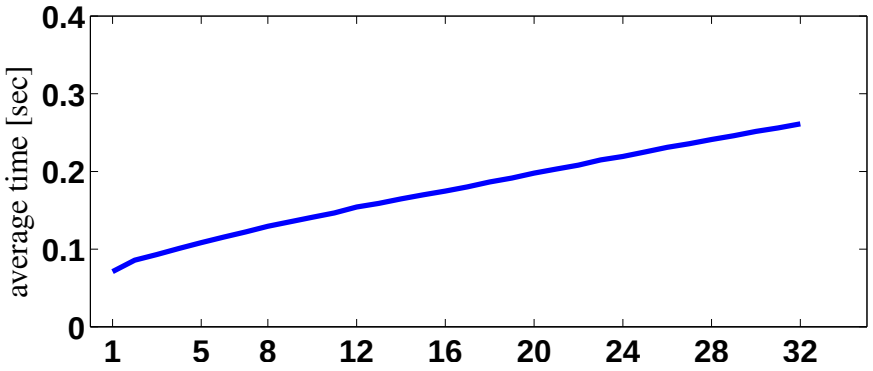


Figure 3.11: The dependence of the query time on the number of linked words for Oxford 105k dataset and 16M vocabulary. The setting is the same as in Table 3.3.

but this might be just an implementation artefact.

We looked at the dependence of the speed of the proposed method as a function of the number alternative words. The relationship shown in Fig. 3.11 is very close to linear plus a fixed overhead. The plot demonstrates that speed-accuracy trade-off is controllable via the number of alternative words.

Finally, the average query time for plain bag-of-words (no alternative words) as a function of the dictionary size was evaluated. To measure directly the speed of traversing the inverted file, the query time without the spatial verification is measured. Results are shown in Figure 3.12.

3.5.3 Results on Other Datasets

The proposed approach has been tested on a number of standard datasets. These include Oxford, INRIA holidays [1], and Paris datasets. In all cases (Table 3.4), the use of the alternative visual words improves the results. On all datasets except the INRIA holidays the method achieves the state-of-the-art results.

The proposed method is designed and trained to improve retrieval of specific object by better matching of features that are projections of *identical physical scene patch*. In the INRIA dataset, it is known that many queries rely on retrieving similar content

[1]The Holidays dataset presented in [JDS08] contains about 5%-10% of the images rotated unnaturally for a human observer. Because the rotational variant feature descriptor was used in our experiment, we report the performance on a version of the dataset, with corrected orientation of the images according to EXIF, or manually (by 90°, 180° or 270°), where the EXIF information is missing and the correct (sky-is-up) orientation is obvious.

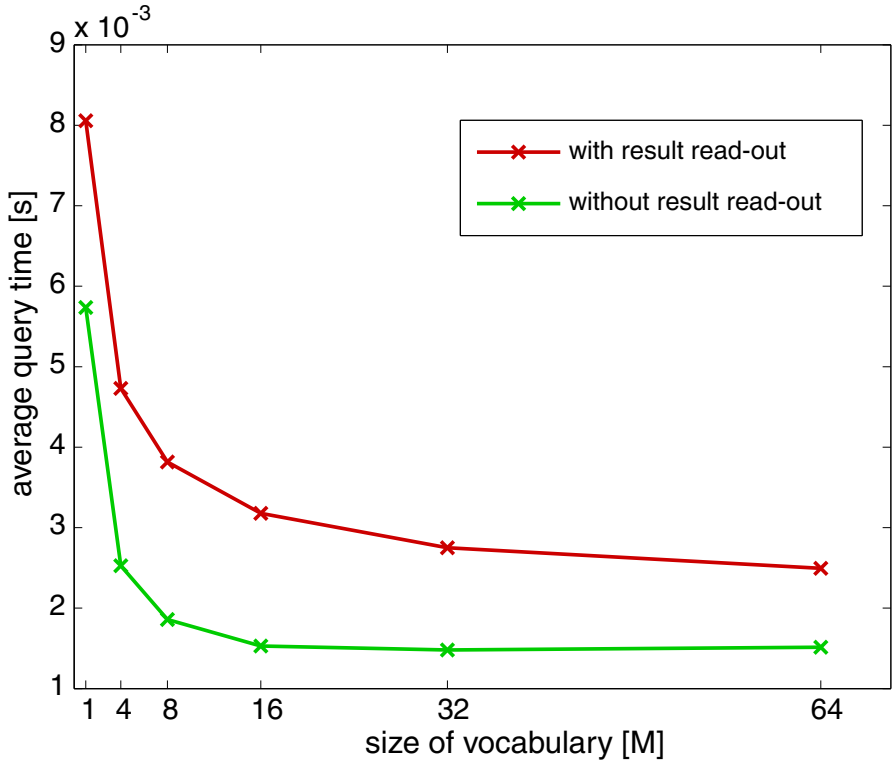


Figure 3.12: The dependence of the query time on the vocabulary size. The times were measured on the Oxford 105k dataset. To measure the speed of inverted file, we are not using spatial verification, alternative words, or query expansion. The green line shows times measured without sorting the documents according the score and copying them to the output.

Dataset	16M L1	16M L16	16M QE	16M L16 QE
Oxford 5k	0.618	0.742	0.740	0.849
Oxford 105K	0.554	0.674	0.695	0.795
Paris 6k	0.625	0.749	0.736	0.824
Paris 106k	0.533	0.675	0.659	0.773
INRIA Holidays rot	0.742	0.749	0.755	0.758

Table 3.4: Results of the proposed method on a number of publicly available datasets for a vocabulary with 16 millions visual words. Four setups are compared: (L16) with 15 alternative words, (L1) without alternative words, with and without (QE) query expansion. (The result for the Oxford 105K is duplicated for completeness.)

rather than on exact feature matching. We consider this property of the dataset to be the reason for relatively small increase in the performance by our method.

3.6 Conclusions

We presented a novel similarity measure for bag-of-words type large scale image retrieval. The similarity function is learned in an unsupervised manner using geometrically verified correspondences obtained with an efficient clustering method on a large image collection.

The similarity measure requires only negligible extra space in comparison with the standard bag-of-words method. Experimentally we show that the novel similarity function achieves mean average precision that is superior to any result published in the literature on the standard Oxford and Paris datasets/protocols. At the same time, retrieval with the proposed similarity function is faster than the reference method.

We showed that using 2 layer hierarchical approach enables to build a large vocabulary, which performs better and faster and proposes the simple balancing method, which helps to keep imbalance factor low.

As a secondary contribution we make available the database of matching SIFT features, together with the source code of the feature detector (Hessian affine) and descriptor used to extract and describe the features [ijc12].

Chapter 4

Query Expansion with Context Learning

In this chapter, we focus on the query expansion (QE) step. Automatic query expansion [CPS+07] has been shown to bring a significant boost in performance [CPS+07, PCI+08, JDS09, PCM09], and all state-of-the-art retrieval results have been achieved by methods that include a QE step. Published QE methods focus on enriching the query model by adding spatially verified features. Retrieval with the "expanded" query follows. It has been observed that if the shortlist has enough true positives, the spatial verification re-ranking almost always correctly identifies relevant images, and, consequently, results for the expanded query are significantly better than the original single image query.

As a first contribution, we improve spatial verification and re-ranking by taking account of already evaluated results. The *incremental spatial re-ranking* (**iSP**) allows verification and subsequent use of images for query expansion that do not have a significant match against the original query, but do match a statistical model gradually built from the query and previously verified images.

As a second contribution, we propose a method that exploits spatial context by incorporating matching features outside the initial query boundary into the query expansion. Since the content outside the query region is not known at query time, the method requires efficient spatial verification of the retrieved images (Fig. 4.1).

In the next section, the novel incremental spatial verification is proposed. The query expansion with context growing is described in Section 4.3. Finally, the performance is evaluated in the last section of this chapter.

4.1 Improving Blind Relevance Feedback in QE

In QE, spatial verification and re-ranking plays the role of blind relevance feedback. Spatially consistent images retrieved with the original query are deemed "relevant", similarly to the images chosen by the user in manual relevance feedback. The selected parts of "relevant" images then contribute to the new, expanded query. The quality of the decision on relevance significantly influences the success of query expansion.

In this section, two improvements of spatial re-ranking are presented. First, we introduce *incremental spatial re-ranking* (**iSP**), where the verification accounts for not

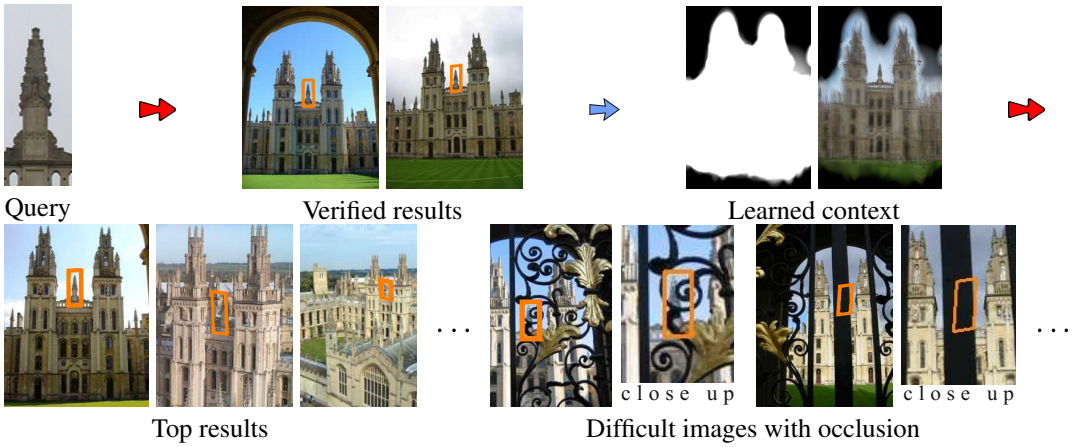


Figure 4.1: **Context expansion**: Consistent context is learned from retrieved images. The context enables successful retrieval (before the first false positive image) and localization of heavily occluded objects.

just spatial agreement with the initial query, but also agreement with all previously verified images. Second, we show that it is beneficial to "grow" the model of the object beyond the boundaries of the initial query, and to examine the spatially consistent neighbourhood of the query.

4.2 Incremental Spatial Re-ranking

In this section, an improvement of the spatial re-ranking (**SP**) phase of the baseline method (see Section 2.3) is proposed. As in the baseline method, the novel incremental spatial re-ranking (**iSP**) starts with the shortlist $\mathcal{S}$ of images ordered by the BoW score. The objective of **iSP** is to form a statistical model of the query object.

Initially, the statistical model M^0 includes only features from the query. Next, images in the shortlist are considered in the order given by BoW scoring. Each image $X \in \mathcal{S}$ is geometrically matched against the current model M^i. If the image matching quality $I_{M^i}(X)$ is greater than θ, the query object model is updated, and M^{i+1} is formed.

The quality function $I_{M^i}(.)$ is defined as the number of geometrically consistent features with the same visual word in image X and model M^i. The threshold θ was set to 15 after extensive preliminary experiments. The updated model M^{i+1} is the union of features in model M^i and features in image X, back-projected using function $f(.)$ onto the query image, clipped by the query bounding box. The final ranking of a shortlisted image is defined by the quality function. The method is described in Algorithm 4.

Since the simplest quality measure described above performed well, no alternatives, *e.g.* accounting for inlier ratio, geometric overlap, or weights of matching features, were evaluated.

Algorithm 4 Incremental spatial re-ranking

Input: query image X_q, shortlist $\mathcal{S}$ of images
Output: ranking $R : \mathcal{S} \leftrightarrow \{1..|\mathcal{S}|\}$, expanded model M^n of the object

$M^0 := X_q$
$Q := [], i := 0$
for $k := 1$ to $|\mathcal{S}|$ **do**
  $X := \mathcal{S}[k]$
  $Q[k] := I_{M^i}(X)$
  if $Q[k] > \theta$ **then**
    $M^{i+1} := M^i \cup f(X)$
    $i := i + 1$
  end if
end for
$R :=$ ranking of the images according to $Q[k]$.

