Qualitative Estimations of Range and Motion Using Spatio-temporal Textural Images

Zhigang Zhu, Guangyou Xu and Dingji Shi
Department of Computer Science, Tsinghua University
Beijing 100084, P. R. China

Abstract

In this paper we model the problem of structure from motion as the range estimation with known motion. First, we approximate the motion within a reasonable time interval as a 3D translation and thus some image transformations are applied to convert an arbitrary motion to a 1D translation . Second we have avoided the feature extraction and correspondence problems by analyzing the epipolar plane image in the Fourier domain. Experimental results with real scene images on campus have shown the efficiency and robustness of the approach.

1: Introduction

For the task of visual navigation of road following, global information is needed to judge if the robot is following the correct path and has reached the appropriate position. One of the most important information is the depth from the robot (observer) to the objects along the route, which can be obtained through a moving observer.

Most of motion analysis algorithms are based on only two images of a sequence . An alternative approach considers image sequence analysis in a multi-dimensional space , or spatio-temporal images [2,3,4]. Generally speaking, there are some significant advantages of this approach. First, motion can be analyzed in the frequency domain. Second more powerful constraints implied by the dynamics of motion can be included. Third, using more than two images , more robust and accurate determination of motion and/or range can be expected. In visual navigation , we have an active observer(robot) and relative static environments. So we can model the problem as range estimation with a priori known motion. In this paper we treat range as 2D oriented texture and infer range information qualitatively using Fourier energy spectrum of large Fourier windows.

2: Motion and range in ST space

The perspective projection p(x,y) of an object point P(X,Y,Z) under a pin hole camera model is

$$(x , y) = (f X / Z, f Y/Z)$$

where f is the focal length. Since the observer(camera) moves, the image coordinates, x and y , as functions of time t, are given by

$$(x(t), y(t)) = (f X(t) / Z(t) , f Y(t)/Z(t)) \quad (1)$$

If a sequence of images is taken continuously and is piled up sequentially with time , we can construct a 3D spatio-temporal image (3D ST image cube). An image point p(x,y,t) of the object point P(X,Y,Z,t) draws a 3D locus inside the 3D cube, so motion and /or range can be regarded as 3D orientation in xyt space.

In our configuration, a mobile robot moves smoothly with known motion on a relative flat ground plane along a route and most objects in the scene are static. The optical axis of the camera is parallel to the ground plane and perpendicular to the motion direction. For simplicity , suppose the robot moves with constant speed -V (V < 0) , then equation (1) becomes

$$(x(t), y(t)) = (f (X + V t) / Z , f Y / Z) \quad (2)$$

where X,Y,Z are the 3D coordinates in the time t = 0. In such a configuration, panoramic view images (PVIs)[1] are the 2D cross-section images perpendicular to x axis and epipolar plane images (EPIs)[2] are the 2D cross-section images parallel to x axis in 3D ST cube. We have used these 2D ST images in the road scene understanding and mobile robot localization[5]. In this paper we focused on the range estimation from xy space(EPI).

Under the condition of equation(2),the depth of the edge point can be calculated as

$$D = Z = f V / (dx/dt) = f V / v \quad (3)$$

where dx/dt is the slope of the straight locus and v is the 1D optical flow. In this way the optical flow in time t is the slope of the locus in that time and range can be estimated by the orientation of the locus in 2D xt space.

3: Motion and range in Fourier domain

3.1: Translation with constant velocity

First we consider the case of an image sequence in which all the objects are in the same depth(range) and the mobile robot is moving with constant velocity -V. From equations (2) and (3) the sequence g(x,t) can be expressed by

1051-4651/94 $04.00 © 1994 IEEE

$$g(x,t) = g(x + vt) \qquad (4)$$

where $g(\cdot)$ is the grey value and v is the 1D image velocity. The Fourier transform of this sequence is

$$G(\xi, \omega) = G(\xi)\,\delta(v\xi - \omega) \qquad (5)$$

The equation states that objects of same range moving with velocity V occupies only a line in the $\xi\ \omega$ space. The slope of the line is -1/v which means that the line in $\xi\omega$ space is perpendicular to the orientation of motion in xt space.

Suppose we have taken image at temporal distance Δt. From the sample theory, if the sampling interval in t direction is smaller than one-half of the reciprocal of temporal bandwidth (product of image velocity |v| and spatial bandwidth $\xi 0$), namely

$$\Delta t < 1/(2\,|v|\,\xi 0) \qquad (6)$$

then the motion and/or range can be recovered unambiguously, and the correspondence problem can be avoided.

