From Single Image Query to Detailed 3D Reconstruction

Johannes L. Schönberger[1], Filip Radenović[2], Ondrej Chum[2], Jan-Michael Frahm[1]
[1]Department of Computer Science, The University of North Carolina at Chapel Hill
[2]CMP, Faculty of Electrical Engineering, Czech Technical University in Prague
jsch@cs.unc.edu, radenfil@cmp.felk.cvut.cz, chum@cmp.felk.cvut.cz, jmf@cs.unc.edu

Abstract

Structure-from-Motion for unordered image collections has significantly advanced in scale over the last decade. This impressive progress can be in part attributed to the introduction of efficient retrieval methods for those systems. While this boosts scalability, it also limits the amount of detail that the large-scale reconstruction systems are able to produce. In this paper, we propose a joint reconstruction and retrieval system that maintains the scalability of large-scale Structure-from-Motion systems while also recovering the often lost ability of reconstructing fine details of the scene. We demonstrate our proposed method on a large-scale dataset of 7.4 million images downloaded from the Internet.

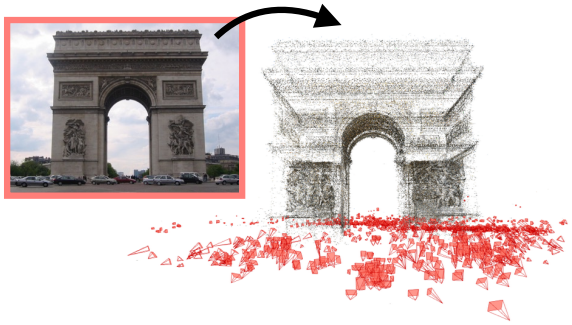

Figure 1. Arc de Triomphe, Paris. 3D reconstruction from a single input image (red inset) using 2,640 views around the landmark from a 7.4M image database. Only imagery is used (no GPS or text). The scene contains 395,431 points and the surface resolution reaches the order of 1mm in the most photographed areas.

1. Introduction

In the last decade, computer vision has made great progress in the areas of image retrieval and 3D modeling. Current image search engines operate on web-scale image collections and are able to localize specific objects and landmarks, and aid user-friendly content browsing. In the field of reconstructing scenes from images and videos, arguably the biggest steps have been made in 3D modeling from unordered Internet photo collections. A natural step forward is to address the problem of obtaining a detailed 3D model of an object depicted in a single, user-provided photograph.

Structure-from-Motion (SfM) systems have been extended from modeling scenes from a few thousand images [29, 30] to modeling city-scale photo collections of millions of images [7, 9]. Early photo collection reconstruction systems leverage exhaustive matching of image pairs to determine possible overlapping image pairs. This is generally quadratic in the number of images and features. Hence, this approach does not scale and is not applicable to datasets containing thousands or even millions of images, which are commonly available. However, exhaustive matching guarantees the discovery of all possible camera overlaps. To achieve scalability, the current state-of-the-art large-scale reconstruction systems abandon the exhaustive pairwise overlap determination. Instead, modern systems leverage image retrieval algorithms [21, 3, 4], or image-clustering techniques to identify overlapping images during reconstruction, as demonstrated by the systems of Agarwal *et al*. [1] and Frahm *et al*. [7]. While the introduction of image retrieval was essential to boosting the scalability of reconstruction methods on large datasets, it also severely impacted the ability to reconstruct fine details of the scene. This problem stems from the fact that the image pairs showing the details are often absent from the retrieval results. This is unsatisfactory, as for applications such as photo field of view extension using unordered photo collections, recently proposed by Zhang *et al*. [36], it is desirable to have the details present in the reconstruction.

The lack of detail is a result of the employed retrieval approaches [21, 3, 4], which are tuned to obtain images similar in scale and appearance. In this paper, we introduce a tightly-coupled retrieval and SfM system for large-scale reconstruction from unordered photo collections of several million images, which not only recovers the coarse geometry of the scene but specifically focuses on modeling scene details. Our approach achieves this by combining SfM with retrieval across differently scaled scene images.

In order to achieve these detailed reconstructions, our

Figure 2. Reconstruction of the Astronomical Clock in Prague, Czech Republic. Left: 3D model obtained from our retrieval and reconstruction system. Images illustrating the range of registered views from overview images to images of a specific architectural detail are shown alongside the model. Right: visualization of the surface resolution from high resolution in red (approximately 1mm surface resolution as obtained from a known object size in 3D) to low resolution in blue.

system has to overcome the following challenges:

- Achieve a more balanced retrieval of overview and detailed images to provide the images needed for fine-detail reconstruction.
- Overcome the registration uncertainties that result from the large resolution differences between overview images and detailed images.