4.3 Outside the Query Boundaries: Incorporating Context

The content outside the query region is not known at the query time. It is clear that learning the query context must be done by the "matching results to results" approach. The process of the *context learning* takes place either after spatial re-ranking, or, in the case of **iSP**, after each update of the query object model. The latter has the advantage that an image may be verified with the help of the context. In this case, implementation of context growing is trivial. As in **iSP**, features are back-projected to query image and are added to the model regardless of whether they are inside or outside the query bounding-box. The extension of the object model beyond the boundary of the original query only requires relaxing this constraint.

At the beginning of the learning phase, the context is identified with the area inside the query boundary. A feature added into the model that is not inside the context is inactive until confirmed by feature(s) from another image with the same visual word and similar geometry. Once a feature is confirmed, it adds the neighbourhood around its center to the context. All the confirmed features in the context are treated as active. The active features are considered the same as those inside the bounding box, and are used in spatial verifications, and, finally, in the query expansion. This is efficiently implemented by spatial binning. The process is summarized in Fig. 4.1.

The progress of context growth for two queries is visualized in Fig. 4.2. The learned model of the query is shown as the mean of elliptic patches associated with its features back-projected to the query. The query bounding box is drawn as an orange rectangle. To save space, the area not covered by the model, or equivalently, the area not covered by a single feature, is cropped. Experiment 2, summarized in Tab. 4.2, shows that including the context improves performance.

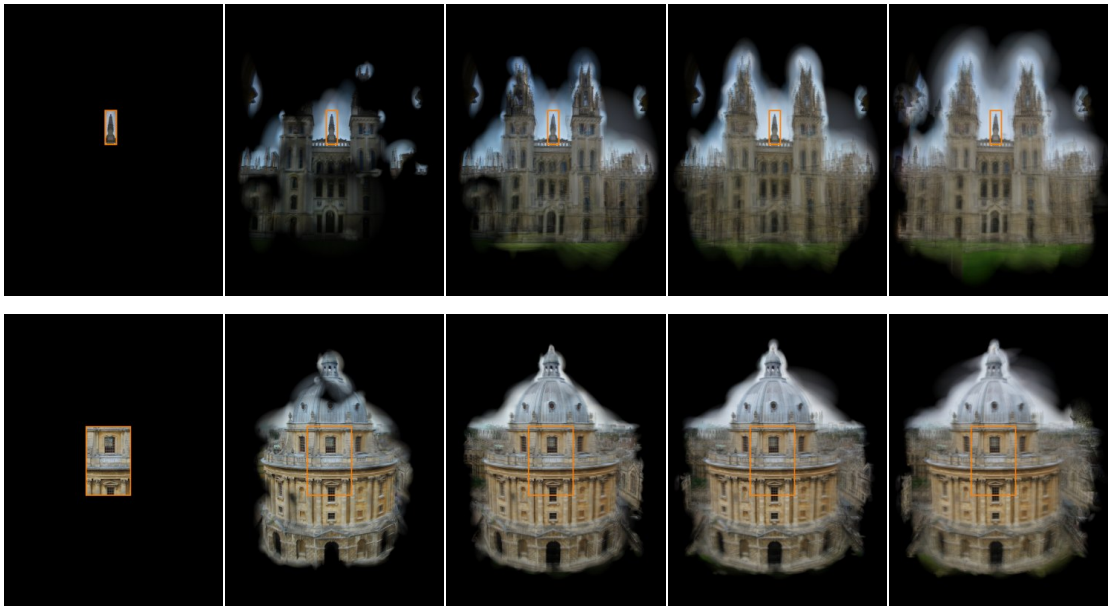


Figure 4.2: The process of context learning. Left column: the original query. Other columns: feature patches back-projected into the context from 2, 5, 10 and 20 spatially verified images.

4.4 Experiments

Datasets and Evaluated Methods

The image retrieval methods proposed in Sections 4.2 and 4.3 were evaluated according to the standard protocol on the Oxford and Paris datasets described in Section 1.4. The performance of all retrieval experiments is measured using the mean average precision (Section 1.3).

The proposed incremental spatial re-ranking and context growing methods are compared with the state-of-the-art image retrieval approaches, see the list in Tab. 4.1. All methods use, in experiments on the Oxford dataset, a 1M visual word vocabulary trained on the Paris dataset and vice versa.

Experiment 1. Evaluation of Incremental Spatial Re-ranking

The experiment compares all image retrieval methods listed in Tab. 4.1 on the Oxford and Paris datasets. We observe that **iSP** outperforms **SP** in all cases; compare the left and right columns of sections I, II and III of Tab. 4.2. The **iSP** improves performance by approximately one half of the query expansion effect; compare columns I right, and II left. Since only the shortlist is accessed, the performance improvement is obtained at a negligible cost compared to issuing a second query. This encourages the use of **iSP** instead of the standard **SP** re-ranking. Additionally, the benefits of **iSP** and query expansion are additive; compare columns I right and II right. Finally, adding context has negligible effect on the Oxford dataset and improves performance on the Paris dataset. This is due to the fact that on the Oxford protocol, queries include entire objects, and there is little gained by growing the context.

SP	BoW scoring, spatial re-ranking, no query expansion, see Sections 2.3 and 2.4
iSP	BoW scoring, incremental spatial re-ranking, no query expansion
SP + avg QE	BoW scoring, spatial re-ranking, average query expansion, see Sections 2.3 and 2.4
iSP + avg QE	BoW scoring, incremental spatial re-ranking, average query expansion
SP + ctx QE	BoW scoring, spatial re-ranking, context query expansion, see Section 4.3
iSP + ctx QE	BoW scoring, spatial re-ranking, incremental spatial re-ranking with context and context query expansion.

Table 4.1: Description of the state-of-the-art (rows 1 and 3) and the proposed methods (rows 2,4,5 and 6).

	I. w/o QE		**II. avg QE**		**III. ctx QE**	
	SP	**iSP**	**SP**	**iSP**	**SP**	**iSP**
Oxford 5k	0.616	0.741	0.785	0.825	0.781	0.827
Oxford 105k	0.553	0.649	0.725	0.761	0.731	0.767
Paris 6k	0.617	0.679	0.720	0.772	0.753	0.805
Paris 106k	0.508	0.556	0.627	0.687	0.653	0.710

Table 4.2: Comparison of image retrieval methods with standard (**SP**) and incremental spatial re-ranking (**iSP**).

Experiment 2. Evaluation of Context Expansion

Next we study the influence of incorporating the context of the query *i.e.* extending the model of the query outside its bounding box. The behaviour is demonstrated on the same datasets by using a novel protocol.

As shown in experiment 1, the effect of context learning is not significant in the case of the Oxford dataset. To model a situation where only a detailed or partial view of the object is available, the following protocol was devised: The query bounding boxes were symmetrically reduced to 10% of their area in nine steps, see Fig. 4.3. The maximum spatial extent of the context was limited to an area 25× larger than the reduced query bounding box.

The results (see Fig. 4.4) show that the performance of the retrieval method using both context and incremental spatial re-ranking (**iSP + ctx QE**) drops below the state-of-the-art (black dashed line in Fig. 4.4) method only after reducing the bounding box area to 40%, (Fig. 4.4b,d), or even to 20% (Fig. 4.4a,c) of the full query bounding box. One of the reasons for the drop in performance is that to keep the number of features in the model, and thus the speed of spatial re-ranking reasonable, we limit the number of images added to the model to ten, which is insufficient to reconstruct the model to the quality of the original query. Also, the results of initial queries on the standard datasets already contain many of true positives, and even the standard query expansion manages to retain a sufficient model of the object.

Some examples of contexts learned for some of the Oxford protocol queries are shown in Figure 4.3.

4.5 Conclusions

The spatial verification and re-ranking step was improved by incrementally building a statistical model of the query object. We show that using this incremental spatial verification the spatial context of the query object can be learned and used in query expansion to improve retrieval performance.

The proposed improvements of query expansion were evaluated on established Paris and Oxford datasets. Experiments show that very similar results are achieved with only part of the original query. According to the standard protocol state-of-the-art results were achieved.

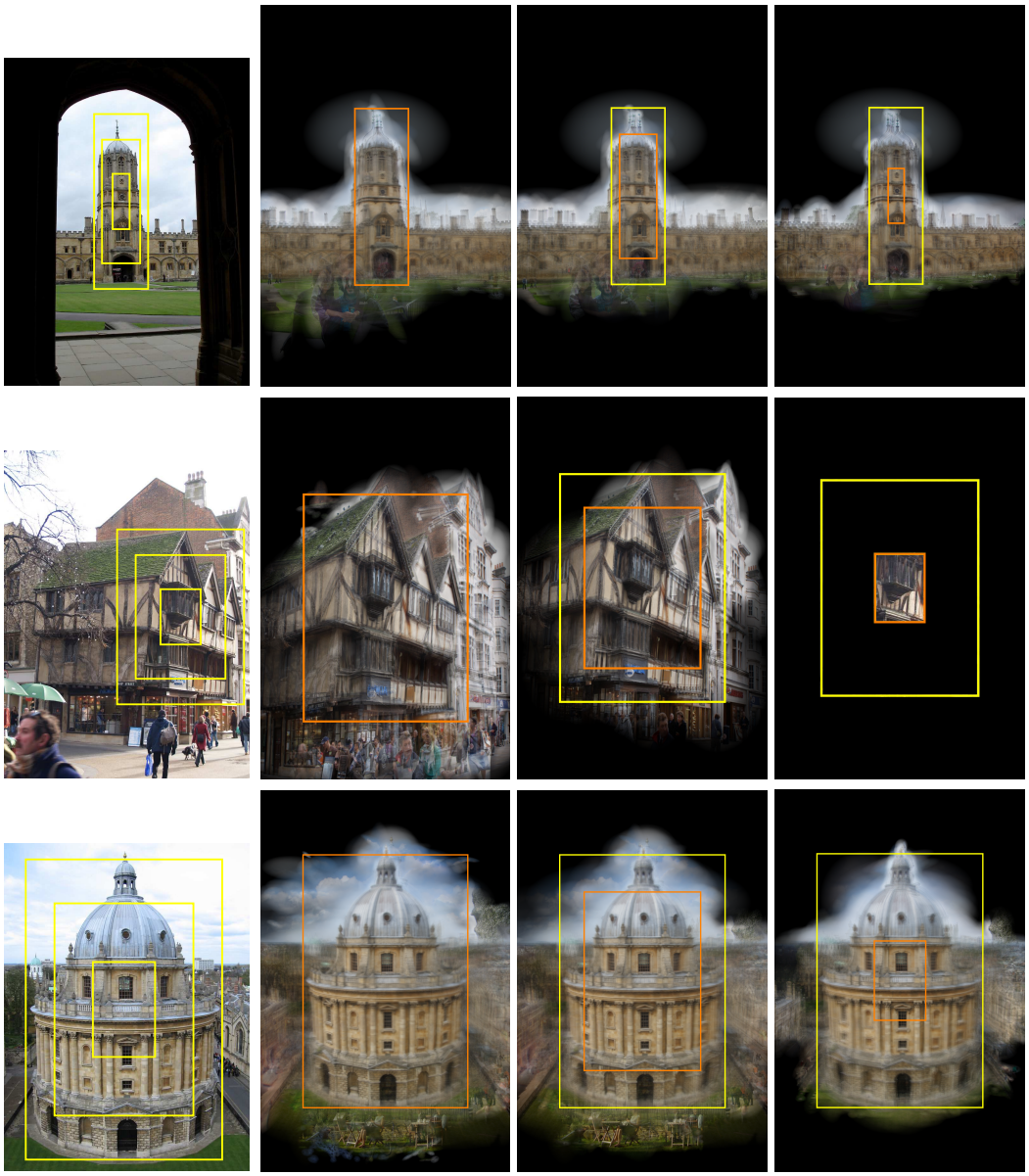
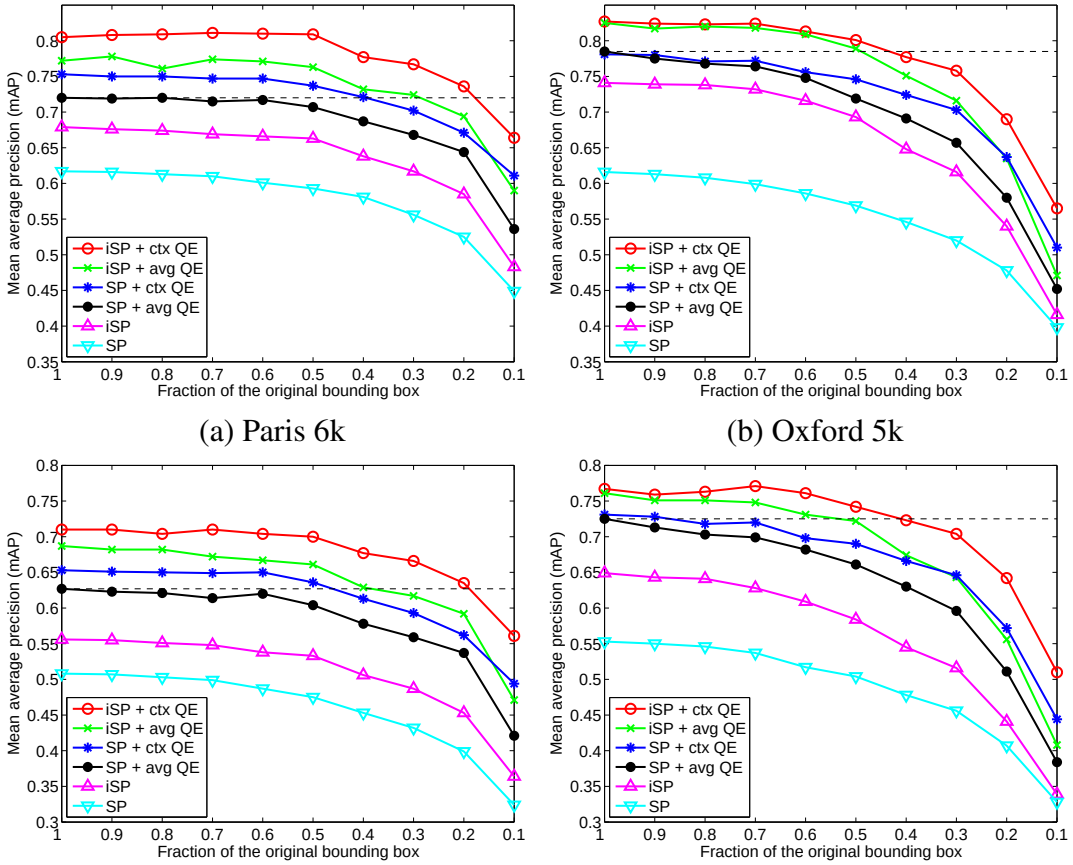


