Affine Structure from Motion

Scientific Computing CS660 Fall '11 Final Project
Angjoo Kanazawa

December 17, 2011

1 Introduction

Over the last few decades, the influence of Scientific Computing has been so prevelent in almost every area of Science and Engineering. It has become a necessary and critical tool for anyone involved in high-level research in Computer Science. Computer Vision is one of the quintessential example of a research area that heavily builds upon methods discovered in Scientific Computing, where the primary interest lies in the analysis and understanding of images which are represented in numerical matrices. This project explores applications of computational algorithms explored in Scientific Computing via tackling the problem of 3D reconstruction of an object from a stream of images.

The 3D reconstruction problem consists of a series of challenges starting from developing the cameral model and feature representation of the image, tracking such features over the image sequences , and finally reconstructing the 3D geometry from the tracked points. In all of these steps, The application of algorithms explored in Scientific Computing is ubiqutous in all of these steps.

The main challenge of recovering 3D geometry of objects and camera motion simultaneously from a set of point correspondences is referred to as the Structure from Motion (SfM) problem. The solution to the problem has a wide range of application in 3D modelling, virtual and augmented reality models in computer graphics, camera calibration and many more. It is a well studied problem with various approaches; this project focuses on the *factorization* method proposed by Tomasi and Kanade [**?**] under the orthographic camera projection model. The algorithm provides a numerically stable closed from optimal solution via the singular-value decomposition technique under certain conditions. [**?**, p. 435]

This project follows the Project 4 of Derek Hoiem's CS 543/ECE 549 course at the University of Illinois at Urbana-Champaign: project description.

2 Background

2.1 Orthographic Projection

A camera model transforms a world point into an image point. An affine camera, often used for its simplicity, preserves up to affine transformation of world points to image points. Basically it is a linear mapping followed by a translation, where points can rotate, scale, and translate but parallelism is preserved.[**?**, p.38]. An *orthographic camera model* is a specific type of affine camera where

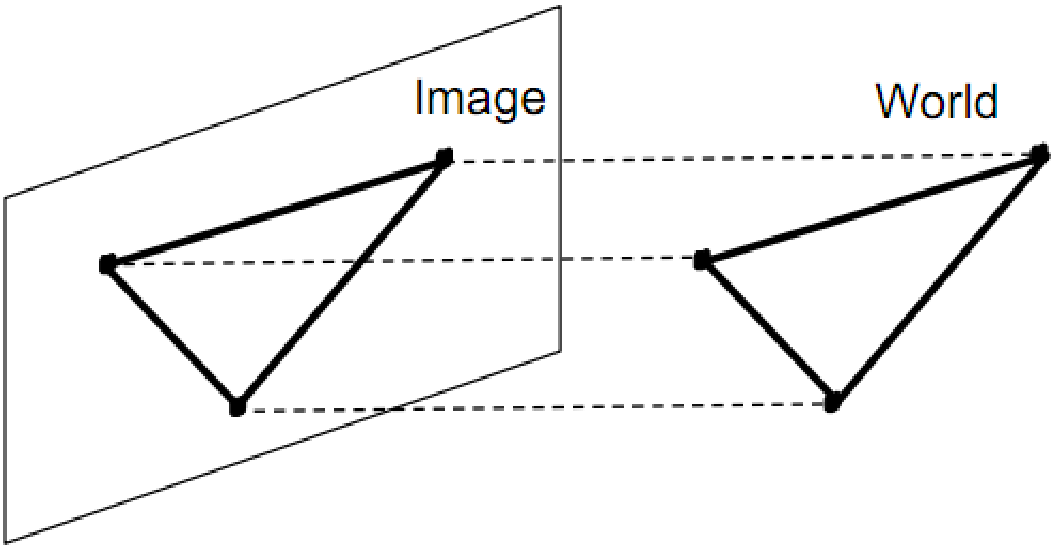

Figure 1: Orthographic Projection

the world points are projected in parallel onto the image plane, where the depth information of the world Z, is simply ignored. Mathematically,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}$$

$$x = PX + t$$

Where P is referred to as the projection matrix that does the mapping and t is the translation vector [**?**, p.172].

2.2 Notations and Assumptions

A "world point" refers to a point, in the 3D coordinate system, $\boldsymbol{X} = (X, Y, Z)^T$, and an "image point" refers to a point projected on to an image plane in the 2D coordinate system, $\boldsymbol{x} = (x, y)^T$.

$I(x, y)$ denotes the pixel intensity value of image I, and $I(x, y, t)$ denotes the pixel value of image I at time t. ∇I is the image gradient $[I_x I_y]$ and H is the image hessian $\begin{matrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{matrix}$

In this project, all projection is orthographic 2.1 and camera motion is affine. The world projected on the image sequences is rigid and the pixel intensities of images over sequences are constant i.e. all frames taken under static environment with no brightness changes. Also for the experiment we assume that there is no occlusion.