We resolve these challenges by proposing a tightly-coupled SfM and retrieval system. Establishing an interactive link between the reconstruction system and the retrieval system enables us to control the retrieval characteristics based on the current state-of-the-art 3D reconstruction. This allows us to specifically retrieve images that are required to overcome the challenges of SfM. Our resulting reconstructions from unordered Internet photo collections show high geometric detail while at the same time conveying the structure of the entire scene. An example reconstruction is shown in Figure 2.

2. Related work

Our system simultaneously leverages retrieval and SfM algorithms to achieve the goal of detailed scene reconstruction. In this section, we discuss the relevant state-of-the-art in both areas, before introducing our method in more detail in the following sections.

Exploring a large unordered image collection by user defined image query – the problem of large-scale image retrieval – made significant progress during the last decade. Most of the approaches pose the problem as a nearest neighbor search in a descriptor space, such as bag-of-words [28, 21, 25, 11], VLAD [12, 2], Fischer vectors [24], or exhaustive matching [34, 27]. Recently, Mikulik *et al*. [18] pointed out that the nearest neighbor image search is not optimal for the user, who is typically looking for new image information rather than for near-duplicate images. Novel formulations and efficient methods for extreme change of scale were proposed in [18] and for detailed image mining in [19]. We extend these ideas to identify initial sets of images suitable for 3D reconstruction and then leverage the obtained reconstructions to suggest further image retrieval goals. Instead of targeting the extreme scale changes that are attractive for a human user, the whole spectrum of scale transitions is sampled, which is more suitable for 3D reconstruction. Further, we propose an efficient retrieval method for content-based crawling around a landmark, mining for views connecting multiple sides of the landmark.

Scene reconstruction from Internet photo collections has been introduced in the seminal paper of Snavely *et al*. [29, 30]. This was the first approach to show that SfM for such diverse and unordered collections of thousands of images is possible. The major limitation of this reconstruction system was its limited scalability due to exhaustive image pair overlap evaluation.

To overcome this lack of scalability, Li *et al*. [13] introduced an appearance-based clustering for grouping the images. This allowed modeling from tens of thousands of images on a single PC. Agarwal *et al*. [1] introduced a cloud computing algorithm to perform modeling from 150,000 images on 62 computers in less than 24 hours. The approach leveraged a vocabulary tree based search with query extension [4] to determine overlapping images, followed by approximate nearest neighbor feature matching. While providing scalability, such an approach severely impairs the retrieval of detailed images for registration and reconstruction. Lou *et al*. [15] proposed a modified vocabulary tree based retrieval enforcing diversity in the retrieval results. The proposed reweighting enhances scene coverage for the reconstruction with SfM, but it does not solve the problem of not retrieving detailed images.

Frahm *et al*. [7, 1] extended the approach of Li *et al*. [13] to scale to the reconstruction from millions of images. However, this approach also suffers from the use of recognition

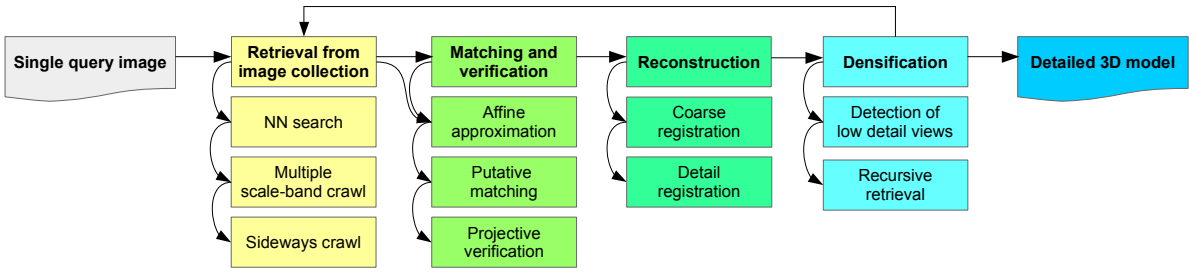

Figure 3. The proposed pipeline that tightly couples image retrieval and SfM.

methods – gist-feature [22] based appearance grouping – that fail to obtain detailed images and thus severely limits the ability to produce fine-detail reconstructions.

Crandall *et al*. [5] proposed a global method that performs SfM based on a MRF optimization. In order to properly initialize this hybrid optimization, their approach requires approximate geo-location priors for the images. While this approach can retrieve geo-located detail images, a large fraction of Internet photo collection photos is not geo-located. Hence, this approach would be very restrictive in our scenario, and it would only register a fraction of the images compared to our approach.

