Direct methods for estimation of structure and motion from three views

G. P. Stein

Artificial Intelligence Laboratory
MIT
Cambridge, MA 02139
gideon@ai.mit.edu

A. Shashua

Institute of Computer Science
Hebrew University of Jerusalem
Jerusalem 91904, Israel
http://www.cs.huji.ac.il/~ shashua/

Abstract

We describe a new 'direct method' for estimating structure and motion from image intensities of multiple views. We extend the direct methods of [5] to three views. Adding the third view enables us to solve for motion, and compute a dense depth map of the scene, directly from image spatio-temporal derivatives in a linear manner without first having to find point correspondences or compute optical flow.

We describe the advantages and limitations of this method and show experiments with real images.

1 Introduction

We present a new method for computing motion and dense structure from three views. This method can be viewed as an extension of the 'direct methods' of Horn and Weldon [5] from two views (one motion) to three views (two motions). These methods are dubbed 'direct methods' because they do not require prior computation of optical flow. We assume small image motions on the order of a few pixels.

Applying the constant brightness constraint [4] to the trilinear tensor of Shashua and Werman [9, 12] results in an equation relating camera motion and calibration parameters to the image gradients (first order only). We get one equation for each point in the image with a fixed number of parameters. This results in a highly over-constrained set of equations.

This method has advantages over both optical flow methods [6][7] and feature based methods [12]. We combine the information from all the points in the image and thus we avoid the aperture problem which makes computation of optical flow difficult. We do not explicitly define feature points. Points with small gradients simply contribute less to the least squares estimation. Information from all points that have gradients is used.

[0] G.S would like to acknowledge DARPA contracts N00014-94-01-0994 and 95009-5381. A.S. would like to acknowledge US-IS BSF contract 94-00120 and the European ACTS project AC074.

These advantages are highlighted in a scene with a set of vertical bars in front of a set of horizontal bars and behind everything a uniform background. Optical flow methods fail because of the straight bars and the aperture problem. Discrete methods fail because the intersections of the lines in the image, which are detected as 'features' do not correspond to real features in space. Many natural scenes such as tree branches or man made objects, window frames, lamp posts and fences give rise to these problems.

Starting with the general uncalibrated model we proceed through a hierarchy of reduced models first by assuming calibrated cameras and then by assuming the Longuett-Higgins and Prazdny small motion model [6]. We then show how to solve the simplified model for the motion parameters.

1.1 Previous Work

The 'direct methods' were pioneered by Horn and Weldon in [5]. A single image pair results in N equations in $N+6$ unknowns, where N is the number of points in the image, so a constraint is needed. Negahdaripour and Horn [8] present a closed form solution assuming a planar or quadratic surface.

This work is based on the work of Shashua and Hanna [11]. Here we describe the results of implementing these ideas in practice. During the course of implementation various subtleties and limitations were discovered.

2 Mathematical Background

2.1 Notations

Let $\boldsymbol{x}$ be a point in 3D space and its projection in a pair of images be p and p'. Then $p = [I; 0]\boldsymbol{x}$ and $p' \cong A\boldsymbol{x}$, the left 3×3 minor of A stands for a 2D projective transformation of the chosen plane at infinity and the fourth column of A stands for the epipole (the projection of the center of camera 0 on the image plane of camera 1). In particular, in a calibrated setting the 2D projective transformation is the rotational component of camera motion and the epipole is the translational component of camera