Figure 4.3: Left column: Examples of the full (100%) bounding box of some Oxford protocol queries (outer rectangle) and the query bounding boxes reduced to 50% and 10%. Columns 2, 3 and 4 depict the context learned from the full, 50% and 10% bounding boxes respectively (the orange rectangles). The yellow rectangle shows the original bounding box. Note the ability of the **iSP + ctx QE** to learn the context even from the smallest query. The method failed on the CORNMARKET 10% (right column, middle) due to the insufficient number of spatially verified images.



(a) Paris 6k (b) Oxford 5k

(c) Paris 106k (d) Oxford 105k

Figure 4.4: The influence of decreasing the query bounding box size on image retrieval methods. The black dashed line is the performance of the state-of-the-art [CPS+07] method with the original bounding box. The performance of the proposed **iSP + ctx QE** is superior to the state-of-the-art method, if the query covers more than 20% of the bounding box on the Paris datasets, and more than 40% of the bounding box on the Oxford datasets. The compared methods are listed in Tab. 4.1.

Chapter 5

Automatic Failure Recovery in Query Expansion

One of the key issue in an image retrieval system based on bag-of-words is the definition of image similarity. The most common approach is to define visual similarity as a normalized sum, over all visual words, using the *tf-idf* weighting scheme [SZ03]. The weight of the word increases proportionally to the number of times it appears in a document, but is offset by the frequency of the word in the corpus. This helps to handle the fact that not every visual word is equally important – has the same discriminability. The *tf-idf* is commonly used no matter which type of vocabulary is used (Section 2.2).

In the image retrieval systems, the use of a similarity measure is typically justified by its probabilistic interpretation. Any similarity measure employing summing over all visual words implicitly assumes that visual words occur independently on each other. This assumption is made because of computational convenience and it is intuitively obvious that it does not hold. Groups of correlated features typically occur on the water surface, on vegetation, images of text, faces, net-like structures, repetitive patterns, and statistical textures [CN08]. If such group is visible on the image query, and is not related to the object of interest, the BoW retrieval, without any special treatment, fails to select

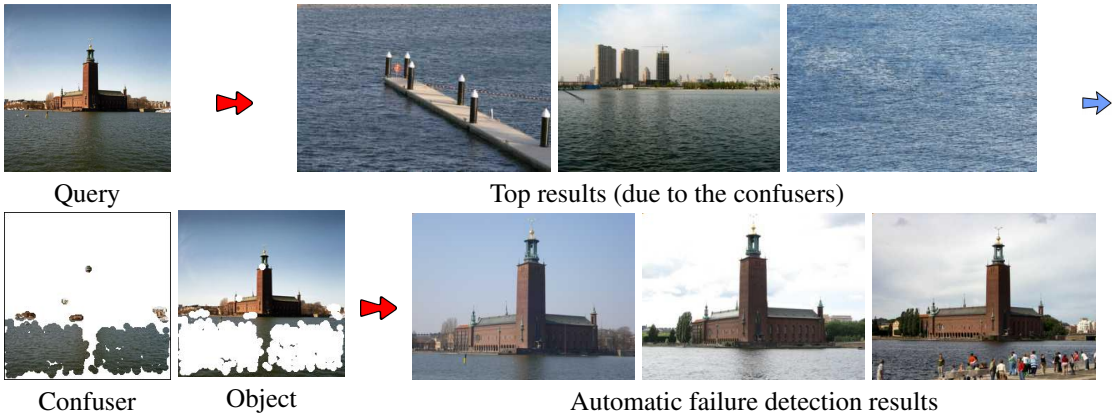


Figure 5.1: **Automatic Failure Recovery**: Initial retrieval results corrupted by confusing water features. The confuser model is learned dynamically. Successful subsequent query using the confuser model.

Figure 5.2: Examples of queries (leftmost images) where standard image retrieval fails to return images of the query object (upper rows of results). The results of the method [CM10b] with removed cooc-sets, which were discovered during the off-line stage. Courtesy of Ondrej Chum [CM10b].

relevant images into the shortlist. This is a consequence of correlated voting for images that contain the same type of 'confusers', which suppresses the relative contribution of the specific object. (see Fig. 5.2)

Moreover if BoW fails to the extent that there are no, or very few, correctly retrieved images in the shortlist, standard QE is no help. Such situations, which arise in the presence of structures with multiple correlated features, have been referenced in the literature as cooc-sets [CM10b] or confusers [KSP10].

In this chapter we show how to detect and recover from the *failing query expansion* situation. Unlike other approaches, the proposed method handles the presence of confusers in the query region on-the-fly, with no prior learning step required. We achieve performance that is comparable to the state-of-the-art without the need for off-line and potentially time-consuming processing that is difficult to execute in a continuously updated database.

5.1 Query Model

We model the query (visual) words as a mixture of words generated by three processes (topics): the object words $\mathcal{O}$, the confuser words $\mathcal{C}$, and the random words $\mathcal{R}$. The three types of words, and their properties, are described in the following paragraph.

We address the retrieval of *specific* objects, defined as a collection of features that preserves appearance and spatial layout over a range of imagining conditions such as viewpoint change, and scale change.

The object words $w \in W_{\mathcal{O}}$ are likely to be observed in images containing the object of interest, *i.e.* $P(w|\mathcal{O})$ is high, $P(w|\mathcal{O}) \gg P(w)$. Moreover, the features associated with words, $w \in W_{\mathcal{O}}$, appear at fixed coordinates with respect to the canonical frame of the object, and thus allow for the geometric consistency check. The confuser words $w \in W_{\mathcal{C}}$ are defined as sets of correlated words, satisfying $P(w|\mathcal{C}) \gg P(w)$. However, confuser words are not significantly spatially consistent[1]. Randomly occurring words, $w \in W_{\mathcal{R}}$, generated from spurious features, and corrupted descriptors form the most frequently occurring class. As reported in [TL09], object features cover as few as 4% of the total features.

Algorithm 5 Automatic failure recovery

Input: query features Q_0
Output: $\langle$ query features, query results, feature mask $\rangle$

Execute query Q_0 including spatial verification
if $\rho(Q_0) > \rho_0$ **then**
  return $\langle$ Q_0, results (Q_0), empty $\rangle$
end if
Learn a set of confuser words $W_{\mathcal{C}}$ (eqn. 5.1)
$Q_N = Q_0 \setminus W_{\mathcal{C}}$
Execute query Q_N including spatial verification
if $\rho(Q_0) > \rho(Q_N)$ **then**
  return $\langle$ return Q_0, results (Q_0), empty $\rangle$
else
  return $\langle$ return Q_N, results (Q_N), mask($W_{\mathcal{C}}$) $\rangle$
end if

5.2 Recovery

We propose a modification, called automatic failure recovery, to the retrieval scheme. First, standard BoW retrieval with spatial verification is performed. The BoW scoring is used to produce a shortlist of documents. The images in the shortlist are checked for spatial consistency with the query features. The shortlist is significantly shorter than the size of the database[2]. If relevant images are included in the shortlist, these are identified by spatial re-ranking. Once relevant documents are retrieved, automatic query expansion techniques are used to improve the object model $\mathcal{O}$. When a significant number of confuser words $\mathcal{C}$ is present in the query, the whole shortlist can be populated by images containing features generated from $\mathcal{C}$, and hence the spatial re-ranking cannot improve the search results. We call this situation a *retrieval failure*. Even though the

[1] We do not aim to solve a philosophical question regarding whether recurring objects, such as phone booths, are objects or confusers. According to our model, appearance and spatially consistent features form objects.

[2] In our experiments, a shortlist of 1000 documents is used.

shortlist does not contain relevant images, it still conveys valuable information. A statistical model of the confusers $\mathcal{C}$ present in the query can be learned from the images in the shortlist, since a vast majority in the shortlist score higher than the relevant images. Once the confuser model $\mathcal{C}$ is known, its influence on the query is suppressed. There are three issues that need to be addressed: (i) efficiently estimate the confuser model $\mathcal{C}$, (ii) down-weight the effect of the confusers to the query result, and (iii) decide if the retrieval failure has arisen. The algorithm is summarized in algorithm 5

Ad (i) The distribution $P(w|\mathcal{S})$ of visual words in the shortlist $\mathcal{S}$ is learned at virtually no cost during the tentative correspondence construction in the spatial re-ranking phase. Features whose visual words appear significantly more frequently than in the database are deemed to be part to the confuser model $\mathcal{C}$:

$$W_{\mathcal{C}} = \{w | P(w|\mathcal{S})/P(w) > r_0\}\,. \tag{5.1}$$

The likelihood ratio threshold was $r_0 = 10$ in our experiments.

Ad (ii) There are many options to reduce the influence of the estimated confusers $\mathcal{C}$. We choose to simply remove the confuser features from the query. This approach, while seeming naive, has been shown to be effective and efficient [CM10b]. If a query expansion, average or any other type, is used after the failure recovery, features that back-project to regions occupied by the confusers are also removed. This prevents back-projected confuser features from entering the expanded query from the result images.

