Video Representation Via 3D Shaped Mosaics

Pedro M. Q. Aguiar *
ISR, Instituto Superior Técnico
Lisboa, Portugal
aguiar@isr.ist.utl.pt

José M. F. Moura
ECE Dep, Carnegie Mellon University
Pittsburgh, PA
moura@ece.cmu.edu

Abstract

We generalize to 3D shaped mosaics the **Generative Video** *representation of video sequences introduced by Jasinschi and Moura. Using a parametric representation of the 3D shape, we recover the 3D shape and 3D motions from the 2D motions in the video sequence. In this paper we consider piecewise planar object shapes under orthography and demonstrate our approach with a real life video clip.*

1 Introduction

Generative Video (GV), introduced by Jasinschi and Moura, e.g., [1], reduces video sequences to world images and ancillary data. The world images represent the background and any moving objects, while the ancillary data describes for example the motions (camera and objects). In the original formulation of **GV**, the world images are modeled as simple planar scenarios. This representation fails when the relative depth of the scene structure is not negligible. In this paper we recover 3D world image representations for the video sequence. Within our framework, the major task is to recover the 3D structure (3D shape and 3D motion) from a 2D video sequence.

Previous related work Consider the case of a single rigid body object, moving relative to the camera, at a large distance when compared to the object depth. In this scenario, to recover the 3D shape by estimating the absolute depth is inaccurate. Tomasi and Kanade, e.g., [2], introduced a method to recover structure from motion without computing the absolute depth as an intermediate step. In the approach of Tomasi and Kanade, the object shape is represented by the 3D position of a set of feature points. The 2D projection of each feature point is tracked along the image sequence. The 3D shape and motion are then estimated by factorizing a measurement matrix whose entries are the set of trajectories of the feature point projections.

Proposed approach We represent parametrically the 3D shape of the rigid body object and apply *Maximum Likelihood* (ML) estimation. The observations are the orthographic projection of the object texture plus noise. The problem is now to estimate from the given sequence of images all the unknowns involved (3D shape parameters, 3D motion parameters, and object texture). We obtain a feasible approximation to the ML estimator by showing that the parametric representation of the 3D rigid body shape induces a parametric model for the optical flow in the 2D image plane. We estimate the optical flow parameters using known techniques and apply *Least Squares* (LS) to resolve the 3D shape and motion parameters from the optical flow parameter estimates. Our technique was introduced in reference [3] where we considered that the world is 2D and the images are 1D projections of the world.

Our method relates to that of Tomasi and Kanade [2] in two very fundamental ways: we estimate directly the 3D shape, instead of computing depth as an intermediate step; and our algorithm leads to the factorization of a measurement matrix. Aside from this, our approach is different. We do not rely on the tracking of feature points. Instead, we use a parametric description of the 3D shape and recover the 3D structure from the parameterization induced in the optical flow. The advantage of our approach is two-fold. First, the tracking of feature points may be unreliable when processing noisy video sequences. The work in reference [2] assumes a very short interval between frames for easy feature tracking. We make no such assumption because large displacements are taken care of by a multiresolution approach to the estimation of the optical flow parameters. Second, we estimate 3D shape and motion from a sequence of few flow parameters instead of needing to process a large set of feature trajectories.

Paper overview Section 2 formulates the problem, derives the ML-based cost function, and solves for the texture estimate. Section 3 describes our approach to the recovery of the 3D shape and 3D motion for piecewise planar shapes. We demonstrate our approach by analyzing a real life video clip in section 4. Section 5 concludes the paper.

2 Problem Formulation

Observation model The frame $\boldsymbol{I}_f$ captured at time f, is modeled as a noisy observation of the orthogonal projection $\mathcal{P}$ of the rigid object $\mathcal{O}$ (assumed segmented)

$$\boldsymbol{I}_f = \mathcal{P}(\mathcal{O}, \boldsymbol{m}_f) + \boldsymbol{W}_f \qquad (1)$$

$\boldsymbol{m}_f$ defines the position and orientation of the rigid object relative to the camera coordinate system. For simplicity, the noise $\boldsymbol{W}_f$ is white, Gaussian.

*The first author was partially supported by INVOTAN.

