Visual Control of a Robot Head

Han Wang Stan Z Li E K Teoh
School of Electrical & Electronic Engineering
Nanyang Technological University, Singapore 2263
e-mail: hw@ntuix.ntu.ac.sg phone: (+65)799-1253 fax: (+65)791-2687

ABSTRACT

We report in this article a real time algorithm TIM that can track and estimate hand gesture, providing reliably 5 degrees of freedom from a single camera for controlling of a robot head. The advantages of this algorithm are: it does not require camera calibration and the syetem will operate under any configuration; secondly it converges fast and reliably which means special hardware, such as a powerful parallel platform, may be dispensed with. Kalman filter is used to track feature points on the finger tips.

1 Introduction

Structure from motion attempts to capture and quantify the relative motion between the viewer and the scene[6, 1, 11, 10]. Commonly adopted approaches need establishing correspondences in tempral space and minimising a certain energy term in the presence of noises incurred by quantisation and observation[3]. These approaches require accurate camera calibration and feature extration. Recent studies show that *time − to − contact* is an important factor in navigation whereby the precise measurement of object depth is not known or not essential. It is known that the optimisation scheme will not always converge to the global minimal due to the presence of many local minima.

Cipolla *et al*[5] retrieved the motion of hand by making use of the motion parallax where he assumed weak perspective and affine transformation. The method is based on the transformation of four marking points on the hand where these points must not be coplanar.

The task of recovering 3D motion of the hand can be simplified with the assumption of known hand model. The three fingers of the right hand have been parameterised and shown in Figure 1. Absolute measures are not required since they will be scaled against the pixel units. We will show that the correspondences from these three finger tips are nonlinear and the motion can be recovered with a simple iteration method.

2 The computation of hand pose under orthographic projection

The motion of the hand is recovered from two consecutive views. In this approach, we assume orthographic projection which is a special case of perspective projection when the focal length is placed at infinity. In practice, the object of concern must be sufficiently small in comparison with the depth. Then, the X and Y axes of the world coordinates is effectively the same as the image coordinates.

Without loss of generality, let $H = \{T, I, M\}$ denote the model of the right hand represented by the three finger tips of Thumb, Index finger and Middle finger (hence the name TIM), where

$$\begin{aligned} T &= (0, -Y_T, 0)^\top \\ I &= (0, 0, -Z_I)^\top \\ M &= (X_M, Y_M, -Z_M)^\top \end{aligned}$$

are relative measurements of the hand parameters obtainable at the initial stage of tracking

to engage the hand model (see Figure 1).

Let x denote a point of the initial hand position (eg. this point can be one of the three finger tips) in the 3D space, and let x′ denote the same point after the motion (R, T). We first rotate the hand by R, then translate it to the origin of the image plane by T, hence the motion can be described as,

$$\mathbf{x}' = R\mathbf{x} + T \tag{1}$$

where the rotation matrix R can be denoted by rotations of $(\theta_x, \theta_y, \theta_z)$ about the X,Y and Z axes,

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & Cx & -Sx \\ 0 & Sx & Cx \end{pmatrix},$$

$$R_y = \begin{pmatrix} Cy & 0 & Sy \\ 0 & 1 & 0 \\ -Sy & 0 & Cy \end{pmatrix},$$

$$R_z = \begin{pmatrix} Cz & -Sz & 0 \\ Sz & Cz & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and we use C_x to denote $\cos(\theta_x)$, S_y to denote $\sin(\theta_y)$. For $R = R_x R_y R_z$, we obtain the following[8],

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix}$$

$$= \begin{pmatrix} C_y C_z & -C_y S_z & S_y \\ C_z S_x S_y + C_x S_z & C_x C_z - S_x S_y S_z & -C_y S_x \\ -C_x C_z S_y + S_x S_z & C_z S_x + C_x S_y S_z & C_x C_y \end{pmatrix} \tag{2}$$

The translational components T can be denoted as (t_x, t_y, t_z). We show the correspondence from the index finger using equation 1,

$$\begin{pmatrix} x_i' \\ y_i' \\ z_i' \end{pmatrix} = -\begin{pmatrix} R_{13} \\ R_{23} \\ R_{33} \end{pmatrix} Z_i + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}. \tag{3}$$