Ad (iii) To check whether a retrieval failure has arisen, we compare the estimated quality $\rho(Q)$ of results of two queries: the original query, Q_0, and the query after the recovery, Q_R. We estimate the quality of the results by the inlier ratios in the top matching results. First, the acceptable result images that each have an absolute and relative non-random number of inliers are selected. The score of the retrieval is then defined as the sum of inlier ratios over the acceptable results. Formally, let $\mathcal{S}_Q$ be a BoW shortlist of query Q, $T_Q(X)$ be a number of tentative correspondences between query Q and image X, and let $I_Q(X)$ be the number of geometrically consistent features between Q and X. The acceptable result of Q is a set of images

$$\mathcal{A}_Q = \left\{ X | X \in \mathcal{S}_Q \;\&\; I_Q(X) > I_0 \;\&\; \frac{I_Q(X)}{T_Q(X)} > \epsilon_0 \right\}.$$

The quality of the shortlist result of query Q is defined as

$$\rho(Q) = \sum_{X \in \mathcal{A}_Q} \frac{I_Q(X)}{T_Q(X)}. \tag{5.2}$$

To avoid wasted computation when improvement is unlikely, the estimated quality of the original query Q_0 is thresholded. If $\rho(Q_0) > \rho_0$, then the hypothesis of the query failure is directly rejected. In the experiments, the following parameters were used: minimal acceptable number of inliers $I_0 = 5$, minimal acceptable inlier ratio $\epsilon_0 = 0.2$, and the failure rejection threshold $\rho_0 = 5$.

	AFR		Cooc [CM10b]		Baseline	
	AP	FP	AP	FP	AP	FP
Stockholm	**0.659**	**16**	0.569	15	0.032	1
Dragon Wall	**0.797**	**56**	0.726	52	0.065	5
St Ignazio	**0.945**	**17**	0.737	14	0.105	2
Colloseum	**0.762**	**514**	0.136	85	0.018	13
Barcelona	**0.895**	**17**	0.789	15	0.053	1
St Mary	**0.943**	**57**	0.895	51	0.020	1
Vatican	**0.957**	**22**	0.870	20	0.130	3
Bridge	0.583	4	**0.716**	**5**	0.143	1

Table 5.1: Quantitative comparison of the proposed method with [CM10b] and the baseline method on the Q8 dataset: Estimated average precision AP and the rank of the first false positive FP. Queries and their confuser models are displayed in Figure 5.3

5.3 Efficiency

The proposed method introduces no extra cost for queries that return reasonable number of matching results (this is the case for almost all images in the standard Oxford and Paris datasets, where the query bounding box is tightly around the query building). For such queries the result is also unaffected because the original query is accepted. For other queries, one extra BoW scoring and spatial re-ranking step is executed. Since the new query is a subset of the original query, this additional step is faster than the original query.

5.4 Experimental Results

In this section, we compare the results of the confuser model learned in the proposed automatic failure recovery step with results obtained by cooc-sets [CM10b]. The quantitative results on the Q8 dataset [CM10b] embedded in a database of over 5 million images are shown in Tab. 5.1. It is not feasible to obtain all true positives, so the average precision is only an upper bound estimate. New positive results have been discovered by the proposed method, and the AP values are not directly comparable to values in [CM10b]. Queries of the Q8 dataset, their confuser models and subsequent queries using the confuser models are displayed on Figure 5.3.

Table 5.1 shows, that for most cases, the two methods give comparable results. For the 'Colloseum' query, the proposed approach gives significantly better results than results obtained by the cooc-sets approach. This is because the cooc-sets approach excludes object-relevant cooc-set features as well as the confuser features, as opposed to the proposed approach which correctly learns the confuser model.

5.5 Conclusion

A method capable of preventing *query expansion failure* caused by the presence of confusers was introduced. In the comparison to the existing off-line method [CM10b]

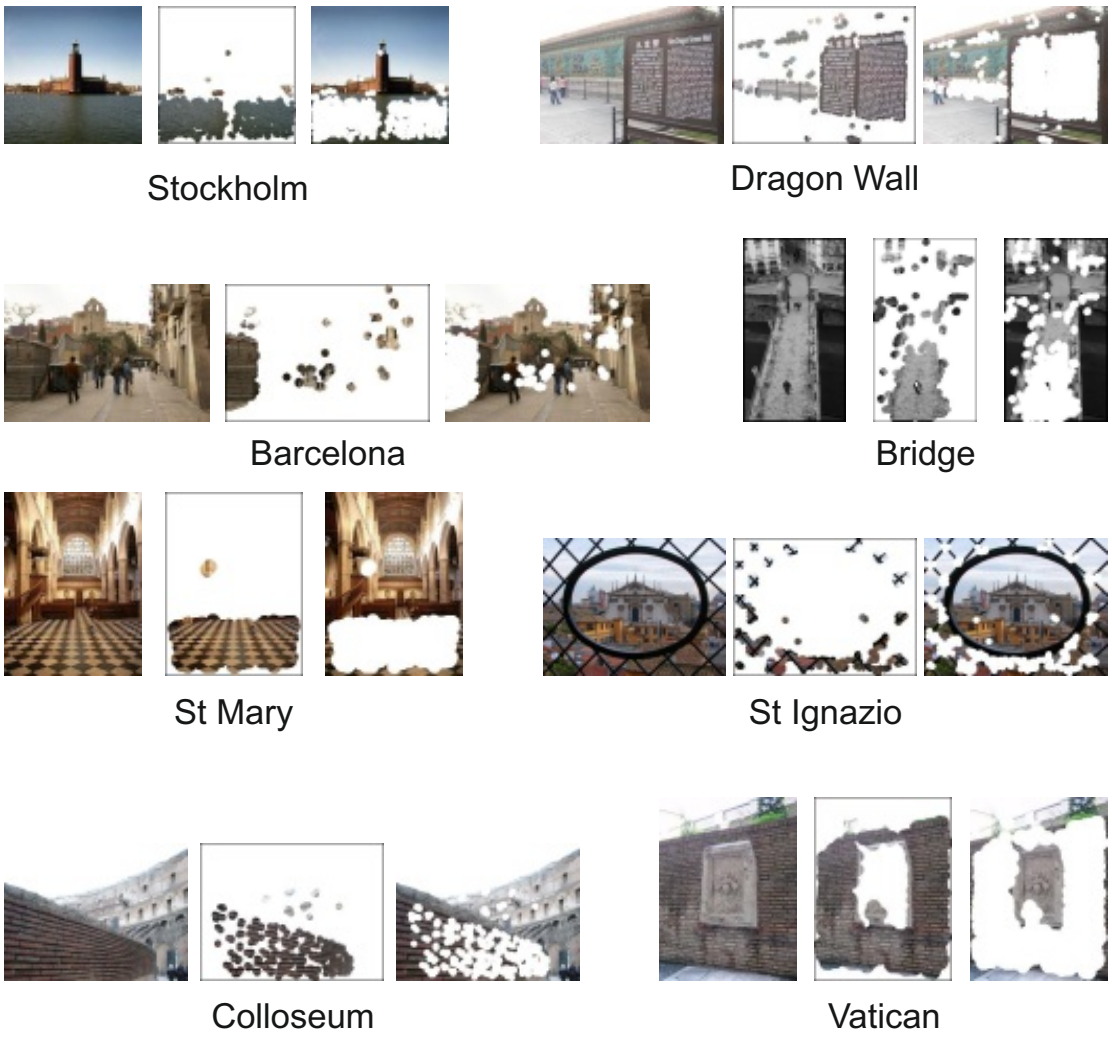


Figure 5.3: Queries of the Q8 database. Each query is displayed with the dynamically learned confuser model and a subsequent query of an object. Quantitative results are displayed in Table 5.1.

it has these pros and cons:

Pros: No pre-learning step is required, so the method is applicable to any dataset and any vocabulary, and additionally, it does not require good training sets that generalize well, or retraining for different vocabularies. The method is specific to the current database and to the current query, so features for some queries that are confusers can be useful for other queries.

Cons: The proposed method requires the execution of the original query, while the cooc-set approach can filter confuser features beforehand. In some queries, the confuser features may represent significant proportion of the features and thus the full query takes longer to execute.

Chapter 6

Novel ranking functions – zooming

Novel problem formulations for large scale image retrieval are proposed in this chapter, 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 query expansion for recall of zoomed images. The proposed methods are tested on a standard Oxford-105k dataset augmented with images of Sagrada Familia.

Later in this chapter, we further generalize the approach. Instead of starting with a user selected region of interest in the query image, the system is searching simultaneously for all interesting parts within the spatial extent of the query.

6.1 Motivation

Most object-retrieval methods take into account the requirements for efficient content-based navigation and browsing of large-scale image collections.

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 Fig. 6.1 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. 6.1 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.

A possible interface to such functionality is as follows. The user issues a query 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. In this case the desired result is "zoom-in" – the answer to "What is this?" question.

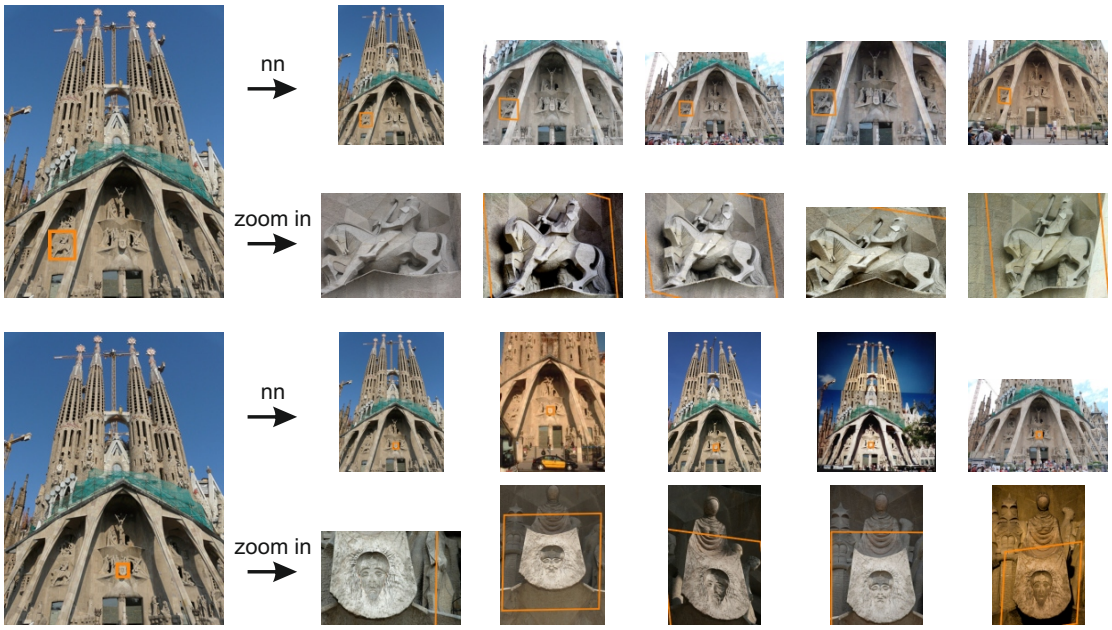
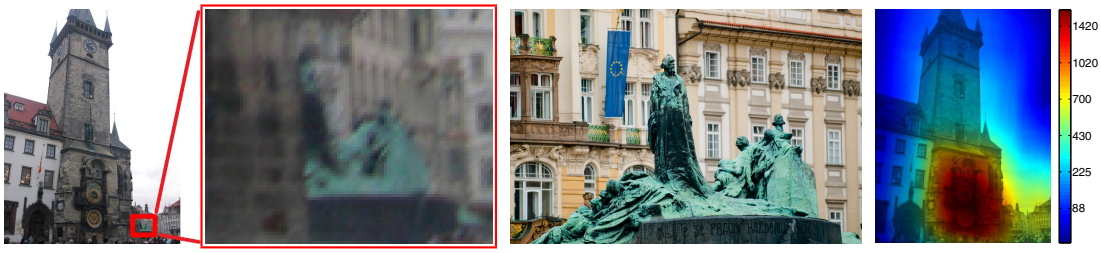


Figure 6.1: Comparison of outputs of the standard and zoom-in approaches. Two queries differing only by bounding-box were issued on the image in the leftmost column. The standard "most similar image" approach (nn, top rows) retrieves nearest neighbor matches, which provide no detailed images local to 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).

Figure 6.2: Comparison of outputs of the standard and the proposed approach. The standard "most similar image" approach (nn, top rows) retrieves nearest neighbor 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).

On the other hand, the user might be interested in a broader contextual query – zoom-out, to answer the *"Where is this?"* question (see Figure 6.2).