3. Overview

Our proposed pipeline (Figure 3) handles image collections of the size of millions of unordered images. A single, user-provided query image serves as the input to our pipeline. In the first stage, the query image is used as the initial seed for image retrieval (Section 4). The retrieval stage first finds nearest neighbor images, followed by a multiple scale-band crawl to obtain additional views at different zoom levels (Section 4.1). Furthermore, we expand the query by retrieving images to the left and right of the query image in order to obtain additional context around the query image (Section 4.2). As a preparation for the subsequent reconstruction stage, we propose an efficient matching method (Section 5), leveraging the by-products of the retrieval stage to intelligently avoid image pairs that do not overlap. The reconstruction stage employs several methods to overcome the challenges of detailed image registration (Section 6). Finally, we perform additional densification (Section 7) by identifying the low-detail parts of the model and recursively retrieving additional images for another round of reconstruction. In comprehensive experiments (Section 9) on a dataset of millions of images, we demonstrate that the method produces large-scale, high-quality models that also capture the fine details of the scene.

4. Image retrieval

The objective of retrieval for 3D reconstruction is to provide a matching graph with a variety of viewpoints (for the stability of the reconstruction), sequences of images providing a smooth transition between extreme viewpoints or scale changes (to be able to connect different parts and to help disambiguate duplicated structures), and mining for images of further structures in space and scale (to extend the reconstruction and improve the level of detail).

The retrieval engine builds upon bag-of-words representation with fast spatial verification [25]. Hessian affine features [17] are detected (1900 features per image on average) and described by the rotation-variant [23] SIFT descriptor [16]. The descriptors are vector-quantized into 16 million visual words using k-means with approximate nearest neighbor search [20]. Fast (several hundred image pairs per second) spatial verification [25] then estimates an approximate affine transformation between query and result images. To enforce transformation consistency (scale change, translation), the scale and position information for each feature is included in the inverted file [11, 31].

In the proposed method, the initial matching graph is obtained from the image collection using the query image as an entry point. Depending on the intended result (either detailed reconstruction of the scene visible in the query image, or detailed reconstruction of the whole neighborhood of the query image) different mining techniques are used to generate the initial matching graph. Once a (partial) reconstruction is available, the same techniques are applied to incrementally extend the reconstruction; see Section 7.

4.1. Multiple scale-bands

To retrieve relevant images of various levels of detail (and/or different amounts of context), we build on the approach of hierarchical query expansion [19]. Unlike in [19], we are not interested in the extreme scale changes, but rather in an image sequence capturing a smooth transition in scale to support stable SfM estimation.

The hierarchical query expansion proceeds as follows. An initial query encouraging change of scale is issued with the query image. To reflect the scale change in image ranking, we use document at a time (DAAT) scoring [31] exploiting geometry stored in an inverted file. The results of this initial query are clustered in scale-space. Each spatial image cluster is then used to issue a new expanded query [4], which retrieves further details at the given location. Figure 4 shows four scale bands – context (zoom out),

Figure 4. Terracotta Army, China. Samples of different scale-bands of the initial query image: context of the query image (zoom out – top left), two examples of mid-level detail (zoom in), and three detailed images for each of the mid-level band (rightmost). Two examples of the left and right side of the query are shown in the bottom left.

original scale, two examples of mid-level detail (zoom in), and three detailed images for each of the mid-level bands.

4.2. Sideways crawl

Retrieving images of multiple scale-bands, starting from a single query image, yields the whole spectrum of image scales, but typically only from a single viewing direction. However, many interesting scene parts are often located around the corners of or next to the observed landmark. In this case, the reconstruction significantly benefits from additional sideways crawling around the landmark, in order to obtain more complete and stable models. The crawling is performed to the left or to the right with respect to the original query image. We propose a novel, efficient retrieval method for content-based crawling around an initial point of view. As a result, we successfully mine images connecting multiple sides of the landmark (see Figure 5), or images with a broader view of the whole area of interest in the case of indoor scenes (see Figure 4).

The sideways crawl retrieval consists of two stages. The first stage allows us to specifically crawl for images in different directions (left and right). In the second stage, the initial set of retrievals is extended by additional images from the desired direction.

The first stage leverages the estimated geometric transformation (an affine transformation in our case) between the query image and the results. When taking a step to the right, for instance, features on the right-hand side of the query image should match a sufficient number of features on the left-hand side of the result image. This can be achieved by geometric re-ranking of the shortlisted results or more efficiently using the DAAT approach [31]. Additional geometric analysis, such as estimating the position of the horizontal vanishing point through homography fitting, can be performed at additional computational effort.

To retrieve a larger set of relevant images that contain novel image information, additional queries are executed using the top ranked images from the first stage. In these expanding queries, only features from areas not visible in the original query image are considered. Finally, retrieved images are merged and then re-ranked based on the amount of viewpoint change.

