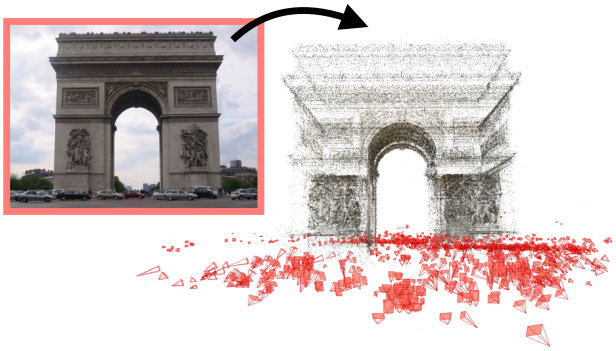
From Single Image Query to Detailed 3D Reconstruction

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



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 [7, 8] to modeling city-scale photo collections of millions of images [4, 5]. Early photo collection reconstruction systems leverage exhaustive matching 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 [2, 3, 6], or image-clustering techniques to identify overlapping images during reconstruction, as demonstrated by the systems of Agarwal et al. [1] and Frahm et al. [4]. 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. [9], it is desirable to have the details present in the reconstruction.

The lack of detail is a result of the employed retrieval approaches [2, 3, 6], 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 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 two components enables us to control the retrieval characteristics based on the current state of the 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 also conveying the structure of the entire scene. An example of the retrieval is shown in Figure 1 and a sample reconstruction is shown in Figure 2.

Figure 1: Image retrieval examples for 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.

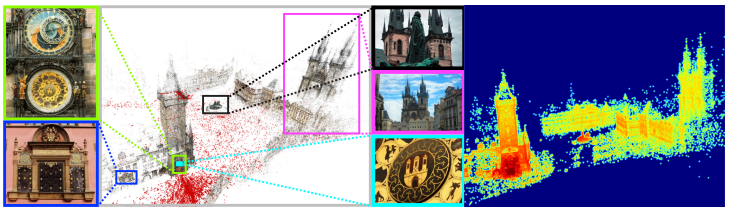


Figure 2: Reconstruction of the Astronomical Clock in Prague, Czech Republic. Left: 3D model obtained from our system. Images illustrating the range of registered views from overview images to images of a specific architectural detail are shown alongside the model. Right: visualization of the surface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

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

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

[3] 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.

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

[5] Jared Heinly, Johannes Lutz Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In *Proc. CVPR*, 2015.

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

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

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

[9] Chenxi Zhang, Jizhou Gao, Oliver Wang, Pierre Georgel, Ruigang Yang, James Davis, J-M Frahm, and Marc Pollefeys. Personal photograph enhancement using internet photo collections. *Visualization and Computer Graphics, IEEE Transactions on*, 20(2):262–275, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.