Building on top of these methods with additional result grouping and spatial verification prior to query expansion, we can solve another task: *Find all "interesting" parts within the spatial extent of the given query*. Two definitions of "interesting" lead to different tasks. The first is to find, for all pixels in the query, the highest resolution images depicting it, Figure 6.3 (left) and Figure 6.4 we call this function *"Highest resolution transform"*. The second is to find regions of interest that are the most often photographed, Figure 6.3 (right). For more examples and comparison of the two task, see also Figures 6.8 to 6.10.



query image 26× zoom example detail frequency

Figure 6.3: An example of a detail (middle) corresponding to the red bounding box in the query image (right). The red bounding box and the close-up are displayed for visualization purposes only and not supplied by the user. The number of images with zoom at least 3× for each pixel of the query image, denoted detail frequency, is shown on the right. The result, calculated from 2630 retrieved images, was obtained by a single application of hierarchical query expansion in less than 25 seconds. The processes required 20150 spatial verifications.

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?". These tasks, despite of having no analogy in the text retrieval can be often more useful to the user than standard search for most similar images.

Figure 6.4: The *"Highest resolution transform"* (right, central part) color-codes the available zoom-in factor at each pixel. For the query (left), the maximum scaling factor is about 30; scaling factors are expressed as ratio of lengths. The retrieved detailed images are shown around the border together with links to the pixels they correspond to. Hovering with the mouse above the query image in the retrieval system interface the images with the highest resolutions could be shown. A dataset-supported superresolution-like functionality could be easily implemented.

6.2 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 difference is in the choice of ranking function. Instead of ordering images according to similarity, it is designed to address new goals: maximizing the detail or maximizing the context.

To encourage 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 re-orders only verified images, *i.e.*, the images for which a geometric transformation was found, preferring zoomed-in or zoomed-out images respectively.

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 [SGP12] 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 [CMPM11] 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 6.

Algorithm 6 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

6.2.1 Ranking functions

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 (represented by a linear function – in our case affinity) 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 expect 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. The largest z_{in} comes first.

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 in the retrieved image 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. Smallest z_{out} comes first.

6.3 Efficient Image Detail Mining

This section describes the method for finding the finest details for every locations in the image and to find regions that are commonly photographed by the crowds. Two things prevent a simple solution of applying the method described in the previous section to every location in the image: computational efficiency and the risk of high false positive rate.

In order to solve those tasks efficiently in a large, unordered image collection, a number of issues has to be tackled. Namely, an efficient retrieval of matching sub-images with significantly different resolution has to be addressed, together with an effective rejection of false matches to prevent topic drifts. Towards this end, we introduce a novel concept of detail mining called hierarchical query expansion.

The results of the method are illustrated in Figures 6.3 and 6.4, which show the query image, a sample of the images of details discovered in the dataset and two visualizations of localized interesting parts of the query image. The color in Figure 6.4a codes the maximal resolution found in the dataset. In Figure 6.3 (right) the color codes the number of images found and back-projected into the query image.

The outputs show what the most interesting details are for the crowds photographing the landmark and which details are worth seeing (taking a picture of). It helps the user to concentrate on interesting details or suggests additional queries. Annotations (such as Flickr tags) of the discovered images can be used for describing parts of the image as in [CM10a]. The output of the proposed detail mining can be also used as a initial step for finding iconic view of the details [WL13].

6.3.1 Hierarchical query expansion

It has been demonstrated many times that the query expansion technique [CPS+07, AZ12] significantly improves on the quality of retrieval performance, especially in the recall. We introduce a novel concept of detail mining called hierarchical query expansion. After the initial query, the image is divided into sub-regions and a new, expanded,

query is issued for each of the sub-regions. The partitioning of the image is naturally driven by the density of the photographed details – the focus of the crowds. Since people tend to take pictures of individual and well aligned objects, regions defined by a number of overlapping images are good candidates for detail mining. There are three issues that need to be addressed in order to efficiently deliver qualitatively appealing results: image coverage, low redundancy, and consistency.

Image coverage and low redundancy. Typically, on well-known landmarks, certain details are photographed significantly more often than others. Considering only the top results without considering their spatial layout, as most of the query expansion approaches do, would result in neglecting details that are available in the image collection, but are depicted in a lower number of photographs. In order to obtain details in all parts of the image, lower ranked images that are not overlapping with higher ranked images are considered.

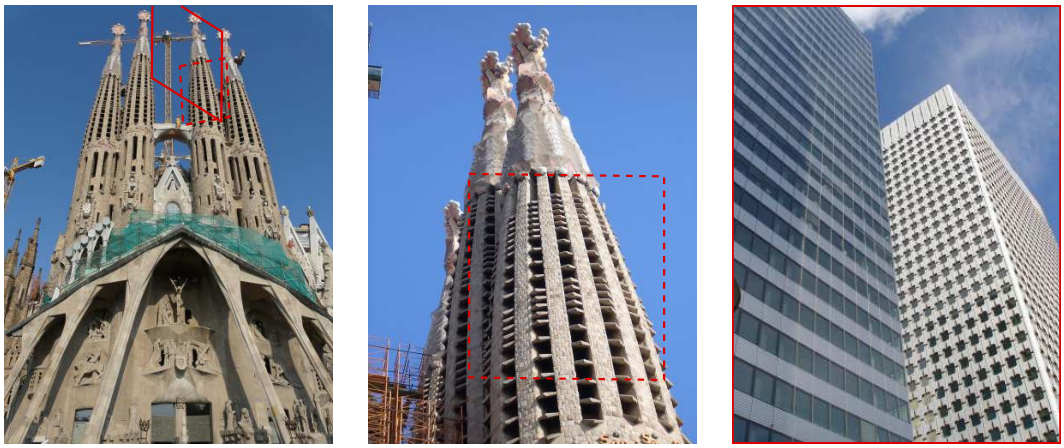
For efficiency, the retrieved images are spatially clustered and large clusters are sub-sampled. Each such cluster provides a simple generative model of a certain part of the image on a higher resolution level than the original query. The clusters are used to issue an expanded zoom-in query, to obtain further details. The procedure can be iterated. However, our experiments suggest that a single application of hierarchical query expansion is sufficient to obtain most of the details present in the database.

Consistency. Since in our approach, the user does not provide the region of interest, a number of seemingly "harmless" and uninteresting regions, such as railings in the corner of the image, can expand into enormous number of false positive images. To eliminate such topic drift, we introduce a novel mechanism to detect and eliminate inconsistencies in the retrieved results. A test is performed as an additional spatial verification between result images to ensure that no false positive will be introduced into any expanded query. In the test, an affine transformation $A_{j,i}$ mapping features from result image j to result image i. In addition, the mappings $A_{q,i}$ and $A_{q,j}$ to the query image q estimated in the initial retrieval phase are used. For a consistent pair of result images i and j, it holds $A_{q,i} \approx A_{q,j}A_{j,i}$. However, for false positive results caused by repeated patterns or bursty features, the three mappings are typically inconsistent, see Figure 6.5.

6.3.2 Expansion regions selection

Images obtained by the zoom-in query (with a minimal scale change of 2) are first filtered by geometric verification against the query image. Only images with at least t_1 inliers are considered. The estimated mapping of the result images to the query is then used to back-project the images. Consequently, the result images are grouped based on location and scale in the query image. Finally, for each group a geometric consistency test is performed, before the expanded queries are issued.

Choice of the t_1 parameter. The number of matching features as a level of confidence of match correctness has been previously used in query expansion techniques [CPS+07]. In our case, when a significant change of scale is required, the parameter t_1 can be set



query result i result j

Figure 6.5: Geometric consistency test. The solid parallelogram in the query image denotes projected image border of result j through transformation estimated between the query and the result j. The dashed parallelogram in the query image is again the border of result j, now transformed by composition of transformations through result image i. The dashed parallelogram in result i is transformed image boundary from result j.

much lower than in standard query expansion. It stems from the fact that the number of features density drops quadratically with the scale of the feature – this is caused by the scale dependent non-maxima suppression in the feature detectors. Therefore, the probability of random geometric match is substantially decreased by the requirement of zooming-in. Experimentally, we have found that as little as two consistent features with a query image ($t_1 = 2$) provides acceptable results. Note that this result is in combination with large vocabulary (16M visual words) and the novel geometric consistency test among the result images. In our experiments, we set $t_1 = 4$.

Result grouping. A simple greedy algorithm is used to group the result images for the hierarchical query expansion. First, a place (a pixel) in the query image covered by the largest number of images is found. The image with the highest estimated scale change covering that pixel is selected as a cluster seed. Images with at least 50% overlap with the seed images are included in the cluster. The cluster is removed and the whole procedure is repeated.

Note that unlike in [WL13], the goal is not to produce an iconic view of the detail, but to group images relevant to certain detail for the purpose of query expansion.

Each cluster is subject to a geometric consistency test. If the size of the cluster is large than 6 images, the 6 images with the largest scale change are used for the query expansion for efficiency reasons. If a cluster contains only a single image, it is discarded, unless it has at least t_2 geometrically consistent features with the query image and thus small probability of being a false match. In the presented experiments $t_2 = 10$. An example of clusters of geometrically consistent images are shown in Figure 6.6.

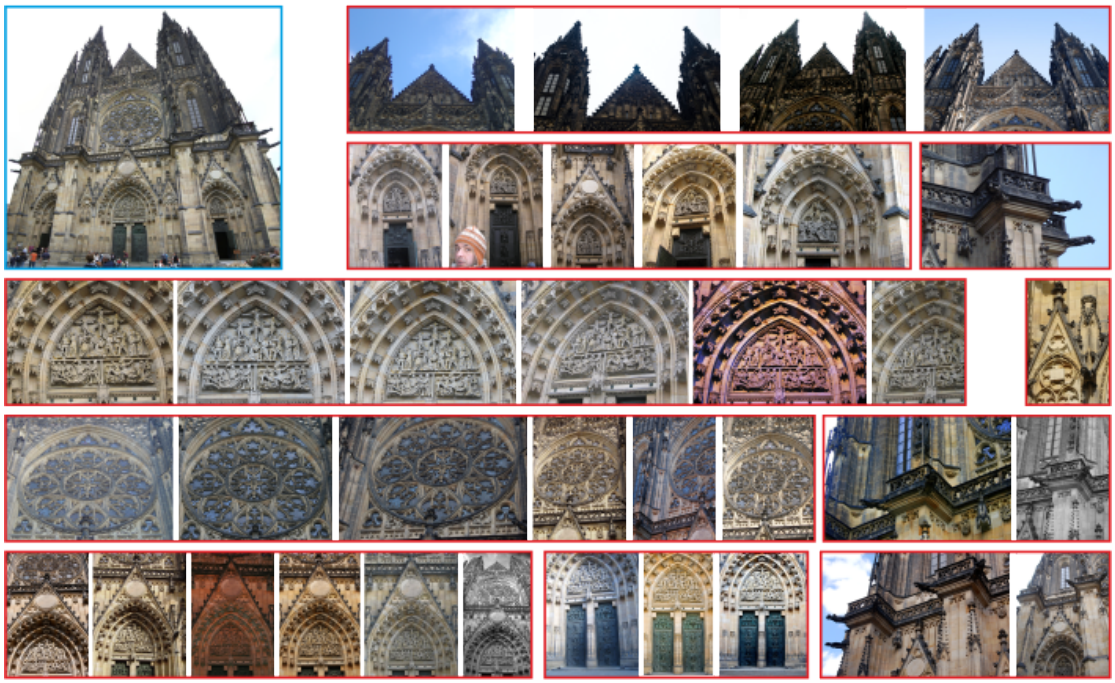


Figure 6.6: Groups of images selected for expanded queries. The query is shown in the top-left corner with a blue border. Groups selected for expanded queries have red borders.

Figure 6.7: A common issue for image clustering methods. Totally unrelated sites linked through an artificial tag superimposed over the images.

6.3.3 Discussion

The proposed method can be seen as a special type of image clustering. In image clustering, false links (*i.e.* links not related to the scene photographed) can be introduced by users inserting visual tags into their images, as depicted in Figure 6.7. These links are difficult to detect and complex heuristic are often used. Our approach naturally eliminates such issue, as a large scale change is required, while the tags, no matter how complex, typically have a fixed scale.