The object $\mathcal{O}$ is described by its 3D shape $\mathcal{S}$ and texture $\mathcal{T}$. We model the shape $\mathcal{S}$ by a parametric description $\mathcal{S}(\boldsymbol{a})$ of the surface of the object, where $\boldsymbol{a}$ is an unknown vector. The texture $\mathcal{T}$ represents the light received by the camera after reflecting on the object surface. Texture depends on the object surface photometric properties, as well as on the environment illumination conditions. We assume the texture at a given surface point does not change with time. The operator $\mathcal{P}$ returns a real valued function defined over the image plane. $\mathcal{P}$ is a nonlinear mapping of $\mathcal{T}$ that depends on the object shape $\mathcal{S}$ and the object position $\boldsymbol{m}_f$. The intensity level of the projection of the object at pixel $\boldsymbol{u}$ on the image plane in terms of $\mathcal{T}$ is

$$\mathcal{P}\left(\mathcal{O}, \boldsymbol{m}_f\right)(\boldsymbol{u}) = \mathcal{T}\left(\boldsymbol{s}_f(\mathcal{S}, \boldsymbol{m}_f; \boldsymbol{u})\right) \tag{2}$$

where $\boldsymbol{s}_f(\mathcal{S}, \boldsymbol{m}_f; \boldsymbol{u})$ is the nonlinear mapping that lifts the point $\boldsymbol{u}$ on image $\boldsymbol{I}_f$ to the corresponding point on the 3D object surface. This mapping $\boldsymbol{s}_f(\mathcal{S}, \boldsymbol{m}_f; \boldsymbol{u})$ is determined by the object shape $\mathcal{S}(\boldsymbol{a})$, and the position $\boldsymbol{m}_f$. To simplify the notation, we will write explicitly only the dependence on f, i.e., $\boldsymbol{s}_f(\boldsymbol{u})$. Let $\boldsymbol{u}_f(\boldsymbol{s})$ be the inverse map of $\boldsymbol{s}_f(\boldsymbol{u})$. The point $\boldsymbol{s}$ on the surface of the object projects onto $\boldsymbol{u}_f(\boldsymbol{s})$ on the image. The mapping $\boldsymbol{u}_f(\boldsymbol{s})$, seen as a function of the frame index f, for a particular surface point $\boldsymbol{s}$, is the trajectory of the projection of that point in the image, i.e., the motion induced in the image, usually referred to as *optical flow*. In the sequel, we refer to the map $\boldsymbol{u}_f(\boldsymbol{s})$ as the optical flow map. The observation model (1) is rewritten by using (2) as

$$\boldsymbol{I}_f = \mathcal{T}\left(\boldsymbol{s}_f(\boldsymbol{u})\right) + \boldsymbol{W}_f \tag{3}$$

ML estimate Given the observation model (3), the 3D shape and the 3D motion of the object $\mathcal{O}$ are recovered from the video sequence $\{\boldsymbol{I}_f, 1 \leq f \leq F\}$ by estimating all the unknown parameters: the 3D shape parameter $\boldsymbol{a}$; the texture $\mathcal{T}$; and the set of 3D positions of the object $\{\boldsymbol{m}_f, 1 \leq f \leq F\}$ with respect to the camera.

With the noise $\boldsymbol{W}_f$ zero mean, white Gaussian, the ML estimate minimizes the cost function C_{ML} defined as

$$C_{\mathrm{ML}}\left(\boldsymbol{a}, \mathcal{T}, \{\boldsymbol{m}_f\}\right) = \sum_{f=1}^{F} \int \left[\boldsymbol{I}_f(\boldsymbol{u}) - \mathcal{T}\left(\boldsymbol{s}_f(\boldsymbol{u})\right)\right]^2 \, d\boldsymbol{u} \tag{4}$$

Texture estimate We show in [4] that the ML estimate $\widehat{\mathcal{T}}(\boldsymbol{s})$ that minimizes C_{ML} is

$$\widehat{\mathcal{T}}(\boldsymbol{s}) = \frac{\sum_{f=1}^{F} \boldsymbol{I}_f(\boldsymbol{u}_f(\boldsymbol{s})) J_f(\boldsymbol{s})}{\sum_{f=1}^{F} J_f(\boldsymbol{s})} \tag{5}$$

where the function $J_f(\boldsymbol{s})$ is the Jacobian of the mapping $\boldsymbol{u}_f(\boldsymbol{s})$, $J_f(\boldsymbol{s}) = |\nabla \boldsymbol{u}_f(\boldsymbol{s})|$. Expression (5) states that the estimate of the texture of the object at the surface point $\boldsymbol{s}$ is a weighted average of the measures of the intensity level corresponding to that surface point.

