Statistical Egomotion Estimation in a Dynamic Environment

Raquel A. Romano

Learning & Vision Group
Artificial Intelligence Laboratory
Massachusetts Institute Of Technology

The Problem: To estimate three-dimensional camera motion from a video sequence of a nonrigid scene viewed by a nonstationary camera.

Motivation: The time-varying imagery captured by a video camera contains a wealth of raw information about the three-dimensional dynamics and structure of the world. Once the camera motion is determined and compensated for, the remaining scene motion may be used to build a global representation of this temporal data. Such tasks as activity detection, 2D mosaic construction, and 3D site modeling become feasible when the multiple sources of image motion are separated.

Difficulty: A camera moving in a dynamic environment produces image sequences that do not obey the geometric constraints typically used for analyzing camera motion in stationary scenes. These conditions also confuse activity detection and motion filters that assume a stationary camera. Consequently, many existing methods are inadequate for interpreting motion in image sequences with both moving objects and a moving camera.

Previous Work: Egomotion computation, the computation of 3D camera motion from a sequence of images, has seen decades of study. Numerous solutions have been proposed for computing the path of a moving camera viewing a static scene [6]. Many of these solutions suffer from inaccuracy in the presence of noisy image measurements, and more importantly, they are inapplicable to imagery captured from a dynamic world. Other attempts have bypassed the 3D motion estimation and have directly addressed the problem of robust 2D motion segmentation in the presence of noisy measurements of multiple moving objects [1, 2].

More recently, researchers have tackled the problem of egomotion computation given an image sequence of a nonstatic world [5]. The problem grows increasingly difficult when significant portions of an image violate the rigidity assumption in a structured way and thus do not satisfy Gaussian noise assumptions. This work addresses precisely the problem of how to estimate the motion of a camera undergoing arbitrary translations and rotations while viewing a scene with multiple independently moving objects.

Approach: We propose a statistical approach which uses a mixture model to describe how multiple sources of 3D object motion, including camera motion, simultaneously induce an observed 2D image motion in textured regions of an image sequence. We have implemented a tracking method that measures 2D image motion by computing the Hausdorff distance between variable size windows in edge maps of an image pair [4]. The Hausdorff tracker returns rigid 2D displacements of possibly nonrigid moving objects as well as the perceived 2D displacement of the static background. Capturing the image motion induced by an object’s global rigid 3D motion rather than its local nonrigid deformations is essential for the accurate estimation of 3D motion.

We use the instantaneous 3D motion model to relate the rigid 3D translation, rotation, and depth of an object at a point in space to the corresponding 2D motion of the imaged point. Each component of the mixture model corresponds to a 3D motion model for a distinct object moving independently in the world. Since the camera moves relative to the stationary regions of the scene, estimation of the perceived 3D motion of the background will yield the parameters of the camera motion. The camera motion is thus represented as just one component of the mixture model. The image motion observed for a moving object will be modeled as a mixture of camera-induced motion and object-induced motion.

Estimating the maximum likelihood parameters of a mixture model requires finding both the parameters of each individual model and the mixing parameters for each data point. In our case, the model parameters are the expected 3D translation and rotation, and the mixing parameters describe the extent to which each 3D motion source is responsible for generating a 2D motion measurement. Iterative algorithms such as expectation maximization allow the simultaneous estimation of both sets of parameters [3].
**Impact:** With the proliferation of digital image and video capabilities in personal computing, there is increasing demand for software that interprets not only still images, but image sequences taken under a variety of conditions without prior knowledge about either the camera motion or the scene dynamics. Potential applications include the creation of virtual environments from video streams, the stabilization of enhanced reality displays in dynamic environments, and activity detection from unmanned ground, air, and space vehicles.

**Future work:** Among the remaining issues under investigation are how to automatically choose and update the number of objects modeled by the mixture and how to fold depth estimates into the motion computation.

**Research Support:** The author is supported by an AT&T Laboratories Graduate Fellowship. This research is also part of the “Forest of Sensors” project funded by DARPA under contract number N00014-97-0363 administered by ONR.

**References:**