The distance from the object origin to the image plane denoted by t_z is unknown from a single camera projection. We drop the corespondence from the depth dimension, resulting in two equations. Correspondences from other two fingers provide additional four equations. Rigid motion is described by 6 independent variables of R and T. Due to the orhographic limitation[7], we can only recover 5 of them with the missing component of translation along the depth dimension. The correspondence equations from the three finger tips are given below,

T:

$$\begin{pmatrix} x_t' \\ y_t' \end{pmatrix} = -\begin{pmatrix} R_{12} \\ R_{22} \end{pmatrix} Y_t + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

I:

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = -\begin{pmatrix} R_{13} \\ R_{23} \end{pmatrix} Z_i + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

M:

$$\begin{pmatrix} x_m' \\ y_m' \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \begin{pmatrix} X_m \\ Y_m \end{pmatrix} - \begin{pmatrix} R_{13} \\ R_{23} \end{pmatrix} Z_m + \begin{pmatrix} t_x \\ t_y \end{pmatrix}. \tag{4}$$

The corespondence equations are nonlinear, therefore a numerical approach is necessary. Dropping the equation of y_m' (only five are required), rearrange the correspondence equations, we obtain,

$$\begin{cases} \theta_x = \arcsin(\dfrac{y_i' - t_y}{Z_i \cos\theta_y}) \\ \theta_y = \arcsin(\dfrac{t_x - x_i'}{Z_i}) \\ \theta_z = \arcsin(\dfrac{x_t' - t_x}{Y_t \cos\theta_y}) \\ t_x = x_m' + Z_m \sin\theta_y \\ \qquad - \cos\theta_y (X_m \cos\theta_z - Y_m \sin\theta_z) \\ t_y = y_t' - y_t(\sin\theta_x \sin\theta_y \sin\theta_z \\ \qquad - \cos\theta_x \cos\theta_z) \end{cases} \tag{5}$$

where a simple iteration method will guarante a quick and unique solution since the

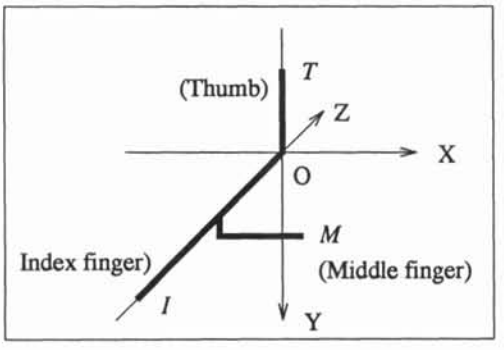

Figure 1: *A parameterised model of the hand with three fingers under orthographic projection. The optical centre is along the Z axis at large distance and the coordinate origin O is located at the cross of the thumb and the index finger. The Kalman filter tracks and predicts the location of finger tips that are marked by black blobs. The hand pose estimation algorithm recovers 5-degrees of freedom.*

Lipschitz condition is satisfied (derivation ommitted). A quadratic solution (Newton's method) is not necessary since the method requires compuation of the Jacobi matrix, and the Kalman filter provides a robust initial guess for the iteration.

3 Tracking and Kalman Filtering

We have implemented a standard Kalman filter for tracking of feature points [4, 2, 9]. Suppose that $[x(k)\ y(k)]^\mathsf{T}$ is the "true" image position of a point to be tracked, so that the system state is $\mathcal{X}(k) = [x(k)\ \dot{x}(k)\ y(k)\ \dot{y}(k)]^\mathsf{T}$. Assuming the state variables are uncorrelated, system dynamics is given by:

$$\mathcal{X}(k+1) = \underbrace{\begin{bmatrix} 1 & t_{k+1} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t_{k+1} \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}(k)} \cdot \mathcal{X}(k) + \underbrace{\begin{bmatrix} w_1(k) \\ w_2(k) \\ w_3(k) \\ w_4(k) \end{bmatrix}}_{\mathbf{w}(k)} \tag{6}$$