6.4 Experiments

To our knowledge there is no standard dataset with an evaluation protocol suitable for testing zooming capabilities. A search engine was built for Oxford-105k dataset (see

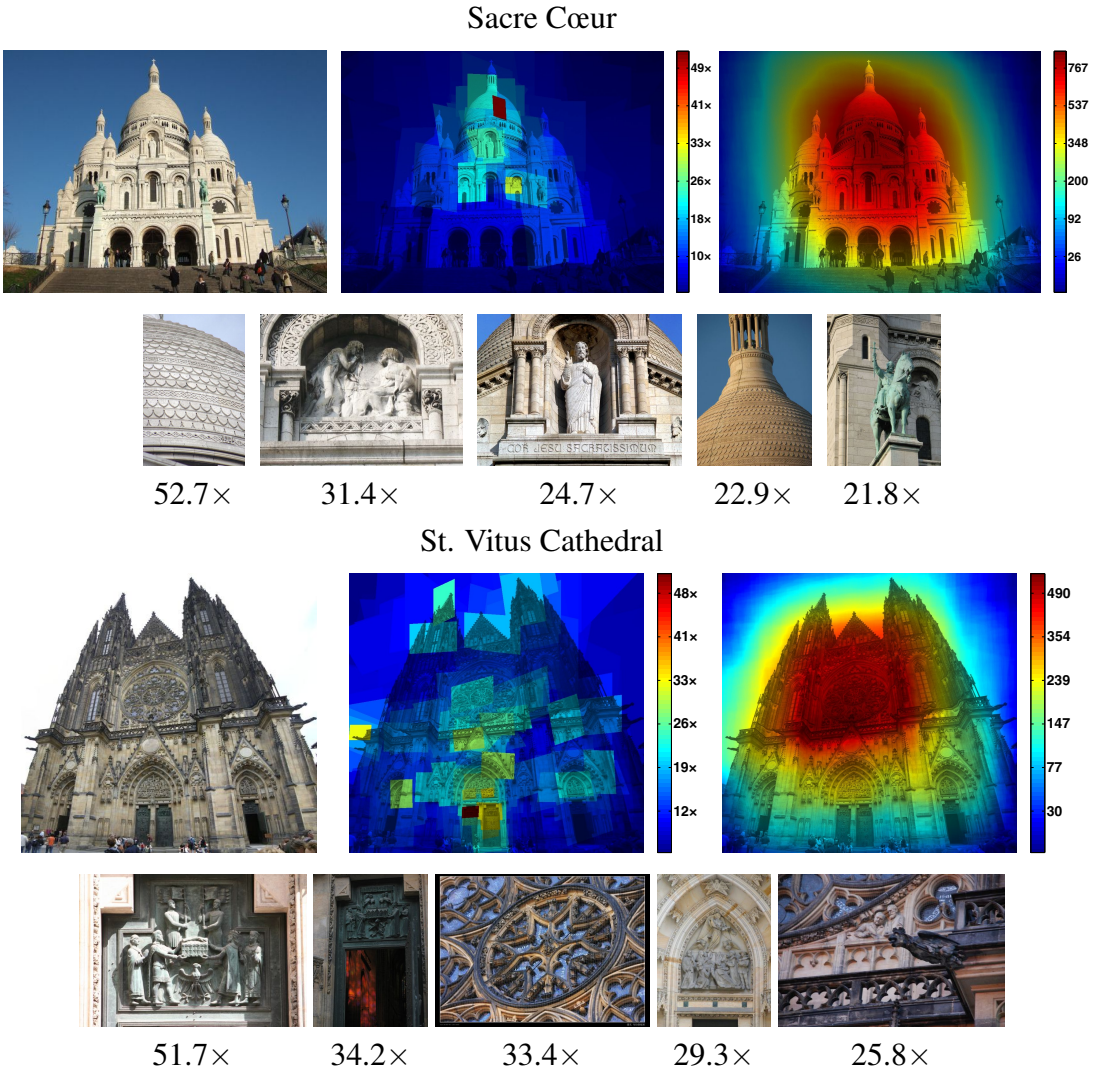


Figure 6.8: Top rows left to right: the original image, the largest scale change for each pixel, and the frequency of the details with zoom larger than 3. Bottom rows show examples (omitting duplicates) of details with the largest relative scale change.

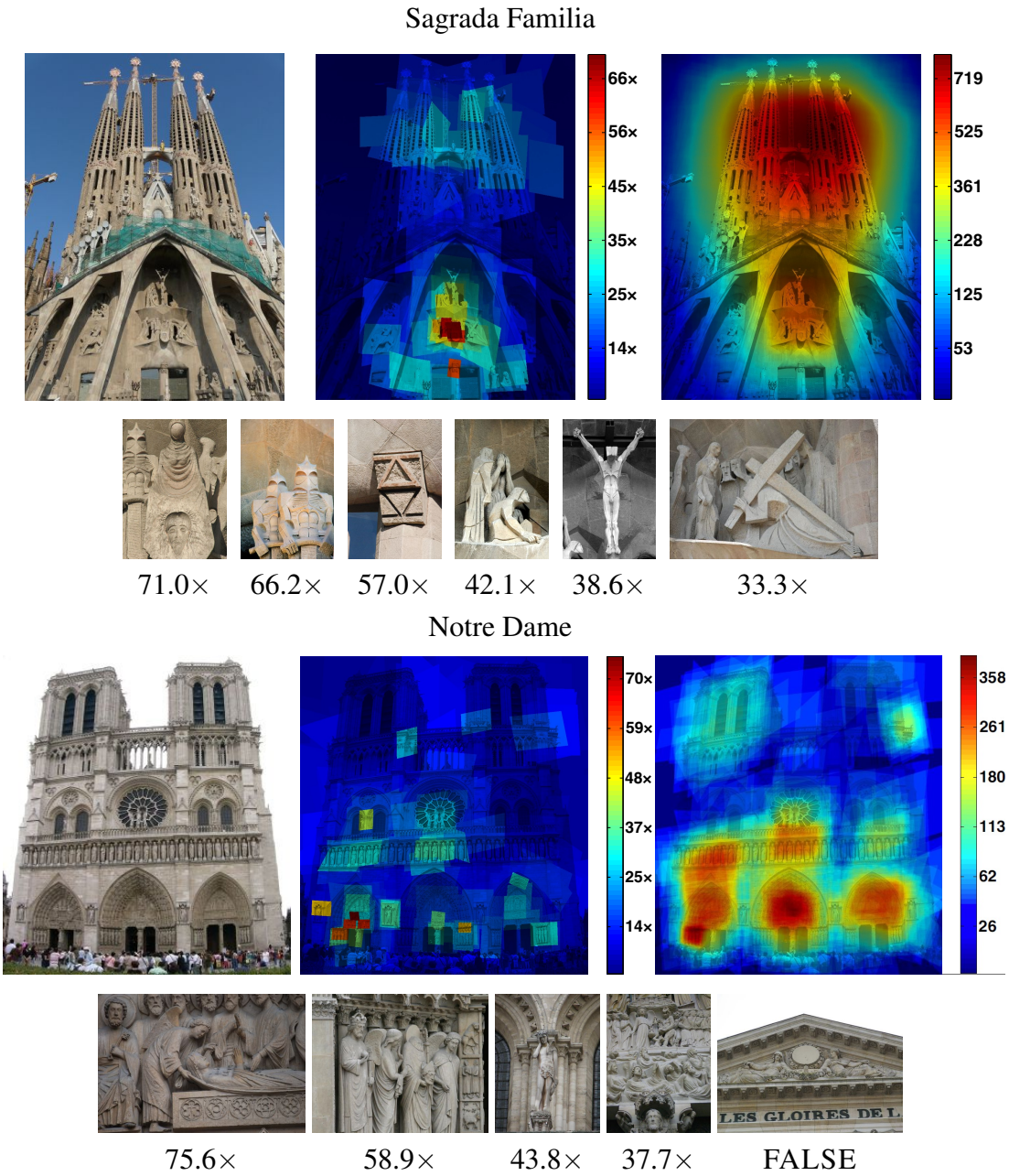


Figure 6.9: Top rows left to right: the original image, the highest scale change for each pixel, and the frequency of the details with zoom larger than 3. Bottom rows show examples of details with their relative scale change. Some false positives, marked FALSE, were detected as details of the query image.

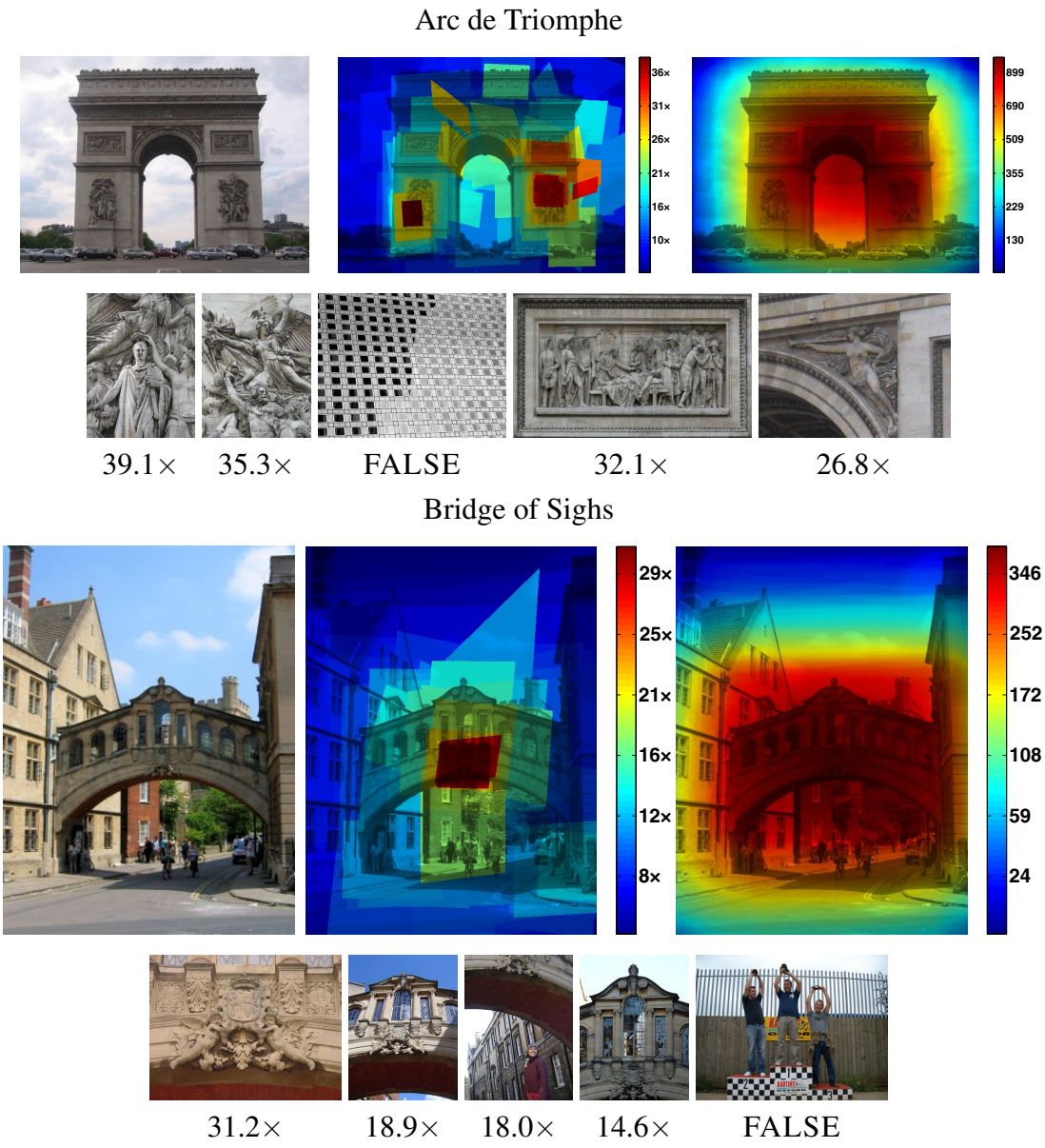


Figure 6.10: Top rows left to right: the original image, the highest scale change for each pixel, and the frequency of the details with zoom larger than 3. Bottom rows show examples of details with their relative scale change. Some false positives, marked FALSE, were detected as details of the query image.

Section 1.4). 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 515,000 images downloaded from Flickr, searching for tags of famous landmarks, European countries and cities, and architectonic keywords.

