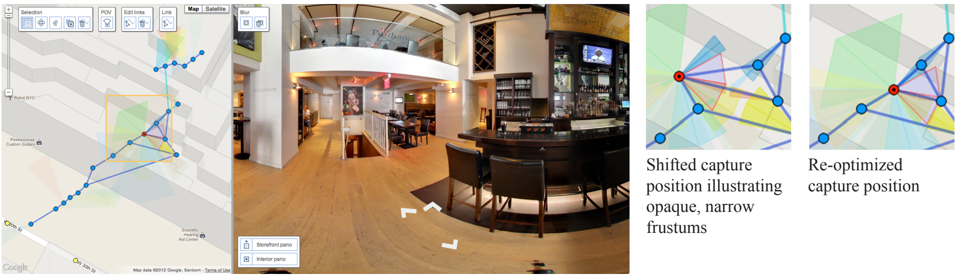
Building Indoor Multi-Panorama Experiences at Scale

Mark Colbert, Jean-Yves Bouguet, Jeff Beis, Spudde Childs,
Daniel Filip and Luc Vincent*
Google, Inc.

Jongwoo Lim†
Hanyang University

Scott Satkin‡
Carnegie Mellon University



Shifted capture position illustrating opaque, narrow frustums

Re-optimized capture position

Figure 1: *Left is a screenshot from the panorama moderation tool. Right is a zoomed selection of the map illustrating how the high-confidence links appear by shifting and re-optimizing a capture point. Note how the less likely links appear lighter and wider.*

1 Introduction

Google Street View has provided millions of users with the ability to visually locate businesses around the world using 360°panoramic imagery. Due to the bulky, custom hardware required to precisely geo-locate the imagery, such as laser scanners and high precision GPS devices, Street View experiences have been limited to large areas that can provide cost-effective collections. This has prevented users from discovering places such as the interiors of small businesses.

Google Business Photos attempts to solve this problem by allowing photographers with commodity DSLR cameras and fisheye lenses to capture panoramas of business interiors and publish the data onto Street View. Photographers collect the data and upload the raw imagery to backend storage. From there, the data is stitched to build 360°panoramas. However, without ancillary capture devices, the images cannot be precisely geo-located. Thus, our method uses a combination of iterative vision-based pose estimation with user-guided optimization enabling photographers to build Street View-like experiences.

2 Iterative Vision-based Pose Estimation

Vision-based pose estimation is a well-studied problem [2004]. In a typical solution, sparse SIFT-like features are extracted from the 360°panoramas and RANSAC is used to determine the best epipole, or direction of motion, between each pair of panoramas. These epipoles provide the fundamental navigation mode in Street View and are represented by the chevrons that link the imagery. To precisely geo-locate panoramas, the images are grouped into clusters in which the poses of panoramas and the feature points are fully reconstructed. The panorama pair containing the largest number of matching sparse features from RANSAC is chosen as a cluster seed. Then, the next best matching panorama to the reconstructed feature points in the cluster is added. This is repeated until all panoramas containing sparse matches above a threshold can be added, after which a new cluster is initialized and loosely connected to an existing cluster using a low-confidence match. Finally, the locations of the panoramas on the map are determined and the connectivity graph of the panoramas is built.

3 User-guided Optimization

While the vision-based pose optimization builds a statistically optimal solution, the underlying result may still be inaccurate or even completely misplaced [1998]. This can arise due to a variety of errors, such as too many false matches from a repeating pattern on a wall or floor, mirrors in the scene, or too few salient feature points. User input is then required to guide the remaining components to build a pose that faithfully reproduces the original geometry.

A map-like display is presented to the user, who is able to manipulate the 3 free parameters for each capture point (latitude, longitude, yaw). In order to aid the user in quickly deciding epipole validity, a visualization mechanism was added indicating the likelihood of a particular estimate to be correct. Unfortunately, it was found that the standard RANSAC score based on inliers did not yield enough useful information. As an alternative, we pick two feature matches, compute the relative planar motion between them, and bin the estimated epipoles into discretized yaw values. These histograms typically fit well to a normal distribution whose standard deviation is inversely correlated to the certainty of an epipole plus a uniform distribution to measure the noise.

This visualization provides a useful guide to assist users in precisely geo-locating each image. However, for simplicity, the user must also be able to adjust subsets of the pose. For instance, if a room was misplaced, the user should roughly move the incorrectly placed data and then re-optimize only the manipulated section. To limit the data size for JavaScript processing, the statistics computed for visualization are used to efficiently optimize subsets of the geometry using gradient descent on the client.

4 Conclusion

The methods and tools presented here have provided photographers the functionality to push thousands of businesses and other locations around the world to Street View. For an average collection of 12 panoramas, the vision-pose estimation typically runs in under 4 minutes and the human-based pose correction takes 7 minutes. The data from this toolchain provides users a novel virtual perspective on many small and medium sized businesses that has previously been unseen, especially for locations that are not easily visible from a driveable street.

References

Hartley, R. I., and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*, second ed. Cambridge University Press, ISBN: 0521540518.

Zhang, Z., and Kanade, T. 1998. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision 27*, 161–195.

*email: {mcolbert,jwlim,satkin,jyb,jeffbeis,djfilip,luc}@google.com
†email: jlim@hanyang.ac.kr
‡email: satkin@cmu.edu

SIGGRAPH 2012, Los Angeles, California, August 5 – 9, 2012.
ISBN 978-1-4503-1435-0/12/0008