To obtain reconstruction of a landmark from all sides, the sideways crawl is repeated, as illustrated in Figure 5. In order to also reconstruct details on all sides of the landmark, each sideways step is followed by multiple scale-band mining. For further examples on the sideways crawl, see an indoor view of the Terracotta Army landmark in China (Figure 4). Having introduced the image-retrieval system, we next detail our SfM system that exploits the unique characteristics of the proposed retrieval system.

5. Matching and geometric verification

The state-of-the-art SfM systems [7, 1] have achieved impressive results on city-scale reconstructions from unordered photo collections. One frequent reason for the lack of detail reconstruction is caused by the nature of conventional image retrieval. That is, when starting from an overview image of the scene, the images of the scene details are not retrieved as nearest neighbors, because of low overlap or a large number of similar views outranking the detail images [19]. The reason for not registering retrieved detail images in SfM is, that SfM will often require transition images that establish connectivity between the detail view with high surface resolution and the overview image with comparably low surface resolution in the area of the detail view. In this paper, we propose a novel combination of a detail oriented retrieval and SfM system to address the challenge of obtaining 3D models from unordered photo collections that provide complete scene coverage and high geometric resolution for the details in the scene.

Our proposed image retrieval method has a major advantage over traditional vocabulary-based and clustering-based

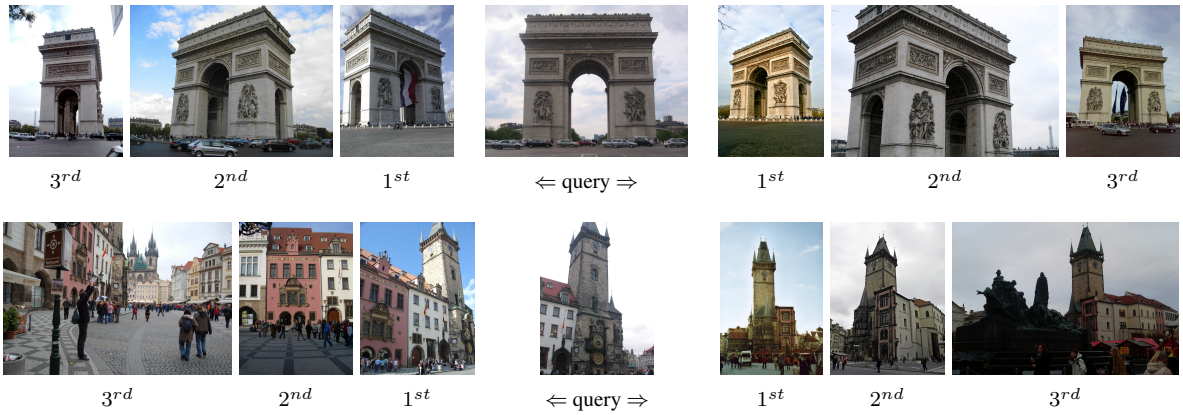

Figure 5. Sideways crawl of Arc de Triomphe (top) and Astronomical Clock (bottom). Sample results of three recursive rightward and leftward queries extending the view of the original query image.

approaches. Since the method performs spatial verification through an affine transformation estimation during retrieval, it obtains a quantitative measure of the scene overlap early on. The number of inliers, treated as a similarity score, is a by-product of robust estimation with RANSAC [6].

Next, our geometric verification of image pairs, *i.e.* the test for a pairwise epipolar geometry, in SfM operates in projective space, estimating an essential or fundamental matrix for moving cameras and a homography for purely rotating cameras. Because the space of affine transformations is a subspace of the space of projective transformations, we can use the existence of an affine transformation between a set of correspondences as a proxy to assume the existence of a projective transformation for a larger set of correspondences between the same image pair. While, in theory, the projective transformations may not exercise all degrees of freedom, the chance of encountering these configurations are extremely low in practice. In fact, for more than 99.9% of image pairs in the experimental datasets (Section 9) we can estimate a valid epipolar geometry or homography when there exists an affine transformation. Hence, if we enforce the existence of an affine transformation, geometric verification in SfM only has to process valid image pairs for reconstruction. This avoids the significant overhead caused by non-overlapping pairs. This makes RANSAC for geometric verification significantly faster, since RANSAC has exponential computational complexity in the number of model parameters and the outlier ratio of the measurements.

Leveraging the early similarity metric, provided by the retrieval system, significantly improves the performance and reliability of our SfM estimation. The reason being that our proposed pipeline can rank the retrieved images by the number of affine transformation inliers and the subsequent geometric verification only spends time on actually overlapping image pairs. This enables us to only match against a limited number of images instead of matching against much larger sets of nearest neighbors as in the case of traditional vocabulary tree based approaches. In all experiments, we attempt to verify a query image to a maximum number of 200 retrieved images, and we empirically found that nearest neighbor images with at least 8 affine transformation inliers to the query image have a very high likelihood of successful registration. Next, we describe the enhancements to our SfM algorithm to reliably obtain reconstructions of the geometric scene details.