We denote $E[\mathbf{w}(k)] = \mathbf{R}$. Note that the observation interval t_k can vary; but we assume that corners move linearly between observations. The measured position of a feature point in the image is denoted by:

$$\mathcal{Y}(k) = \begin{bmatrix} x_m(k) \\ y_m(k) \end{bmatrix} \tag{7}$$

in which case

$$\mathcal{Y}(k) = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{C}} \cdot \mathcal{X}(k) + \underbrace{\begin{bmatrix} v_1(k) \\ v_2(k) \end{bmatrix}}_{\mathbf{v}k} \tag{8}$$

with $E[\mathbf{v}(k)] = \mathbf{Q}$. The prediction step of the Kalman filter is given by:

$$\hat{\mathcal{X}}(k+1|k) = \mathbf{A}(k) \cdot \hat{\mathcal{X}}(k|k)$$
$$\mathbf{P}(k+1|k) = \mathbf{A}(k)\mathbf{P}(k)\mathbf{A}^\mathsf{T}(k) + \mathbf{Q}(k),$$

where $\mathbf{P}(k)$ is the state covariance matrix at step k after k measurments. The state update step is given by:

$$\hat{\mathcal{X}}(k+1) = \mathbf{A}(k)\hat{\mathcal{X}}(k) + \mathbf{K}(k+1)[\mathcal{Y}(k+1) - \mathbf{C}\cdot\mathbf{A}\hat{\mathcal{X}}(k)]$$

in which the filter gain $\mathbf{K}$ is given by:

$$\mathbf{K}(k+1) = \mathbf{P}(k+1|k)\,\mathbf{C}^\mathsf{T}\,[\mathbf{C}\,\mathbf{P}(k+1|k)\,\mathbf{C}^\mathsf{T} + \mathbf{R}(k)]^{-1}$$

Finally, the state covariance matrix is updated by:

$$\mathbf{P}(k+1) = \mathbf{P}(k+1|k) - \mathbf{K}\cdot\mathbf{C}\cdot\mathbf{P}(k+1|k).$$

The track initialisation computes:

$$\hat{\mathcal{X}}(1) = \begin{bmatrix} x_1 \\ (x_1 - x_0)/t_1 \\ y_1 \\ (y_1 - y_0)/t_1 \end{bmatrix}.$$

4 Experiments

The algorithm TIM is able to track and estimate hand pose at video rate using a modest workstation (Sun Sparc-10) with on additional processing power. Our experiment is limited by the frame grabber to reach for frame rate. The computed pose is being fed into a pan-tilt device that can follow the action of the hand. The algorithm is stable and it does not require camera calibration. The parameterisation of the hand is straightforward, it in fact only requires relative measurements at the beginning of the tracking.

5 Summary

Given the correspondence of three points with a single camera, the motion is recoverable upto 5 degrees of freedom. We made assumptions in this approach that the hand model is known and is parameterised and that the projection is orthographical–setting free the need for camera calibration. Translational motion is measured in units of pixel. What we are interested in is the relative motion hence the accurate measure is not required. This is possible as long as the hand is placed at a distance to the camera, *ie* the object spans a small space against its distance from the viewer.

Possible extension to this work is to model the hand using perspective projection and recover from it the full 3D motion including the translational component along the optical axis. This is possible since the total number of unknowns are six which matches exactly the number of equations: each point correspondence provides two equations and there are three tracked points.

Acknowledgements

Authors would like to thank Professor D Mital for his support of experimental equipements and research lab. We also like to thank the technicians in Research Lab III for their speedy and kind assistance.

References

[1] K S Arun, T S Huang, and S D Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9:698–700, 1987.

[2] Y Bar-Shalom and T E Fortman. *Tracking and Data Association.* Academic, 1988.

[3] R. J. Blissett. Retriving 3D information from video for robot control and surveillance. *Electronics & Communication Engineering Journal*, pages 155–163, August 1990.

[4] S M Bozic. *Digital and Kalman Filtering.* Edward Arnold, UK, 1979.

[5] R. Cipolla, Y Okamoto, and Y Kuno. Robust structure from motion using motion parallax. In *Proc. IEEE ICCV*, pages 374–382, 1993.

[6] T S Huang and O D Faugeras. Some properties of the E matrix in two-view motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:1310–1312, 1989.

[7] T S Huang and C H Lee. Motion and structure from orthographic projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(5):536–540, May 1989.

[8] K Kanatani. *Geometric Computation for machine vision.* Oxford Press, 1993.

[9] S Lee and Y Kay. A Kalman filter approach for accurate 3-D motion estimation from a sequence of stereo images. *Computer Vision, Graphics and Image Processing: Image Understanding*, 54(2):244–258, Sept 1991.

[10] H C Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293-10:133–135, 1981.

[11] S. Ullman. *The interpretation of visual motion.* MIT Press, New Jersey, 1979.