3.2: Translation with constant acceleration

If the mobile robot moves with an initial velocity V and acceleration A , then equation (4) becomes

$$g(x,t) = g(x + v\,t + 1/2\ a\ t^2). \qquad (7)$$

where v = fV/Z and a = fA/Z are the image velocity and acceleration respectively. Its Fourier transform is

$$G(\xi, \omega) = G(\xi)\,H(\omega - v\,\xi) \qquad (8)$$

where $H(\omega)$ is the Fourier transform of $h(t) = \exp(j\pi a\xi\ t^2)$, ie.

$$H(\omega) = \begin{cases} \delta(\omega) & \text{if } a = 0 \\ 1/\sqrt{(a\,\xi)}\,\exp(j\pi/4)\,\exp(-j\,\pi\,\omega^2 / a\,\xi) & \text{if } a \neq 0. \end{cases} \qquad (9)$$

It indicates that $G(\xi, \omega)$ is much more complex when $a \neq 0$. Therefore, instead of using $G(\xi,\omega)$ of equation (8) directly, we transform the ST image expressed by equation (7) to the form of equation(4) by a temporal resampling process.

4: Image rectification and resampling

4.1: Piecewise translation plane model

In practical application, the robot could not move in a pure translational motion with a constant speed. For an arbitrary motion , the discrete motion sequence can be expressed by

$$S = \{ (Ri,Ti),\ i = 0,1,2,\ldots \} \qquad (10)$$

where (Ri,Ti) is the motion parameters from frame i to i+1 and is known at a priori. In our smoothing egomotion model, we approximate the motion in a reasonable time period $i \in [0, N]$ as a piecewise 3D translation. The optical center of the observer forms the 3D motion trajectory and the Y axis forms the "motion surface". Within the time period N, this surface is fitted as a plane in 3D space, namely translation plane P, and the 3D motion trajectory is approximated as a pure 3D translation vector T inside plane P.

4.2: Gaze transformation

Within the time period N, If the optical axis Z is not perpendicular to the translation plane P and/or X axis is not parellel to vector T, the software gaze transformation is used to reproject the sensor image to the plane parallel to P. In fact, gaze transformation is a image perspective rectification through a software camera rotation $R = (r_{ij})_{3x3}$, which is irrelevant to the depths of 3D points. The relation between new and old image coordinates is

$$U' = (f R_1 \cdot U / R_3 \cdot U,\ f R_2 \cdot U / R_3 \cdot U,\ f) \qquad (11)$$

where $U' = (x',y',f)$, $U = (x,y,f)$ and $R_i = (r_{i1},r_{i2},r_{i3})$, i=1,2,3. After the image rectification, the motion of the objects is approximately a 1D translation parellel to X axis , with small translation components in Y and Z directions ignored. In this way ranges can be determined in 2D ST images (EPIs).

4.3: Temporal resampling

After gaze transformation, suppose the mobile robot moves along X axis with velocity -V(t) which is the function of time t, we have

$$g(x,t) = g(x + f\ (V(t)/Z)\ t). \qquad (12)$$

If V(t) is not a constant, uniform temporal sampling interval (eg. 1/30 s) of the image sequence will result in an EPI with curved loci, which are not intuitive in the Fourier domain. So we resample the resulted EPI along t axis with a velocity-dependent time intervals. In the continuous case, substituting

$$t' = V(t) / V_0\ t \qquad (13)$$

in equation(12), we have

$$g(x,t') = g(x + f\,(V_0 / Z)\,t') = g(x + v_0\,t') \qquad (14)$$

where $v_0 = f\,V_0 / Z$. It can be seen that equation(14) has the same form as equation(4) which means that the temporal resampled ST image is equivalent to the ST image of translation with the constant velocity V_0. It should be noticed that the temporal resampling process is also independent to the range of the object.

5: Qualitative range from temporal texture

5.1: Algorithm

For the textured and complex natural scene with tree and grass, edges of the original images and hence the loci of the EPI are not very strong and contrast varies significantly. However the temporal texture formed in the EPI of natural scene is characterized by very strong orientation features, which represent the ranges of the objects in the scenes. So instead of tracking the loci sequentially, we detect the main texture orientation of the EPI qualitatively by means of Fourier energy spectrum in a sequence of overlapping Fourier windows(OFWs) along the time axis of EPI. The energy spectrum of an OFW of size MxM can be expressed as

$$P(\xi, \omega) = |G(\xi, \omega)|^2 \qquad (15)$$

where $G(\xi, \omega)$ is the Fourier transform of xt image in OFW and $\xi, \omega = 0,1,...,M-1$.

