Geometry and Statistics of Visual Space-Time

Cornelia Fermüller, Patrick Baker and Yiannis Aloimonos*

Abstract

Although the fundamental ideas underlying research efforts in the field of computer vision have not radically changed in the past two decades, there has been a transformation in the way work in this field is conducted. This is primarily due to the emergence of a number of tools, of both a practical and a theoretical nature. One such tool, celebrated throughout the nineties, is the geometry of visual space-time. It is known under a variety of headings, such as multiple view geometry, structure from motion, and model building. It is a mathematical theory relating multiple views *(images) of a scene taken at different viewpoints to* three-dimensional models *of the (possibly dynamic) scene. This mathematical theory gave rise to algorithms that take as input images (or video) and provide as output a model of the scene. Such algorithms are one of the biggest successes of the field and they have many applications in other disciplines, such as graphics. One of the difficulties, however, is that the current tools cannot yet be fully automated, and they do not provide very accurate results. More research is required for automation and high precision. During the past few years we have investigated a number of basic questions underlying the structure from motion problem. Our investigations resulted in a small number of principles that characterize the problem. These principles, which give rise to automatic procedures and point to new avenues for studying the next level of the structure from motion problem, are the subject of this paper.*

1 Introduction: The problem

We are given a number of images of a scene taken at different viewpoints and the goal is to create 3D models of the scene in view. What is a geometric model of image formation? To make an image, we first pick a point in space and consider all the light rays passing through this point. Then we cut these rays with a surface. For the standard pinhole camera, this surface is a plane and images are formed by central projection on a plane (Fig. 1a). The focal length is f and the coordinate system $OXYZ$ is attached to the camera, with Z being the optical axis, perpendicular to the image plane. Image points are represented as vectors $\mathbf{r} = [x, y, f]^{\mathrm{T}}$, where x and y are the image coordinates of the point in the coordinate system oxy, with $ox \| OX$, $oy \| OY$ and O the intersection of the axis OZ with the image plane, and f is the focal length in pixels. A scene point $\mathbf{R}$ is projected onto the image point

$$\mathbf{r} = f \frac{\mathbf{R}}{\mathbf{R} \cdot \hat{\mathbf{z}}} \tag{1}$$

where $\hat{\mathbf{z}}$ is the unit vector in the direction of the Z axis.

If we cut the rays with a sphere, we obtain a spherical eye with a full field of view (Fig. 1b). In the case of video, the camera is moved to different locations while acquiring new images. Thus, video acquired by a moving camera amounts to a collection of images of a scene, i.e., projections onto an imaging surface, acquired from different viewpoints. Figuring out a model for the scene and the movement in the scene becomes a problem of relating the different projections (images) to each other.

In general, when a scene is viewed from two positions, there are two concepts of interest:

(a) The 3D transformation relating the two viewpoints. This is a rigid motion transformation, consisting of a translation and a rotation (six degrees of freedom). When the viewpoints are close together, this transformation is modeled by the 3D motion of the eye (or camera).

(b) The 2D transformation relating the pixels in the two images, i.e., a transformation that given a point in the first image maps it onto its corresponding one in the second image (that is, these two points are the projections of the same scene point). When the viewpoints are close together, this transformation amounts to a vector field denoting the velocity of each pixel, called an image motion field.

Perfect knowledge of both transformations described above leads to perfect knowledge of models of space and action. Regarding models of space, this is easy to understand. Knowing exactly how the two viewpoints and the

*The authors are with the Computer Vision Laboratory and Center for Automation Research at the University of Maryland, College Park, MD 20742-3275, USA. E-mail: fer@cfar.umd.edu