6.4.1 Design choices.

Following the results from previous chapters, multi-scale Hessian-affine features [MTS$^+$05] were detected and described by the SIFT descriptor [Low04]. Two level fine vocabulary with 16 millions visual word was used (see Chapter 3). In this case the vocabulary is learned on all 620,000 images (nearly 1.3×10^9 SIFT descriptors).

As in [PCM09], 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 [SGP12].

6.4.2 Zoom-in

To demonstrate the method, we chose two queries from Sagrada Familia and nine queries from the Oxford dataset. The queries and the top results retrieved with the zoom-in method are shown in Figure 6.11. 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 6.1 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 neighbor (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).

6.4.3 Scale change

This experiment shows scale change in the highest ranked images for three different settings. The standard retrieval system, the zoom-in and the method with query expansion designed for discovering as many details as possible. Figure 6.12 shows that last method retrieve a large portion of detailed images. Figure 6.13 shows that retrieved images with the zoom-in resp. detail mining method are more informative than images from standard retrieval.

In the case of the astronomical clock from Figure 6.4, the displayed images – local maxima in the resolution, are in our method ranked between first 400 (10 of them in first 22) in comparison to ranks from 8469-392533 in standard retrieval. As the length of the shortlist is limited because of efficiency, these images are not even considered for verification in standard retrieval and thus are surrounded by false positives.

Figure 6.11: 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 6.1: 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.

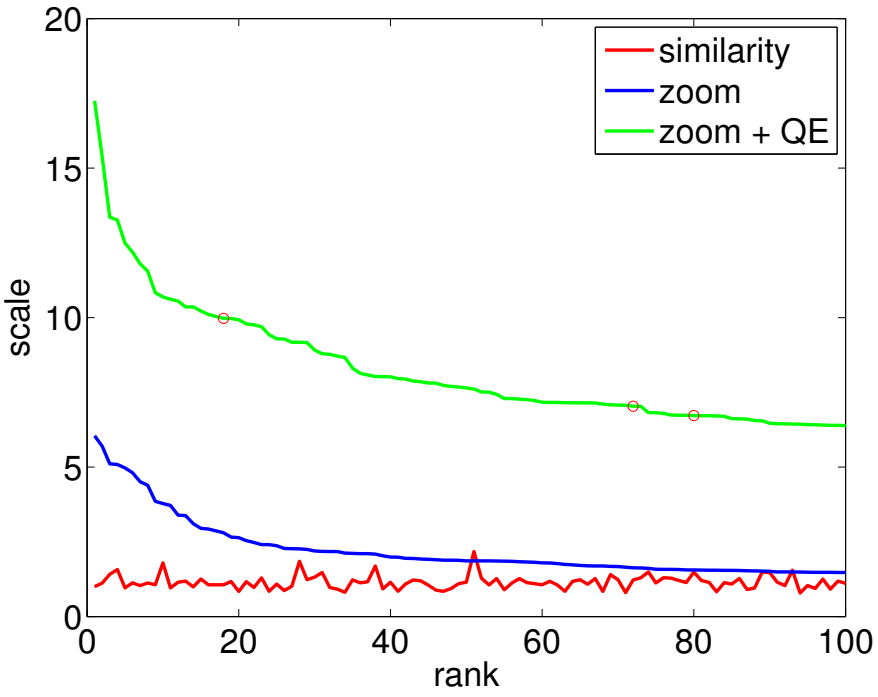


Figure 6.12: Scale change the for 100 top scored images. Comparison of the standard nn, zoom-in and zoom-in with query expansion (QE) methods. Red circles mark false positives. The query and the first few results are shown in Figure 6.13.

Figure 6.13: A comparison of the highest ranked images for three different settings. The query image on the left is used in each case. The first line shows top 11 results of the standard system optimizing average precision – (i.e. similarity). The second line shows the top 15 images optimizing zoom-in. The last two lines show the top 23 images after query expansion of the chosen groups of images. Note that while NN search retrieves many very similar results, the result of our approach are more informative but also more prone to false positives (the image marked red with rank 17 is a false positive).

6.4.4 Maximum scale versus frequency

Figures 6.8 to 6.10 show further examples and a comparison of the two methods. The maximum scale is typically achieved by images of some interesting detail or eventually by a false match, as shown in Figures 6.9 and 6.10. The false matches are rare and are caused by the query expansion. The spatial consistency test is not performed on the final results to reduce the response time.

On the other hand, the frequency distribution is dominated by a relatively small scale change from the query image. Most of such images show people in front of the landmark with a part of the building in the background. The biggest difference between the location of the details and frequently photographed spots is in the Arc de Triomphe, where many people have their photo taken upwards with the arc above them.

6.5 Conclusions

We have formulated novel problems for large scale image retrieval demonstrated that the classical ranking of the images based on similarity is only one of 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 novel retrieval methods were proposed adding zooming-in and out capabilities and answering 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.

Next, We have formulated novel problems that straddle the boundary between image retrieval and data mining: (i) given a query image, find images for every pixel location with maximum resolution *"Highest resolution transform"* and (ii) return the frequency with which a pixel is photographed.

To solve these problems we introduced two novel methods: hierarchical query expansion method that exploits the DAAT inverted files and a new geometric consistency verification step that is sufficiently robust to to prevent topic drift.

Experiments show that the proposed methods are able to retrieve surprisingly fine details on the tested landmarks, even those that are hardly noticeable by inspection in the query image.

Chapter 7

Conclusions

The problem of large-scale content-based image retrieval has been studied. We have contributed to number of components of the standard bag-of-words approach.

We presented a novel similarity measure for measuring image feature distances. The similarity function is learned in an unsupervised manner using geometrically verified correspondences obtained with an efficient clustering method on a large image collection. The similarity measure requires no extra space in comparison with the standard bag-of-words method. Experimentally we showed that the novel similarity function achieves mean average precision that is superior to any result published in the literature on the standard Oxford and Paris datasets. At the same time, retrieval with the proposed similarity function is faster than the reference method.

We showed that using two-layer hierarchical approach for construction enables to built a larger vocabulary, which performs better and faster. We propose a simple balancing method, which helps to keep imbalance factor low.

Next, we proposed two modifications to the query expansion: First, the spatial verification and re-ranking step was improved by incrementally building a statistical model of the query object and its spatial context. Experiments show that the relevant spatial context significantly improves retrieval performance and achieves state-of-the-art results if it is used in the query expansion. Second, a method capable of preventing *query expansion failure* caused by the presence of confusers was introduced. Unlike other approaches, the proposed method handles the presence of confusers in the query region on-the-fly, with no prior learning step required. We achieve performance that is comparable to the state-of-the-art without the need for off-line and potentially time-consuming processing that is difficult to execute in a continuously updated database.

Finally, we formulated novel problems of image retrieval which despite the fact of having no analogy in text-retrieval can be often very useful to the user. We proposed methods for answering the queries such as *What is this?* and *Where is this?*, the method for discovering all possible details from the given picture and the method which creates a heat-map of the frequency with which is every pixel of the given query image photographed.

Bibliography

[ALR03] Y. Aasheim, M. Lidal, and K.M. Risvik. Multi-tier architecture for web search engines. *Web Congress, 2003. Proceedings. First Latin American*, 2003. 15

[ASS+09] S. Agarwal, N. Snavely, I Simon, S. Seitz, and R. Szeliski. Building Rome in a day. In *Proc. ICCV*, 2009. 3, 27, 28

[AZ12] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, pages 2911–2918, 2012. 14, 62

[BDH03] L.A. Barroso, J. Dean, and U. Holzle. Web search for a planet: The google cluster architecture. *Micro, IEEE*, 23, 2003. 15

[BSAS95] Ch. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. *NIST SPECIAL PUBLICATION SP*, pages 69–69, 1995. 16

[BTVG06] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006. 13, 14

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, ISBN: 020139829, 1999. 12, 15, 34

[CBK+11] D. M. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 737–744, 2011. 3

[CM10a] O. Chum and J. Matas. Large-scale discovery of spatially related images. *IEEE PAMI*, 32:371–377, 2010. 3, 27, 29, 62

[CM10b] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. CVPR*, 2010. 13, 51, 53, 54

[CMK03] O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. In *Pattern Recognition*, pages 236–243. Springer, 2003. 16

[CMP08] J. Cech, J. Matas, and M. Perdoch. Efficient sequential correspondence selection by cosegmentation. In *Proc. CVPR*, 2008. 28

[CMPM11] 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. 10, 11, 60

[CN08] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008. 50

[CPM09] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *CVPR*, 2009. 29

[CPS+07] 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. 3, 16, 19, 33, 42, 49, 62, 63

[fac] facebook. URL: `http://facebook.com`. 2

[fli] Flickr. URL: `http://www.flickr.com`. 2

[FSN07] F. Fraundorfer, H. Stewénius, and D Nistér. A binning scheme for fast hard drive based image search. In *Proc. CVPR*, 2007. 31

[FTVG04] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *Proc. ECCV*, 2004. 28

[gooa] Google+. URL: `http://plus.google.com`. 2

[goob] Google Goggles. URL: `https://www.google.com/search?q=google+gogles`. 3

[gooc] Google Streetview. URL: `http://books.google.com/help/maps/streetview/`. 2

[GR01] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer, 2001. 28

[HBW07] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *Proc. ICCV*, 2007. 24

[hol] INRIA Holidays dataset. URL: `http://lear.inrialpes.fr/jegou/data.php`. 6

[ijc12] Learning a Fine Vocabulary data files, 2012. URL: `http://mikulik.sk/publications/ijcv2012/index.html`. 41

[ins] Instagram. URL: `http://instagram.com`. 2

[JDS08] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008. 6, 13, 39

[JDS09] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Proc. CVPR*, 2009. 13, 19, 35, 42

[JDS10] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, 2010. 18, 24, 31, 34, 35

[JDS11] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE PAMI*, 33(1):117–128, 2011. 15

[JDSP10] H. Jégou, M. Douze, C. Schmid, and P. Parez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010. 13, 14, 17

[JH99] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999. 14

[JHS07] H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Proc. CVPR*, 2007. 13

[KSP10] J. Knopp, J. Sivic, and T. Pajdla. Avoding confusing features in place recognition. In *Proc. ECCV*, 2010. 51

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 12, 13, 14, 17, 24, 28, 69

[LWZ+08] X. Li, C. Wu, C. Zach, S Lazebnik, and J. M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. ECCV*, 2008. 27

[MCM13] A. Mikulik, O. Chum, and J. Matas. Image retrieval for online browsing in large image collections. In *Similarity Search and Applications*, 8199, pages 3–15. Springer Berlin Heidelberg, 2013. 11

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC.*, volume 1, pages 384–393. BMVA, September 2002. 13

[ML09] M. Muja and D.G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009. 31

[MM07] K. Mikolajczyk and J. Matas. Improving sift for fast tree matching by optimal linear projection. In *Proc. ICCV*, 2007. 24

[MPCM10] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In *Computer Vision – ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2010. 10

[MPCM13] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning vocabularies over a fine quantization. *International Journal of Computer Vision*, 103(1):163–175, 2013. 10

[MS02] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, pages 128–142. Springer, 2002. 13

[MS04] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004. 13

[MTS+05] 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. 13, 69

[nok] Nokia San Francisco. URL: `http://www.nn4d.com/site/global/developer_resources/nokia_data_share/overview/p_overview.jsp`. 2

[NS06] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006. 13, 17, 18, 31

[OJA09] Whyte O., Sivic J., and Zisserman A. Get out of my picture! internet-based inpainting. In *Proceedings of the 20th British Machine Vision Conference, London*, 2009. 3

[OT06a] 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. 15

[OT06b] 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*, 155:23–36, 2006. 17

[pan] Panoramio. URL: `http://www.panoramio.com/`. 2

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007. 6, 13, 15, 19, 31, 35

[PCI+08] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR*, 2008. 6, 17, 18, 19, 42

[PCM09] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009. 15, 19, 35, 38, 39, 42, 69

[pic] Picasa. URL: `http://picasa.google.com`. 2

[PLSP10] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proc. CVPR*, 2010. 13

[SB97] G. Salton and Ch. Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, pages 355–364, 1997. 16