6. Reconstruction of Details

Detailed scene reconstruction depends on accurate and reliable camera registration, which is especially challenging for the highest-resolution images in a photo collection. There are three major reasons for this: the dependence of incremental SfM on the order of camera registrations, the reduced redundancy of measurements during registration of the detailed views, and the often challenging geometric configurations for these views. In the following, we examine these challenges and describe our solutions to them.

Generally, the quality of SfM results are dependent on three main factors: First, to attain reliable and precise estimates, no parameter should rely on just a minimal or small set of measurements to enable the compensation of measurement noise (detailed views usually have significantly reduced correspondences to other images). Second, reliability provides us with the ability to detect outliers and determines the degree to which undetected outliers affect our estimates (outlier detection for detailed views is challenging due to their reduced redundancy). Third, the uncertainty of measurements propagates to the uncertainty of the estimated parameters, and the viewing geometry impacts the stability of estimation (very oblique or distant views generally have significantly higher measurement and thus registration uncertainties).

First, incremental SfM is heavily dependent on the order of camera registrations, due to the non-linear nature of bundle-adjustment. This effect becomes especially im-

portant for detailed scene reconstruction, since the different levels of detail are usually only sparsely connected or connected with forward motion views. Forward motion is a particularly challenging situation for SfM, caused by unstable viewing geometry. In our experience, seeding the reconstruction with one of the detailed views and then incrementally growing the model to include less detailed views fails in most of the cases or results in inferior-quality models. We therefore seed the reconstruction with an image that sees a maximal fraction of the scene, effectively ruling out the detailed and extremely zoomed-out views. From there, we gradually extend the model, avoiding abrupt scale and viewpoint changes by ranking cameras for registration based on the amount of currently visible scene structure.

Second, we discuss the effect of reduced redundancy that is often encountered for the images observing the details of the scene. Since images of detailed structure only see a small fraction of the scene, there are generally much fewer images that observe the same features. In this case, conventional nearest neighbor matching produces fewer and shorter tracks, as it fails to match a significant portion of image pairs from the already small set of images that see the same structure. Hence, bundle-adjustment must deal with a significantly reduced redundancy of the measurements, resulting in less accurate and less reliable camera resectioning and structure estimation. Redundancy, and thus the reliability in bundle-adjustment (or, more generally, in maximum likelihood estimation), is determined as the difference of the number of independent observations minus the effective degrees of freedom [32]. Hence, we are provided with two opportunities for improving redundancy: increasing the number of observations and reducing the degrees of freedom. Inherently, the employed retrieval system mitigates the effect of reduced redundancy, in that it reveals significantly more overlapping image pairs. In this manner, we are able to build significantly more and significantly longer tracks than before, which leads to significantly increased redundancy of the measurements compared with standard retrieval systems, such as the vocabulary tree based approaches. Additionally, we restrict ourselves to a relatively simple camera model with a total of 8 degrees of freedom (3 orientation, 3 translation, 1 focal length, and 1 first-order polynomial radial distortion parameter with fixed principal point at the image center).

Third, SfM methods face distinct challenges for difficult geometric configurations of the scene and/or the cameras. Many images of the geometric details of the scene are taken at high zoom levels, *i.e*. with large focal length. In this case, the viewing rays are close to parallel, resulting in high registration uncertainty along the viewing direction. Given that a relatively large displacement along the viewing direction causes only a small change in reprojection error. Hence, we need to have a good initial estimate of the focal length before starting the non-linear refinement in order to achieve convergence to the correct solution. For this purpose, we can use focal length information extracted from EXIF data, if available. However, crowd-sourced photos often lack this information due to modifications, such as resizing, cropping, *etc*. For large zoom factors, it is not enough to simply assume a default focal length [10, 26], inferred from the image dimensions, and use it as an initial estimate for a non-linear refinement. Rather, it is necessary to exhaustively sample an *a priori* specified space of focal lengths during 2D-3D pose estimation. If EXIF information is missing, we uniformly sample 50 focal lengths for opening angles between $[5°, 130°]$ using P3P RANSAC and use the solution with the highest number of inliers, followed by a non-linear refinement of the pose.

Figure 6. Single query images used for landmark reconstruction. From left to right: Astronomical Clock, Bridge of Sighs, Terracotta Army, Arc de Triomphe, Notre Dame, Sagrada Familia.

