

Technical Perspective

Visual Reconstruction

By Carlo Tomasi

THE PAGES THAT FOLLOW boast impressive numbers: 496 processors with a total of 1,984 gigabytes of memory and 62 terabytes of disk digested nearly 460,000 Flickr pictures of Rome, Venice, and Dubrovnik. After 2.5 days, the processors output the detailed three-dimensional geometry and colors of famous landmarks and monuments in these cities. To computer vision researchers, these automatic visual reconstructions—awesome in their detail, magnitude, and fidelity—are a dream come true, never mind the occasional gaps that give the resulting scenes a faintly war-torn look.

It took decades to get here. In 1959, Edgar Hynes Thompson, then Professor of Photogrammetry at University College London, worked out the algebra for the smallest instance of the geometric side of visual reconstruction: If we take two pictures of the same scene from different viewpoints, the image coordinates of five world points are enough to compute where both points and cameras are in space. To this end, we need to know where each of the five points in one image shows up in the other, a task—called point correspondence—that in those days was performed by human operators. In 1934, Thompson himself, a young Captain of the British Royal Engineers, had designed a double microscope with reference grids and moving tables, the Cambridge Stereo-Comparator. An operator peering into the microscope wrote down coordinates of corresponding points in the two photographs. This elaborate apparatus did not just satisfy military exactness: Applied mathematicians soon proved visual reconstruction to be numerically ill-conditioned, thereby requiring extremely accurate data and carefully calibrated cameras to yield reasonable results.

In 1981, to lessen this difficulty, the British theoretical chemist and cognitive scientist Hugh Christopher Longuet-Higgins developed the first of a class of algorithms that use a large number of point pairs to solve an ap-


proximate but convex least-squares version of visual reconstruction. Unfortunately, the resulting estimates are statistically inconsistent, meaning that the output error does not vanish even as the amount of input data grows indefinitely. The modern way out is to compute an initial solution by one of the approximate methods— together with robust estimation techniques to confront omnipresent data outliers—and then refine that solution through numerical, local optimization. This refinement, called bundle adjustment, operates efficiently on large but sparse matrices with techniques that can be traced back to the 1880 work on nested dissection—a divide-and-conquer heuristic based on graph partitioning methods—by the German geodesist Friedrich Robert Helmert. Fortunately, bundle adjustment restores statistical consistency, thus opening the way to automatic computation on common imagery.

The topic of point correspondence—the other grand challenge of visual reconstruction—originated with the advent of digital cameras. People can easily judge if two image details look similar to each other, or if two pictures as a whole depict the same scene. Computers, however, find either task very difficult. In 1999, David Lowe, a profes-

How can visual reconstruction possibly work with images taken by unknown, disparate, uncalibrated cameras under varying weather, lighting, and exposure settings?

sor of computer science at the University of British Columbia, showed how to describe image details well for computing point correspondences. Together with fast data structures for approximate nearest-neighbor search, Lowe's feature descriptors are the workhorse of visual matching in this paper—and of much else in computer vision.

Yet the automated, bulk photogrammetry described here still seems an improbable achievement in the face of the difficulties of geometry and correspondence mentioned earlier. How can visual reconstruction possibly work with images taken by unknown, disparate, uncalibrated cameras under varying weather, lighting, and exposure settings?

In a way, success reveals as much about the input as it does about the computation. In a telltale statistic, only about 20% of the input images were eventually used in the reconstructions, the others being discarded at the many stations along the processing pipeline: Does the scene in this image match that of any other image in the set? Can individual features in this image be placed in accurate correspondence with those of other images? Is the resulting cloud of 3D points consistent with computed camera positions? Are colors similar enough across images to allow for texture mapping? A picture is about as likely to join the final elite as a high school senior is to make it into Duke or Cornell. Success, then, is in part tied to a sort of converse Murphy's Law that seems to hold for massive collections of tourist photographs: If something can go right, it will. If high-quality images are needed, taken under similar weather conditions and exposure settings, and from appropriately separate viewpoints that provide just the right coverage, then there are enough pictures out there that such a set will be found—if you know how. 

Carlo Tomasi (tomasi@cs.duke.edu) is professor and chair of computer science at Duke University.

© 2011 ACM 0001-0782/11/10 \$10.00