Although the video rate image sequences generally do not meet the temporal sampling condition , objects in natural scene, such as trees and grass, contain spatial frequency over a wide range. So we can choose the band of the epipolar wave according to the temporal sampling intervals and the velocity and depth of the object, namely

$$\xi m < 1/2|v|\Delta t = D / 2 f V \Delta t \qquad (16)$$

Using polar coordinates (r, ϕ) instead of (ξ, ω), we obtain $P(r, \phi)$, where $r_{i+1} - r_i = M/2n$, $\phi_{j+1} - \phi_j = \pi/n$, $i,j = 0,1,...,n-1$, and n is the sample number for r and ϕ. By computing the histogram with variable ϕ , we have

$$P_d(\phi) = \sum_{r=r_l}^{r_h} P(r, \phi) \ , \ \phi = \phi_1, \phi_2, ..., \phi_{n-1} \qquad (17)$$

where bandpass filter is achieved through the selection of r_l and r_h ($0 < r_l < r_h \leq r_{n-1}$). For the single oriented texture(ie. single range within the Fourier window), $P_d(\phi)$ has a single peak , hence the direction of the temporal texture can be expressed as

$$\theta = \{ \phi : P_d(\phi) = \max \} \qquad (18)$$

where we only need $0 < \phi < \pi/2$ for the case of EPI texture . The relative range D can be computed reliably using equations (3) :

$$D = f V / v = - f V (\Delta t / \Delta d) \tan(\theta) \qquad (19)$$

where Δt(s/frame) is the temporal sampling interval and Δd(mm/pixel) is the interpixel distance.

5.2: Experiments

The above procedures have been implemented on the SUN386i workstation with DT color image processing boards. The PVI and EPI were constructed continuously while the mobile robot was moving along the route(Fig. 1). Fig. 2 shows the Fourier energy spectrum of EPI in successive non-overlapping blocks of size 64*64 and the relative ranges of the points in the scene. In this experiment, the band of the bandpass filter is $r \in [10,32]$. Theoretical analysis and simulation experiments show that the angle resolution for ϕ is about 2 degrees. It can be seen the robust range estimation were obtained using Fourier approach.

6: Discussions

In order to obtain more accurate and robust estimation of ranges, we choose a large Fourier window, eg. MxM=64x64, so the Fourier analysis is a time-consuming procedure. Consider the overlapping the successive Fourier windows, we have implemented a very fast Fourier transform(VFFT) algorithm. The time complex for the 2D VFFT is O(MxM), which is significantly smaller than O(MxMxlog$_2$M) of 2D FFT.

Currently we have applied Walsh transform (DWT) instead of Fourier transform(DFT) to extract the range information from temporal texture images. Even if DFT is generally considered to be superior to DWT in energy compaction, there are two advantages of DWT when it is used in this specific application --- much lower computing complex and double higher orientation resolution . (The experiment results are not shown here).

References

[1].Zheng,J.Y., Barth, M. and Tsuji,S., Panoramic representation of scenes for route understanding. Proc. IEEE 10th ICPR,1990: pp 161-167.

[2]. Baker,H.H., and Bolles,R.C., Generalizing epipolar plane image analysis on the spatio-temporal surface. Proc. IEEE CVPR, 1988 pp2-9.

[3]. Allmen, M. and Dyer, C.R., Computing spatiotemporal surface flow, Proc IEEE 3rd ICCV, 1990: p47-50.

[4]. Heeger,D.J., Optical flow from spatio-temporal filters, ICCV-87:p181-190.

[5].Zhu,Z.G., Xu, G.Y., Peng, J. and Shi, D.J., " Route understanding using spatio-temporal images viewed through a cross window," Proc. ACCV, Osaka, Japan,1993: pp 35-38 .

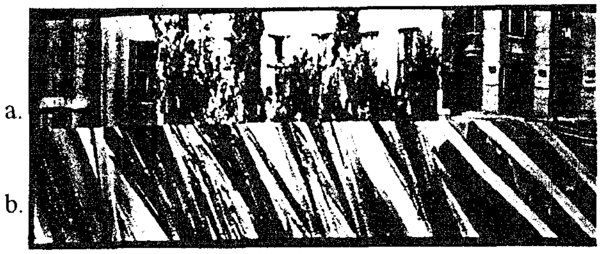

*Fig. 1. (a). PVI and (b). EPI of 128*512*

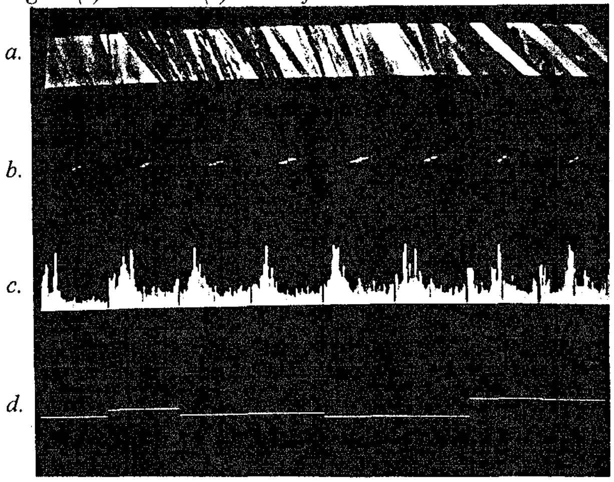

Fig.2. Range from temporal texture (a). g(x,t) (b). P (ξ, ω) (c).P_d (φ) (d). 6(t)