However, even after these modifications, the camera registration occasionally still fails due to low redundancy, bogus EXIF information, or unfortunate configurations. We detect these cases in order to avoid a cascade of misregistrations due to faulty triangulations from an initially bad camera. These cases can be detected in different stages of the SfM pipeline. First, we detect a small number of inliers during RANSAC pose estimation. Second, a non-linear refinement of an initial registration is performed; faulty registrations typically result in high cost in the non-linear refinement of the pose. Hence, we reject camera registrations that display any of the above properties. Third, bundle-adjustment usually converges to a local minimum for faulty registrations. As a result, it tries to minimize the cost through the use of extreme camera parameters. Whenever we refine the structure and motion in bundle-adjustment, we filter images that have abnormal camera parameters (opening angle outside $[5°, 130°]$, absolute value of radial distortion parameter greater than 1).

7. Densification

Our proposed combined SfM and retrieval system achieves significantly more detailed reconstructions than the state-of-the-art reconstruction systems. However, some parts of the scene naturally have low detail. This occurs when the initial set of images, which is obtained by image retrieval without considering the full 3D scene information, does not provide a sufficient number of detailed images, or even none at all, in certain areas of the structure.

However, we wish to produce complete models with high detail across all parts of the structure. To overcome this lim-

Figure 7. Dense reconstruction of Arc de Triomphe details.

itation, we introduce an incremental strategy to extend the initial 3D model by explicitly mining for high detail in low-resolution parts of the originally reconstructed 3D model. To avoid redundant detail mining everywhere, we propose to perform detail mining on demand after the initial 3D reconstruction. After the initial reconstruction we are able to determine the density of the obtained sparse model and identify the low resolution parts of the 3D model. Then, we attempt to densify the reconstruction only for those parts.

To find images that cover the low-resolution parts, we first determine the highest model-resolution of every 3D point by calculating the spatial extent of every image observation in the world coordinate frame, *i.e*. back-projecting the image pixel to the 3D point into the world coordinate frame. As multiple images from different distances and at different zoom levels potentially see the same structure, the 3D point is assigned the maximum resolution of all of its observations. Given the resolution of the entire structure and the camera poses, we can then identify images that cover the low-resolution parts of the scene. Each image typically only sees a fraction of the entire scene. The median of the observed 3D point resolutions in an image provides us with a meaningful measure of the overall contribution of an image to the surface resolution it contributes, independent of distance to the individual structure or the zoom level. In a final step, we sort all images by their median resolutions and iteratively query for more detail for the top images until no further details are found or a sufficient resolution is achieved. Finally, we connect the new retrievals to the existing reconstruction by only matching the new images using the strategy described in Section 5.

8. Duplicate scene structure

Duplicate, symmetric, or repetitive scene structure is a common pattern in urban environments, posing challenges for incremental SfM due to a potential cascade of camera mis-registrations and faulty triangulations [35, 8, 33]. These camera mis-registrations are caused by symmetric scene structure that is erroneously retrieved and registered by RANSAC based alignment [8].

The existing solutions for the correction of the problem caused by symmetric scene structure is formulated as post-processing of registered camera triplets[35] or post-processing of the entire model [8, 33]. The major drawback of all of these approaches is, however, that mis-aligned cameras and faulty 3D points could potentially cause unstable models. Moreover, incremental SfM might prematurely stop the extension of the model when there are too many conflicting observations in some parts of the scene. Ideally, such mis-registrations are avoided during the incremental extension of the model.

We found, that, if a more gradual set of transition images is provided (such as in a video), the potential for the confusion caused by symmetric structures is significantly reduced. Given the ability of our retrieval system to crawl for images in different directions, we are able to provide a more gradual sequence of images to the SfM system. Hence, in practice we observed a significantly increased robustness to symmetric structures. For example, Arc de Triomphe, Notre Dame, and St. Vitus consistently produced symmetry issues with traditional retrieval approaches, while our system does not suffer from these effects.

9. Experimental Results

In our experiments, we use a generic database of over 7.4 million images downloaded from Flickr through keywords of famous landmarks, cities, countries, and architectural sites. We use single images as seeds for the retrieval (Table 1 and Figure 6) and the subsequent reconstruction.

For the retrieval, the maximum number of verifications per query is set to 5000, the average timings for different types of queries are summarized in Table 2. Combining the different query types, we can retrieve a set of images for a given query image in the order of minutes. The retrieved collection is a concise set of images (less than 0.1% of the entire database) with a relatively small number of irrelevant images, evidenced by the ratio of registered over retrieved images. Please note, that the registered images are all part of the same connected component as the query image. The efficient matching (Section 5) allows us to build the individual models in a matter of a few hours, since we only need to perform matching on the retrieved image pairs which are a fraction of the possible image pairs.

