# Scientific Computing CS660 Fall '11 Final Project Proposal
# Affine Structure from Motion: an Application of Singular Value Decomposition

Angjoo Kanazawa

December 17, 2011

## 1 Introduction

For this project I propose to explore an application of SVD in Computer Vision domain. Structure from Motion (SfM) is the problem of recovering 3D scene geometry and camera motion from a sequence of 2D images. SfM has a wide range of application including 3D model reconstruction of real-world objects, 3D motion matching for computer graphics, virtual and augmented reality models, camera calibration and many more.

SfM is a well studied problem with multitude of approaches and problem statements. The project will focus on the *factorization* method proposed by Tomasi and Kanade, which recovers the structure and motion of video sequences under the orthographic projection model. With this model, the shape and motion can be recovered simultaneously using a Singular Value Decomposition [**?**]. This project will follow the project 4 of Derek Hoiem's CS 543/ECE 549 course at the University of Illinois at Urbana-Champaign: project description.

## 2 Outline

The outline of the project is as follows:

1. Description of the orthographic camera projection model
2. Assumptions and problem statement
3. Tomasi-Kanade Factorization method
4. Implementation of SfM: recovering a 3D point cloud of a short video sequence using the factorization method (using supplimental materials from the UIUC course web site)
5. Results and analysis

# 3 Overview of the Factorization Method

This will be the main part of the project, but to demonstrate how matrix factorization theorems we covered in class will be used, here is an overview of the Tomasi-Kanade Factorization Method:

## 3.1 Affine Cameras

Projective geometry illustrates relationship between a single 2D image point and it's corresponding 3D world point as (written in homogeneous coordinates for numerical stability reasons):

$$\begin{pmatrix} fx \\ fy \\ w \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

$$x = PX$$

Where $f$ is the focal length of the camera, $P$ is the projection matrix of the camera, and the 2D location of the point in inhomogeneous coordinate is $(fx/w, fy/w)$. For SfM, given $m$ images of $n$ fixed points, we have the equation

$$x_{ij} = P_i X_j \ i = 1, \ldots, m, j = 1, \ldots, n.$$

Orthography is a special case of perspective projection, where the 3D world points are projected in parallel onto the image plane i.e. the distance from the center of projection to the image is infinite and $Z$ has no influence. i.e. we have the equation

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

## 3.2 Factorization Method

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.