By inserting the texture estimate $\widehat{\mathcal{T}}$ in (4), we express C_{ML} in terms of the optical flow mappings $\{\boldsymbol{u}_f(\boldsymbol{s})\}$. After manipulations, see [4], C_{ML} is written as

$$C_{\mathrm{ML}} = \sum_{f=2}^{F} \sum_{g=1}^{f-1} \int \left[I_f(u_f(s)) - I_g(u_g(s))\right]^2 \frac{J_f(s) J_g(s)}{\sum_{h=1}^{F} J_h(s)} \, ds \tag{6}$$

Eliminating the dependence on the texture, we are left with a cost function that depends on the *structure* (3D shape $\mathcal{S}(\boldsymbol{a})$ and 3D motion $\{\boldsymbol{m}_f\}$) only through the *motion* induced in the image plane, i.e., through the optical flow mappings $\{\boldsymbol{u}_f(\boldsymbol{s})\}$. Recall that $\boldsymbol{u}_f$ depends on the shape $\mathcal{S}$ and the motion $\boldsymbol{m}_f$.

Summary of the approach To recover the 3D shape and the 3D motion of the object $\mathcal{O}$ from the image sequence, we do not attempt the direct minimization of C_{ML} over the parameters $\boldsymbol{a}$ and $\{\boldsymbol{m}_f\}$. Rather, we exploit the constraints induced on the optical flow by the orthogonality of the projection $\mathcal{P}$, the rigidity of the motions (rigid object), and the parameterization of the surface shape of the object. The constraints induced on the optical flow enable us to parameterize the optical flow mapping $\boldsymbol{u}_f(\boldsymbol{s})$ in terms of a parameter vector $\boldsymbol{\alpha}_f$ as $\boldsymbol{u}(\boldsymbol{\alpha}_f; \boldsymbol{s}), 1 \leq f \leq F$. The parameter vector $\boldsymbol{\alpha}_f$ is directly related to the 3D shape parameter $\boldsymbol{a}$ and the 3D position $\boldsymbol{m}_f$, as will be shown in section 3, i.e., $\boldsymbol{\alpha}_f = \boldsymbol{\alpha}(\boldsymbol{a}, \boldsymbol{m}_f)$. The steps of our approach to recover the 3D structure, i.e., the 3D shape parameter $\boldsymbol{a}$ and the set of 3D positions $\{\boldsymbol{m}_f\}$, are summarized as: **i)** Given the image sequence $\{\boldsymbol{I}_f, 1 \leq f \leq F\}$, estimate the set of time varying optical flow vectors $\{\boldsymbol{\alpha}_f, 1 \leq f \leq F\}$ parameterizing the optical flow mappings $\{\boldsymbol{u}(\boldsymbol{\alpha}_f; \boldsymbol{s})\}$. **ii)** From the sequence of estimates $\{\widehat{\boldsymbol{\alpha}}_f, 1 \leq f \leq F\}$, solve for the shape parameter vector $\boldsymbol{a}$ and the object positions $\{\boldsymbol{m}_f, 1 \leq f \leq F\}$.

3 Piecewise Planar Surface

Attach a coordinate system to the object given by the axes labeled by x, y, and z. We consider objects whose shape is given by a piecewise planar surface with K patches. The shape parameter vector $\boldsymbol{a}$ collects the coefficients $\boldsymbol{a} = \left\{a_{00}^k, a_{10}^k, a_{01}^k, 1 \leq k \leq K\right\}$ where

$$1 \leq k \leq K; \qquad z = a_{00}^k + a_{10}^k x + a_{01}^k y \tag{7}$$

describes the shape of the patch k.

To capture the 3D motion of the rigid object, we attach a coordinate system to the camera given by the axes u, v, and w, see figure 1. We express the object position at time instant f in terms of $\left(t_{uf}, t_{vf}, t_{wf}, \theta_f, \phi_f, \psi_f\right)$ where $\left(t_{uf}, t_{vf}, t_{wf}\right)$ are the coordinates of the origin of the object coordinate system with respect to the camera coordinate system (3D translation), and $(\theta_f, \phi_f, \psi_f)$ determine the orientation of the object coordinate system relative to the camera coordinate system (3D rotation).