The individual components of the retrieval system have different effects on the obtained results. The sideways crawl helps to increase the extent of the 3D model and to disambiguate the repeated / symmetric structures. When the reconstruction is executed without sideways crawl (using a state-of-the-art pipeline), Arc de Triomphe, St. Vitus, and Notre Dame have symmetry issues and are not reconstructed as complete – for frontal images the top 100 images are still frontal. Zoom-out provides more context and we have observed that it also helps to disambiguate sym-

	Retrieved	Reg.	Pairs	Points
Astronomical Clock	10,443	8,163	572,412	830,238
Bridge of Sighs	2,077	1,018	70,473	117,182
Terracotta Army	2,781	2,099	113,747	167,715
Arc de Triomphe	3,744	2,640	179,346	395,431
Notre Dame	4,978	2,081	164,871	304,339
Sagrada Familia	11,783	7,129	617,362	364,510

Table 1. Details of reconstructed models with the number of retrieved and registered images, the number of verified image pairs, and the number of reconstructed 3D points.

Query type	Time
NN query, no QE	1 sec
NN query, with QE	5 sec
Multi scale-band crawl	2.8 min
Sideways crawl	5.6 sec

Table 2. Average query duration for the retrieval engine.

metric structures. For instance, if executed without zoom-out, the sides of Notre Dames left tower are cross-matched. Zoom-in does not increase the correctness of matching, but it significantly increases the level of detail.

To quantify the amount of detail reconstruction, we determine the spatial resolution of every 3D point as described in Section 7. The surface resolution is mapped to jet color map for visualization, with red referring to the highest and blue to the lowest resolution. Figures 2 and 8 show the resolutions for a variety of scenes. Moreover, Figure 7 is an example of dense reconstruction using multi-view-stereo.

Another experiment shows that the choice of the query image is not critical. Seeding the Arc de Triomphe scene with two different images from opposing sides of the building results in models with 2640 and 2721 images (intersection over union 92%), which are visually near-identical.

To compare our system against full pairwise reconstruction, we injected the Dubrovnik6k dataset [14] with 6,036 images into the 7.4M image database. Starting from a single query image, our pipeline reconstructs 87% (4430 images) in the first 3 retrieval-SfM iterations, w.r.t. to full pairwise reconstruction on the isolated dataset (5102 registered images). Both approaches result in similar visual quality, but faster runtime for our pipeline, even though it operates on 7.4M images (compared to 6K for the pairwise approach).

10. Conclusion

We propose a novel tightly coupled image retrieval and SfM system for the reconstruction of detailed 3D models from unordered Internet photo collections. Our method is able to seed the reconstruction from just a single image. The tight integration of reconstruction and retrieval enables us to retrieve image data suitable for reconstruction that the current state-of-the-art systems in 3D reconstruction from photo collections do not recover. We demonstrate our method on a large variety of scenes from a collection of 7.4 million images downloaded from the Internet.