2.3 Problem Statement

Given F frames of sequencial images (videos), obtain P trajectory of image points for all F: $\{x_{fp} = (u_{fp}, v_{fp})^T \forall F, P$ and solve for X_p, the world coordinate of all P points. these observation.

3 Structure for Motion

There are three main components for 3D reconstruction of an image stream. First we need to select a subset of "interesting" pixels from the original frame. These points are then tracked through out the rest of the squence. This set of point correspondence across all frames is then fed into the factorization algorithm. Appropriate keypoint selection and accurate feature tracking are critical for a successful 3D reconstruction.

3.1 Keypoint Selection

It's not computationally possible to solve the world coordinate for every pixel of image sequences. We need to select keypoints in the image that are distinct, reliable and meaningful.

Harris corner detector is an optimal feature selection for the tracker that is used for this project. (It's optimality is discussed in section 3.2). The idea is that we should easily recognize an interesting point by looking through a small window, where shifting a window in any direction should give a large change in intensity i.e. we accept a point $\boldsymbol{x}$ if SSD of the displacement of the window by $(u, v)^T$, $E(u, v) = \sum_{(x,y)\in W}[I(x+u, y+v) - I(x, y)]^2$, is large in all direction. Assuming $\boldsymbol{d} = (u, v)^T$ is small, the taylor series expansion of I is

$$I(x+u, y+v) \approx I(x, y) + I_x u + I_y v + \mathcal{O}(d^T d)$$

where $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$.

Then,

$$\begin{aligned}
E(u, v) &= \sum_{(x,y)\in W} [I(x+u, y+v) - I(x, y)]^2 \\
&= \sum_{(x,y)\in W} [I(x, y) + I_x u + I_y v - I(x, y)]^2 \\
&= \sum_{(x,y)\in W} (\begin{pmatrix} I_x & I_y \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix})^2 \\
&= \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \\
&= d^T H d.
\end{aligned}$$

Since the two eigenvalues of H ,λ_1, λ_2, denotes the amount of change in the direction of its corresponding eigenvector, we accept a window if $min(\lambda_1, \lambda_2) > \tau$, where τ is predefined threshold [?]. Figure 2 illustrates the components of the Hessisan of frame one.

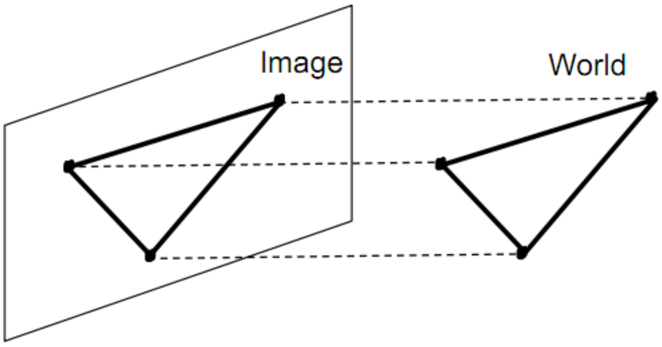

Figure 2: Plot of the image gradients of the first frame

3.2 Feature Tracking

Now using the detected keypoint, we track these points over all frames using the Kanade-Lucas-Tomasi (KLT) Tracker [?]. The basis of KLT tracker is that although image intensities change as camera moves, images taken at near time are strongly related to each other because they tend to refer to the same scene given the assumption of local brightness constancy. This means that under an ideal static environment, image at time $t+1$ can be obtained by moving every pixl in the at time image t by a suitable amount [?].

This displacement can be either a translation or an affine mapping or the combination of both, but since the world is rigid we will only consider translation. Mathematically, given $\boldsymbol{x}$ at frame t, we want to find a displacement vector, $\boldsymbol{d} = (u, v)^T$, that minimizes the dissimilarity

$$I(\boldsymbol{x} + \boldsymbol{d}, t+1) - I(\boldsymbol{x}, t) = 0 \tag{1}$$

Again, the taylor expansion of $I(\boldsymbol{x} + \boldsymbol{d}, t+1) = I(x+u, y+v, t+1) \approx I(x, y, t+1) + I_x u + I_y v + \mathcal{O}(d^T d)$. Substituting that back to (1),

$$\begin{aligned} 0 &= I(\boldsymbol{x} + \boldsymbol{d}, t+1) - I(\boldsymbol{x}, t) \\ &= \nabla I(\boldsymbol{x}) \cdot \boldsymbol{d} - I_t(\boldsymbol{x}) \end{aligned}$$

Where $I_t(\boldsymbol{x}) = I(\boldsymbol{x}, t) - I(\boldsymbol{x}, t+1)$. Here we have 2 unknowns but only one constraint. In fact even with the brightness constancy assumption, it is difficult to track a single point unless the point is extremely distinctive [?]. Therefore we consider minimizing the dissimilarity within a small window (typically around 15 x 15) and obtain an overdetermined linear system of equations:

$$\sum_{\boldsymbol{x} \in W} \nabla I(\boldsymbol{x}) \cdot \boldsymbol{d} = \sum_{\boldsymbol{x} \in W} I_t$$
$$A\boldsymbol{d} = \boldsymbol{t}$$

