# Photo tourism: exploring photo collections in 3D

**Richard Hu**

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 835–846, New York, NY, USA, 2006. ACM

The authors present a system for interactively browsing a large unstructured collection of photos of a scene using a 3D interface. The approach is based on an image modeling technique, called structure from motion(SfM), which computes the camera pose for each photograph, along with a 3D model of the underlying scene represented as point cloud. The resulting information enables all the novel browsing options presented in the paper, including image-based rendering, different modes of navigation, and annotation transfer.

With a collection of photos as input, the system attempts to reconstruct the pose of each camera and a sparse geometry of the underlying scene. The system first detects keypoints in each image using the SIFT keypoint detector, which is invariant to image transformation. Next, for each pair of images, the keypoints are matched between pairs of images and then linked together into a track. Once the tracks are computed, the system applies the SfM technique to recover camera parameters and 3D point for each track. This problem can be formulated as a non-linear least squares problem such that the sum of differences between the projection of the 3D point and the corresponding image features observed by the estimated camera pose is minimized. To avoid obtaining bad local minima, the authors take an incremental approach, greedily adding cameras with most feature matches followed by a global refinement of the parameters.

Once the photo collection is registered, users can browse the photos with the photo explorer interface. The scene is rendered by displaying the underlying scene model as 3D points, and photos, represented by their cameras' location and orientation, are displayed as pyramids in the 3D space. The inferred geometric information also allows transition between photos to be more visually compelling by emphasizing the spatial relationship between the photos. A planar morphing technique is applied to provide smooth transition between the source and target photos.

The inferred geometric information also enables different modes of navigation. The relation based navigation allows users to find zoom-out, zoom-in, similar, neighbor, and detail images, based on the relative position and size of the bounding box for the underlying 3D points in the images. Object based navigation is also supported such that users can find photos of a specific object in the scene by dragging a 2D rectangle in a photo or the 3D point cloud. Also, because the camera poses are known for each photo, stabilized slide show is enabled by finding photos with similar camera poses. Since the images are mapped to the common 3D point cloud, annotation transfer is possible at the region level for each image.

There are several contributions of this paper. First, this paper provides the first successful demonstration of SfM technique applied to a large set of images acquired from the Internet. Second, the system presented is robust and combines techniques from image modeling, image rendering, and scene navigation. Such a robust system has high application value and ultimately leads to a streaming web-based service called Photosynth. Most importantly, the paper demonstrates numerous possibilities of rendering and navigating a set of images based on a unifying model.

The paper also has some limitations. First, the system is not able to register a significant portion of photos. For half of the datasets, the system is able to register only about 20 percent of the photos. Part of the reason for such high failure rate is because it is difficult to detect useful features in some challenging scenes, such as textureless scenes, scenes with repeating structures, and cluttered scenes. Second, the running time of the SfM procedure is prohibitive. It takes approximately 2 weeks for a scene with a few thousand photos. A more efficient algorithm or a distributed approach is needed to scale up the system. There are also a few other minor limitations in different aspects of the system, such as the fact that it is currently not able to construct an accurate metric model, constructs only one connected component from the photos, etc. Being the first paper introducing the system, these minor limitations simply suggest more rooms for improvements rather than serious flaws.

One interesting extension to the system is to detect regions that correspond to the underlying model but are blocked by unwanted objects (e.g. people), and replace the unwanted objects with pixels that correspond to the underlying scene. This is similar to image completion; however, since the system maintains a correspondence of the feature points between images and a mapping of each image to the underlying 3D points of the model, it is possible to automatically detect occluded regions and efficiently search for the right replacement.