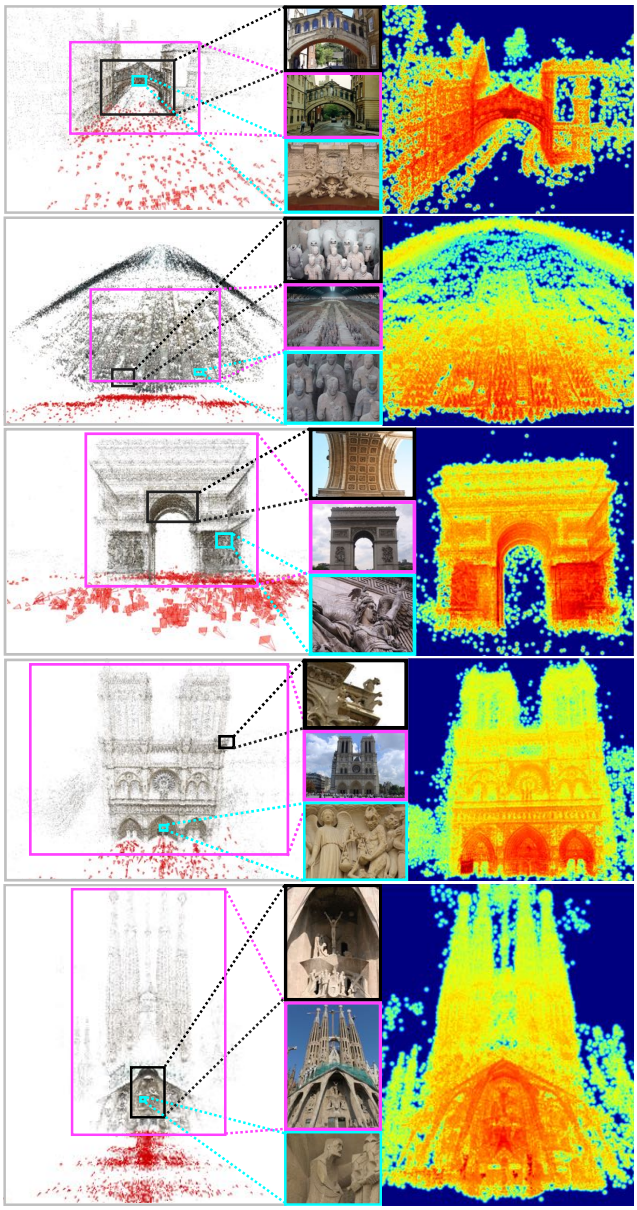

Figure 8. Reconstructions from top to bottom: Bridge of Sighs, UK; Terracotta army, China; Arc de Triomphe, France; Notre Dame, France; Sagrada Familia, Spain. Left: 3D model obtained from our retrieval and reconstruction system. Middle: registered images illustrating the range of views from overview images to images of a specific architectural detail. Right: visualization of the surface resolution from high resolution in red (approximately 1mm surface resolution) to low resolution in blue.

Acknowledgment This material is based upon work supported by the National Science Foundation under Grant No. IIS-1252921, IIS-1349074, IIS-1452851, CNS-1405847, and by the US Army Research, Development and Engineering Command Grant No. W911NF-14-1-0438. F. Radenović and O. Chum were supported by MSMT LL1303 ERC-CZ grant.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building Rome in a Day. *Comm. ACM*, 2011. 1, 2, 4

[2] R. Arandjelović and A. Zisserman. All about VLAD. In *Proc. CVPR*, 2013. 2

[3] O. Chum and J. Matas. Large-scale discovery of spatially related images. *IEEE PAMI*, 32:371–377, 2010. 1

[4] 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. 1, 2, 3

[5] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. SfM with MRFs: Discrete-Continuous Optimization for Large-Scale Structure from Motion. *IEEE PAMI*, 35(12), 2013. 3

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[7] J. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a Cloudless Day. *Proc. ECCV*, 2010. 1, 2, 4

[8] J. Heinly, E. Dunn, and J.-M. Frahm. Correcting for duplicate scene structure in sparse 3d reconstruction. In *Proc. ECCV*, volume 8692, pages 780–795. 2014. 7

[9] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In *Proc. CVPR*, 2015. 1

[10] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. CVPR*, pages 2599–2606, June 2009. 6

[11] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008. 2, 3

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010. 2

[13] 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*, pages 427–440. Springer Berlin Heidelberg, 2008. 2

[14] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Computer Vision ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 791–804. Springer Berlin Heidelberg, 2010. 8

[15] Y. Lou, N. Snavely, and J. Gehrke. MatchMiner: Efficient Spanning Structure Mining in Large Image Collections. *Proc. ECCV*, 2012. 2

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[17] 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*, 65:43–72, 2005. 3

[18] A. Mikulik, O. Chum, and J. Matas. Image retrieval for online browsing in large image collections. In *Proc. SISAP*, 8199, pages 3–15, 2013. 2

[19] A. Mikulik, F. Radenović, O. Chum, and J. Matas. Efficient image detail mining. In *Proc. ACCV*, 2014. 2, 3, 4

[20] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009. 3

[21] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006. 1, 2

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 3

[23] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR*, 2009. 3

[24] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proc. CVPR*, 2010. 2

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007. 2, 3

[26] T. Sattler, C. Sweeney, and M. Pollefeys. On sampling focal length values to solve the absolute pose problem. In *Proc. ECCV*, 2014. 6

[27] J. L. Schönberger, A. C. Berg, and J.-M. Frahm. PAIGE: PAirwise Image Geometry Encoding for Improved Efficiency in Structure-from-Motion. In *Proc. CVPR*, 2015. 2

[28] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470 – 1477, 2003. 2

[29] N. Snavely, S. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. *SIGGRAPH*, 2006. 1, 2

[30] N. Snavely, S. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 2007. 1, 2

[31] H. Stewénius, S. H. Gunderson, and J. Pilet. Size matters: exhaustive geometric verification for image retrieval. In *Proc. ECCV*, pages 674–687. Springer, 2012. 3, 4

[32] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment a modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer Berlin Heidelberg, 2000. 6

[33] K. Wilson and N. Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Proc. ICCV*, 2013. 7

[34] C. Wu. Towards linear-time incremental structure from motion. In *3D Vision-3DV 2013, 2013 International Conference on*, pages 127–134, 2013. 2

[35] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3d structure and motion? *Proc. CVPR*, 0:1–8, 2008. 7

[36] C. Zhang, J. Gao, O. Wang, P. Georgel, R. Yang, J. Davis, J.-M. Frahm, and M. Pollefeys. Personal photograph enhancement using internet photo collections. *Visualization and Computer Graphics, IEEE Transactions on*, 20(2):262–275, 2014. 1