[SBS07] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *Proc. CVPR*, 2007. 3

[SGP12] H. Stewénius, S. H. Gunderson, and J. Pilet. Size matters: exhaustive geometric verification for image retrieval. In *Proc. ECCV*, pages 674–687. Springer, 2012. 12, 60, 69

[SSS06] N. Snavely, S. Seitz, and R. Szeliski. Photo Tourism: exploring photo collections in 3D. In *Proc. ACM SIGGRAPH*, pages 835–846, 2006. 3

[SZ02] 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. 13

[SZ03] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, pages 1470 – 1477, Oct 2003. 2, 7, 13, 14, 15, 17, 33, 50

[TAJ10] R. Tavenard, L. Amsaleg, and H. Jégou. Balancing clusters to reduce response time variability in large scale image search. Research Report RR-7387, INRIA, 09 2010. 31

[TL09] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop LAVD*, 2009. 52

[WHB09] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *Proc. CVPR*, 2009. 13

[WL11] T. Weyand and B. Leibe. Discovering favorite views of popular places with iconoid shift. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1132–1139, 2011. 3

[WL13] Tobias Weyand and Bastian Leibe. Discovering details and scene structure with hierarchical iconoid shift. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013. 62, 64

[www] World Wide Web size. URL: `http://www.worldwidewebsize.com/`. 2

[Zhu04] M. Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2004. 5

Appendix A Resumé in Czech language

Tato práce se zabývá vyhledáváním obrázků a specifických objektů v obrazových databázích. Na vstupu uživatel zadá obrázek objektu resp. scény a vyhledávací stroj vrátí obrázky stejného objektu resp. scény z databáze. Teze se zaměřuje na *bag-of-words* přístup, který je jedním z nejefektivnějších pro tento typ úlohy. Specifický objekt může pokrývat pouze část obrázku nebo může být z části překrytý jiným objektem. Práce vylepšuje více částí standardních bag-of-wors postupů.

Nová similaritní funkce je definovaná pro bag-of-words vyhledávání obrázků. Tato funkce je naučená bez učitele, oproti standardní metodě nevyžaduje extra pamětový prostor a je více diskriminabilní než eukleidovský L2 *soft assignment* nebo *Hamming embedding*. Navrhovaná similaritní funkce dosahuje na standardních databázích vyšší *mean average precision* než všechny dosud publikované výsledky v literatuře.

Jsou studovány účinky velmi jemné kvantizace u velkých vizuálních slovníků (až 64 milionů slov) a ukazuje se, že výsledky vyhledávače specifických objektů se zlepšují se zvyšujícím se množstvím slov. Toto pozorování je v rozporu s předešlými publikovanými výsledky. Dále ukazujeme, že s velikostí slovníku se zvyšuje rychlost *tf-idf* skórování.

Všechny *state-of-the-art* výsledky vyhledávačů publikované v literatuře byly dosažené s použitím *query expansion* kroku, který zásadně vylepšuje kvalitu vyhledávání. Přinášíme tři modifikace automatické query expansion: (i) metodu předcházející selhání query expansion kroku vzniklou přítomností tzv. *confusers*, (ii) vylepšenou geometrickou verifikaci, která inkrementálně vytváří statistický model objektu s přibývajícími verifikovanými obrázky a (iii) učení relevantního geometrického kontextu, který zásadně zlepší výsledky, pokud je využitý v query expansion.

Všechny tři vylepšení query expansion kroku byly testované na standardních databázích Paříž a Oxford, kde dosáhly state-of-the-art výsledky.

Nakonec byly formulovány nové úlohy vyhledávání. Ukázali jsme, že klasické uspořádání výsledků založené na podobnosti obrázků odpoví pouze jeden z možných dotazů uživatele. Místo vyhledávání nejvíce se podobajících obrázků navrhujeme metody přibližování a oddalování, které zodpoví otázky *"Co je to?"* a *"Kde je to?"*.

Formulujeme další dvě úlohy: (i) pro každý pixel zadaného obrázku nalezni jeho maximální rozlišení v obrázcích dané databáze a (ii) pro každý pixel zadaného obrázku vrať četnost jeho výskytu v databázi. Tyto úlohy postavené na zoom-in a zoom-out metodách vyžadují dvě nové techniky: hierarchickou "query expansion" a verifikaci geometrické konzistence na nalezených obrázcích, která je dostatečně robustní, aby předešla odklonění se od původního objektu v průběhu vyhledávání. Experimenty ukazují, že navrhované metody dokáží najít překvapivě drobné detaily na testovaných obrázcích, a to i detaily jenom těžko vyditelné pouhým okem.

Appendix B Author's Publications

Journal Paper

[MPCM13] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning vocabularies over a fine quantization. *International Journal of Computer Vision*, 103(1):163–175, 2013.

Conference Papers

[MPCM10] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 1–14, Heidelberg, Germany, September 2010. Foundation for Research and Technology-Hellas (FORTH), Springer.

[MCM13] A. Mikulik, O. Chum, and J. Matas. Image retrieval for online browsing in large image collections. In Nieves Brisaboa, Oscar Pedreira, and Pavel Zezula, editors, *Similarity Search and Applications*, 8199, pages 3–15, Heidelberg Germany, October 2013. Springer.

[CMPM11] O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *CVPR 2011: Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 889–896, Los Alamitos, USA, June 2011. IEEE Computer Society, IEEE Computer Society.

Papers not Related to the Thesis

[MMPC10] A. Mikulik, J. Matas, M. Perdoch, and O. Chum. Construction of precise local affine frames. In Roy Sterritt, editor, *20th International Conference*

on Pattern Recognition (ICPR'2010), page 5, 10662 Los Vaqueros Circle, Los Alamitos, California USA, August 2010. IEEE Computer Society.

[MBD+08] A. Mikulik, S. Basovnik, M. Dekar, P. Jusko, D. Obdrzalek, R. Pechal, T. Petrusek, and R. Pitak. Logion - a robot which collects rocks. In La Ferté-Bernard, editor, *Proceedings of the International Conference on Research and Education in Robotics - EUROBOT 2008*, SRH Hochschule Heidelberg, May 2008. MATFYZPRESS.

[BMMO10] S. Basovník, L. Mach, A. Mikulík, and D. Obdržálek. Detecting scene elements using maximally stable colour regions. *Research and Education in Robotics-EUROBOT 2009*, pages 107–115, 2010.

Appendix C SCI Citations of Author's Work

Mikulík, A. - Perďoch, M. - Chum, O. - Matas, J.: Learning a Fine Vocabulary. In Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Proceedings, Part III. Heidelberg: Springer, 2010, p. 1-14. ISSN 0302-9743.ISBN 978-3-642-15557-4. [*cited 17 times*]

- Arandjelovic R, Zisserman A: Smooth Object Retrieval using a Bag of Boundaries. In 2011 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2011. ISSN

- Girod B, Chandrasekhar V, Chen DM, et al.: Mobile Visual Search [Linking the virtual and physical worlds]. In IEEE SIGNAL PROCESSING MAGAZINE, 2011. ISSN 1053-5888

- Greuter M, Rosenfelder M, Blaich M, et al.: Obstacle and Game Element Detection with the 3D-Sensor Kinect. In RESEARCH AND EDUCATION IN ROBOTICS - EUROBOT 2011, 2011. ISSN 1865-0929

- Jegou H, Tavenard R, Douze M, et al.: SEARCHING IN ONE BILLION VECTORS: RE-RANK WITH SOURCE CODING. In 2011 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2011. ISSN 1520-6149

- Johns E, Yang GZ: From Images to Scenes: Compressing an Image Cluster into a Single Scene Model for Place Recognition. In 2011 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2011. ISSN

- Kuang YB, Astrom K, Kopp L, et al.: Optimizing Visual Vocabularies Using Soft Assignment Entropies. In COMPUTER VISION - ACCV 2010, PT IV, 2011. ISSN 0302-9743

- Kuang YB, Byrod M, Astrom K: Supervised Feature Quantization with Entropy Optimization. In 2011 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS (ICCV WORKSHOPS), 2011. ISSN

- Qin DF, Gammeter S, Bossard L, et al.: Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors. In 2011 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2011. ISSN 1063-6919

- Wang ZX, Zhao Q, Chu D, et al.: SELECT INFORMATIVE FEATURES FOR RECOGNITION. In 2011 18TH IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), 2011. ISSN 1522-4880

- Arandjelovic R, Zisserman A: Three things everyone should know to improve object retrieval. In 2012 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012. ISSN 1063-6919

- Gao K, Zhang YD, Luo P, et al.: Visual Stem Mapping and Geometric Tense Coding for Augmented Visual Vocabulary. In 2012 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012. ISSN 1063-6919

- Chen YZ, Dick A, Li X: Visual Distance Measures for Object Retrieval. In 2012 INTERNATIONAL CONFERENCE ON DIGITAL IMAGE COMPUTING TECHNIQUES AND APPLICATIONS (DICTA), 2012. ISSN

- Shen XH, Lin Z, Brandt J, et al.: Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking. In 2012 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012. ISSN 1063-6919

- Dillard SE, Henry MJ, Bohn S, et al.: Coherent Image Layout using an Adaptive Visual Vocabulary. In IMAGE PROCESSING: MACHINE VISION APPLICATIONS VI, 2013. ISSN 0277-786X

- Chen YZ, Dick A, Li X, et al.: Spatially aware feature selection and weighting for object retrieval. In IMAGE AND VISION COMPUTING, 2013. ISSN 0262-8856

- Qi SY, Luo YP: Relevance of Useful Visual Words in Object Retrieval. In FIFTH INTERNATIONAL CONFERENCE ON DIGITAL IMAGE PROCESSING (ICDIP 2013), 2013. ISSN 0277-786X

- Xie Y, Jiang SQ, Huang QM: Weighted visual vocabulary to balance the descriptive ability on general dataset. In NEUROCOMPUTING, 2013. ISSN 0925-2312

Chum, O. - Mikulík, A. - Perďoch, M. - Matas, J.: Total Recall II: Query Expansion Revisited. In CVPR 2011: Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2011, p. 889-896. ISSN 1063-6919.ISBN 978-1-4577-0393-5. [*cited 9 times*]

- Torii A, Sivic J, Pajdla T: Visual localization by linear combination of image descriptors. In 2011 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS (ICCV WORKSHOPS), 2011. ISSN

- Arandjelovic R, Zisserman A: Three things everyone should know to improve object retrieval. In 2012 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012. ISSN 1063-6919

- Chen YZ, Dick A, Li X: Visual Distance Measures for Object Retrieval. In 2012 INTERNATIONAL CONFERENCE ON DIGITAL IMAGE COMPUTING TECHNIQUES AND APPLICATIONS (DICTA), 2012. ISSN

- Chen YZ, Li X, Dick A, et al.: Boosting Object Retrieval With Group Queries. In IEEE SIGNAL PROCESSING LETTERS, 2012. ISSN 1070-9908

- Shen XH, Lin Z, Brandt J, et al.: Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking. In 2012 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012. ISSN 1063-6919

- Feng DY, Yang J, Liu CX: An efficient indexing method for content-based image retrieval. In NEUROCOMPUTING, 2013. ISSN 0925-2312

- Gong MY, Sun LF, Yang SQ, et al.: Find where you are: a new try in place recognition. In VISUAL COMPUTER, 2013. ISSN 0178-2789

- Chen YZ, Dick A, Li X, et al.: Spatially aware feature selection and weighting for object retrieval. In IMAGE AND VISION COMPUTING, 2013. ISSN 0262-8856

- Chen YZ, Li X, Dick A, et al.: Ranking consistency for image matching and object retrieval. In PATTERN RECOGNITION, 2014. ISSN 0031-3203

Mikulík, A. - Perďoch, M. - Chum, O. - Matas, J.: Learning Vocabularies over a Fine Quantization. International Journal of Computer Vision. 2013, vol. 103, no. 1, p. 163-175. ISSN 0920-5691. [*cited once*]

- Sluzek A: Inverted Indexing in Image Fragment Retrieval using Huge Keypoint-Based Vocabularies. In 2013 11TH INTERNATIONAL WORKSHOP ON CONTENT-BASED MULTIMEDIA INDEXING (CBMI 2013), 2013. ISSN 1949-3983