Geometric Analysis of Motion in Video

Raquel Romano

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

Problem: This work addresses the problem of finding geometric representations for the multiple motions present across many frames of a video sequence. We examine several geometric models that capitalize on the continuity and redundancy of points tracked across frames in a sequence.

Motivation: The ability to automatically understand multiple motions in video streams when both the camera and the foreground objects are moving is relevant to a variety of video manipulation applications. Multi-frame motion models make possible a wealth of applications that might involve warping, stitching, editing, modifying, tracking, synthesizing, indexing, or querying video.

Previous Work: Historically, motion analysis methods have focused on fitting geometric models to pairs or triplets of images in an image sequence. Computational simplicity and mathematical completeness are good reasons for analyzing only 2 or 3 images at a time, but modeling the geometry of a video stream as a chain of image pairs has the drawback that the sets of matching points from non-consecutive frames are not explicitly constrained to satisfy a single geometric model. They are only implicitly related through the chain of pairwise models. Such pairwise models make sense when analyzing a collection of images taken from widely spaced and oriented camera poses because only some views overlap enough to provide enough point matches for reliable estimates. In video, however, an abundance of scene points are visible in every frame, so n-tuples of matching image points should fit an explicit multiple-view model of rigid motion. The problem is how to find a geometric model for a multiple-frame sequence that is both tractable to compute and has the right numbers of degrees of freedom for the underlying camera motion. Several approaches to fitting general, geometric models to many frames may be found in [2, 5, 7].

When the camera motions are not rich enough for general motion models, one recourse is to make simplifying assumptions about the motions. Because video streams are taken from a continuously moving camera, simplified motion models are often good local approximations to a globally more general motion, and can be modeled using differential or instantaneous models in which 3D motions and their 2D projections are represented as time-varying functions [1, 4]. By approximating these functions with low-order Taylor polynomials, we may obtain low degree of freedom multiple-view models that do not require 3D analysis.

Approach: Our approach to video motion analysis is to constrain the geometry of points tracked across many views without explicitly computing 3D scene structure and camera motion. Our most general model is a collection of interdependent fundamental matrices whose building block is a novel parameterization for the "third" fundamental matrix in a triplet of images. Suppose we have three non-collinear cameras viewing a scene, and the epipolar geometry between two pairs of the triplet has been reliably estimated. Then the fundamental matrix relating the third pair of views is partially determined by the first two matrices. Ordinarily, the fundamental matrix has 7 degrees of freedom, 4 for the epipoles and 3 for the epipolar transformation, the collineation mapping the epipolar lines in one view to those in the other. However, the two known epipolar geometries constrain the epipoles of the unknown view pair to lie on a straight line, and hence only 1 parameter of each epipole is free. In addition, the known fundamental matrices completely determine one matching line of the unknown view pair's epipolar transformation, so the collineation has only 2 degrees of freedom.

We have developed a technique for expressing the 4 parameters of the third fundamental matrix in terms of the two known epipolar geometries. These parameters are then computed from point matches using a nonlinear minimization, and are used to construct the fundamental matrix. This method is embedded in several algorithms that explore how to choose the order of fundamental matrix estimation given many views, and how a novel frame may be integrated into the model by constraining it to agree with a subset of frames with previously estimated relative epipolar geometries.

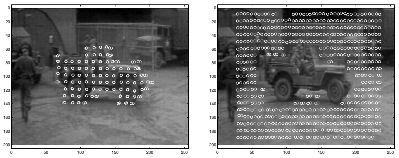
Such a collection of dependent epipolar geometries forces points viewed in many images to adhere to a consistent rigid 3D motion.

While dependent, parameterized fundamental matrices form a single, general geometric model of the relative motion between the camera and the background, we have found that simplified motion models over multiple frames can be good indicators of distinct 3D motions. Figure 1 shows how by making first-order approximations to instantaneous motion models and simplifying assumptions about the complexity of motion over many frames, we can classify interpolated optical flow tracks into regions of distinct motions.

Difficulty: The question of how many parameters accurately model the motions in a video sequence is of great importance. If a fully general model is chosen but the motion is linear or the optical center is stationary, the model is under-constrained. If a simplified 3D motion model is fit, but the camera motion is very rich, the model will not capture the full complexity of the motion.

Future Work: A difficult and pervasive problem is how to decide how general or specific the motion model for a particular video sequence should be. For any given model, it is easy to find a video sequence for which the motions or scene structure violate any simplifying assumptions, or for which the underlying motions are too simple to be properly constrained by a general model. Model selection [6] is one approach to finding the most general model for which the motions in a sequence are not degenerate. We would like to find methods that neither require a brute-force search of all potential models, nor an a priori specification of the types of motions present in the sequence.

(a)



(b)

Figure 1: (a) Every other frame of a ten-frame input sequence. Both the camera and the foreground jeep are moving and points are tracked over many frames using interpolated optical flow. (b) Segmentation of the foreground (left) and the background (right) regions. White circles indicate the locations where optical flow is measured.

Research Support: This research is supported by DARPA under ONR grant N00014-97-0363 and by the AT&T Laboratories Fellowship Program.

References:

[1] Kalle Astrom and Anders Heyden. Continuous Time Matching Constraints for Image Streams. *IJCV*, 1998.

[2] S. Avidan and A. Shashua. Threading Fundamental Matrices. *ECCV*, 1998.

[4] Michal Irani. Multi-Frame Optical Flow Estimation Using Subspace Constraints. *ICCV*, 1999.

[5] David Nister. Reconstruction from Uncalibrated Sequences with a Hierarchy of Trifocal Tensors. *CVPR*, 2000.

[6] Phil Torr, Andrew W. Fitzgibbon and Andrew Zisserman. Maintaining Multiple Motion Model Hypotheses through Many Views to Recover Matching and Structure. *ICCV*, 1998.

[7] Bill Triggs. Matching Constraints and the Joint Image. *ICCV*, 1995.