The point (x, y, z) projects at time instant f on the

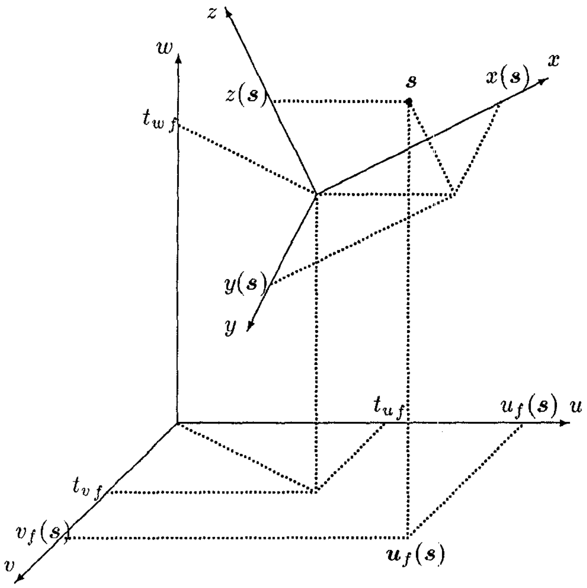

Figure 1: Object and camera coordinate systems.

image coordinates $\boldsymbol{u}_f = [u_f, v_f]^T$ given by

$$\boldsymbol{u}_f = \begin{bmatrix} i_{xf} & i_{yf} & i_{zf} \\ j_{xf} & j_{yf} & j_{zf} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_{uf} \\ t_{vf} \end{bmatrix} \tag{8}$$

where the matrix that multiplies $[x, y, z]^T$ is a submatrix of the well known 3D rotation matrix, determined by the angles $(\theta_f, \phi_f, \psi_f)$:

$$i_{xf} = \cos\theta_f \cos\phi_f \tag{9}$$
$$i_{yf} = \sin\phi_f \tag{10}$$
$$i_{zf} = -\sin\theta_f \cos\phi_f \tag{11}$$
$$j_{xf} = \sin\theta_f \sin\psi_f - \cos\theta_f \sin\phi_f \cos\psi_f \tag{12}$$
$$j_{yf} = \cos\phi_f \cos\psi_f \tag{13}$$
$$j_{zf} = \sin\theta_f \sin\phi_f \cos\psi_f + \cos\theta_f \sin\psi_f \tag{14}$$

Expression (8) shows that the orthogonal projection is insensitive to the translation along the w axis. This reflects the fact that under orthography the absolute depth can not be estimated. Only the set of positions $\left\{\boldsymbol{m}_f = \left\{t_{uf}, t_{vf}, \theta_f, \phi_f, \psi_f\right\}, 1 \leq f \leq F\right\}$ can be estimated from the image sequence.

3.1 Optical flow

We show that the optical flow mappings $\{\boldsymbol{u}_f(\boldsymbol{s})\}$ are described parametrically. Choose the coordinate $\boldsymbol{s} = [s, r]^T$ of the texture function to coincide with the coordinates $[x, y]$ of the object coordinate system. Also, the object coordinate system and the camera coordinate system coincide in the first frame. A point with coordinate $\boldsymbol{s} = [s, r]^T$ in the object surface projects on $\boldsymbol{u} = [x, y]^T = [s, r]^T = \boldsymbol{s}$ in the first frame, so $\boldsymbol{u}_1(\boldsymbol{s}) = \boldsymbol{s}$. At time f, that point projects according to (8). For a point $\boldsymbol{s}$ that falls in patch k of the object surface, we have

$$\boldsymbol{u}_f(\boldsymbol{s}) = \begin{bmatrix} i_{xf} & i_{yf} & i_{zf} \\ j_{xf} & j_{yf} & j_{zf} \end{bmatrix} \begin{bmatrix} s \\ r \\ a_{00}^k + a_{10}^k s + a_{01}^k r \end{bmatrix} + \begin{bmatrix} t_{uf} \\ t_{vf} \end{bmatrix} \tag{15}$$

Define the set of parameters $\boldsymbol{\alpha}_f^k = \left\{\alpha_{fmn}^{uk}, \alpha_{fmn}^{vk}\right\}$ as the coefficients of the powers of s and r above,

$$\alpha_{f00}^{uk} = t_{uf} + i_{zf} a_{00}^k, \qquad \alpha_{f00}^{vk} = t_{vf} + j_{zf} a_{00}^k \tag{16}$$
$$\alpha_{f10}^{uk} = i_{xf} + i_{zf} a_{10}^k, \qquad \alpha_{f10}^{vk} = j_{xf} + j_{zf} a_{10}^k \tag{17}$$
$$\alpha_{f01}^{uk} = i_{yf} + i_{zf} a_{01}^k, \qquad \alpha_{f01}^{vk} = j_{yf} + j_{zf} a_{01}^k \tag{18}$$