Using the normal equations, we solve this least linear squares problem:

$$A^T A\boldsymbol{d} = A^T \boldsymbol{t}$$

$$\begin{pmatrix} I_x(x_1) & \dots & I_x(x_{|W|}) \\ \vdots & & \vdots \\ I_y(x_1) & \dots & I_y(x_{|W|}) \end{pmatrix} \begin{pmatrix} I_x(x_1) & \dots & I_y(x_1) \\ \vdots & & \vdots \\ I_x(x_{|W|}) & \dots & I_y(x_{|W|}) \end{pmatrix} \boldsymbol{d} = \begin{pmatrix} I_x(x_1) & \dots & I_y(x_1) \\ \vdots & & \vdots \\ I_x(x_{|W|}) & \dots & I_y(x_{|W|}) \end{pmatrix} \begin{pmatrix} I_t(x_1) \\ \vdots \\ I_t(x_{|w|}) \end{pmatrix}$$

$$\sum_{\boldsymbol{x} \in W} \begin{pmatrix} I_x^2 & I_x I_y \\ I_y^2 & I_y I_x \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \sum_{\boldsymbol{x} \in W} \begin{pmatrix} I_x I_t \\ I_y I_t \end{pmatrix}$$

3.2.1 Numerical Stability

We can dolve for $\boldsymbol{d}$ in In analyzing of the numerical stability of the solution of (**??**), we can discuss the optimality of the keypoint selection. Note that

3.3 The Factorization Method

4 Results and Analysis

5 Conclusion

Affine cameras combine the effect of affine transformation in the 3D space, orthographic projection, and an affine transformation in the 2D image space. i.e.

$$P = [\text{3 by 3 affine transformation}] \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} [\text{4 by 4 affine transformation}]$$

For factorization, we can explicitly denote the mapping and translation of the projection matrix P, and write it as a linear combination of mapping and translation. In inhomogeneous coordinates, this projection is $x = RX + t$, where t is the translation and R is the rotation/orientation of the camera.

For SfM, the image sequence is represented by a $2F \times P$ *measurement matrix* W, where $w_{fp} = (x_{fp}, y_{fp})^T$, and P is the number of points tracked through F frames.

To get rid of the translation term, we center the image points by subtracting the mean (centroid) of image points, and assume that the world coordinate system is at the centroid of the 3D points. Then, the orientation (rotation) of the camera at frame f is represented by orthonormal vectors $i_f, j_f, k_f \in \mathbf{R}^3$, where each vector corresponds to the x, y, and z-axis of the image plane respectively. Under orthography with the z axis along the optical axis, these vectors over F frames are collected into a *motion matrix* $M \in \mathbf{R}^{2F \times 3}$

$$M = \begin{pmatrix} i_1^T \\ \vdots \\ i_F^T \\ j_1^T \\ \vdots j_F^T \end{pmatrix}$$

We let $S_p = (X_p, Y_p, Z_p)^T$ be the 3D coordinates of feature p in the fixed world point with the same origin. These vectors are collected into a *shape matrix* $S \in \mathbf{R}^{3\times P}$ s.t. $S = \begin{pmatrix} s_1 & \cdots & s_p \end{pmatrix}^T$. Using this notation, for a single frame we get

$$\begin{pmatrix} x_{fp} \\ y_{fp} \end{pmatrix} = \begin{pmatrix} i_f^T \\ j_f^T \end{pmatrix} \begin{pmatrix} X_p \\ Y_p \\ Z_p \end{pmatrix}$$

$$w_{fp} = M_f S_p$$

So for all frames, we have the equation

$$W = MS$$

Our goal is to estimate $\hat{M}$ and $\hat{S}$, s.t. $\hat{W} = \hat{M}\hat{S}$, our estimated measurement matrix, and the actual W is minimized i.e.

$$\min_{M,S} ||W - \hat{M}\hat{S}||^2$$

a least squares problem. [**?**] proved that under this model, the rank of W is 3. So we can achieve the least squares approximation by factoring W by SVD. Namely,

$$\begin{aligned} W &= U\Sigma V^T \\ &\approx U_3\tilde{\Sigma}V_3^T \text{ because } W \text{ is rank 3} \\ &= \begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ u_{2F,1} & u_{2F,2} & u_{2F,3} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} \begin{pmatrix} v_{1,1} & \cdots & \cdots & v_{1,p} \\ v_{2,1} & \cdots & \cdots & v_{2,p} \\ v_{3,1} & \cdots & \cdots & v_{3,p} \end{pmatrix} \end{aligned}$$

Where possible solution is to choose $\hat{M} = U_3\tilde{\Sigma}^{1/2}$ and $\hat{S} = \tilde{\Sigma}^{1/2}V_3^T$. This solution can be refined by eliminating affine ambiguity, but this algorithm is numerically stable and it is guaranteed to converge to the global minimum of the least squares problem.