The optical flow between frames $\boldsymbol{I}_1$ and $\boldsymbol{I}_f$ in the image of the surface patch k is written in terms of the optical flow parameter vector $\boldsymbol{\alpha}_f^k$ as

$$\boldsymbol{u}_f(\boldsymbol{s}) = \boldsymbol{u}(\boldsymbol{\alpha}_f^k; s, r) = \begin{bmatrix} \alpha_{f00}^{uk} + \alpha_{f10}^{uk} s + \alpha_{f01}^{uk} r \\ \alpha_{f00}^{vk} + \alpha_{f10}^{vk} s + \alpha_{f01}^{vk} r \end{bmatrix}$$

The optical flow parametrization above is usually referred to as *affine motion model.* The ML estimation of $\left\{\boldsymbol{\alpha}_f^k\right\}$ leads to the minimization of C_{ML} given by (6) with respect to the set of vectors $\left\{\boldsymbol{\alpha}_f^k, 1 \leq f \leq F, 1 \leq k \leq K\right\}$ parameterizing the mappings $\{\boldsymbol{u}_f(\boldsymbol{s}) = \boldsymbol{u}(\boldsymbol{\alpha}_f; \boldsymbol{s})\}$. In practice, this is a highly complex task. A more feasible and practical solution decouples the estimation of each vector $\boldsymbol{\alpha}_f$ from the estimation of the remaining vectors $\boldsymbol{\alpha}_g, g \neq f$, by simplifying the cost function (6). Instead of using all possible pairs of frames, compare all frames with respect to the first frame $\boldsymbol{I}_1$. Also, neglect the weighting term $\frac{J_f(\boldsymbol{s}) J_g(\boldsymbol{s})}{\sum_{h=1}^{F} J_h(\boldsymbol{s})}$, and obtain the usual expression for the optical flow estimation:

$$\left\{\widehat{\boldsymbol{\alpha}_f^k}\right\} = \arg \min_{\left\{\boldsymbol{\alpha}_f^k\right\}} \sum_{k=1}^{K} \int_{\mathcal{R}_k} \left[\boldsymbol{I}_f\left(\boldsymbol{u}(\boldsymbol{\alpha}_f^k; \boldsymbol{s})\right) - \boldsymbol{I}_1(\boldsymbol{s})\right]^2 ds$$

We compute the optical flow parameter vector estimates $\left\{\widehat{\boldsymbol{\alpha}_f^k}\right\}$, by using known techniques, see references [5] and [6].

3.2 3D Structure from 2D Optical Flow

The set of equations (16-18) defines an overconstrained system with respect to the 3D shape parameters $\left\{a_{00}^k, a_{01}^k, a_{10}^k,\ 1 \leq k \leq K\right\}$ and to the 3D positions $\left\{t_{uf}, t_{vf}, \theta_f, \phi_f, \psi_f,\ 1, \leq f, \leq F\right\}$. The estimate $\left\{\hat{a}_{mn}^k\right\}$ of the object shape and the estimate

$\left\{\widehat{t_{uf}}, \widehat{t_{vf}}, \widehat{\theta_f}, \widehat{\phi_f}, \widehat{\psi_f}\right\}$ of the object positions are the *Least Squares* (LS) solution of the system. We first solve for the translation, leading to a closed-form solution. Then, replace the translation estimate and solve for the remaining motion parameters and shape parameters by using a two-step iterative method.

Translation estimation The translation components along the camera plane at instant f, t_{uf} and t_{vf} only affect, respectively, the set of parameters $\left\{\alpha^{uk}_{f00}\right\}$ and $\left\{\alpha^{vk}_{f00}\right\}$. If the parameters $\left\{a^k_{00}\right\}$ and $\{\theta_f, \phi_f, \psi_f\}$ are known, the LS estimate of $\left\{t_{uf}, t_{uf}\right\}$ is given by

$$\widehat{t_{uf}} = \frac{\sum_{k=1}^{K} \hat{\alpha}^{uk}_{f00} - i_{zf} \sum_{k=1}^{K} a^k_{00}}{K} \tag{21}$$

$$\widehat{t_{vf}} = \frac{\sum_{k=1}^{K} \hat{\alpha}^{vk}_{f00} - j_{zf} \sum_{k=1}^{K} a^k_{00}}{K} \tag{22}$$

Without loss of generality, we choose the object coordinate system in such a way that $\sum_{k=1}^{K} a^k_{00} = 0$ (the first frame only restricts the rotation and the component of the translation parallel to the image plane, so we have freedom to move the coordinate system along the axis orthogonal to the image plane). With this choice, we obtain the estimates

$$\widehat{t_{uf}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\alpha}^{uk}_{f00}, \qquad \widehat{t_{vf}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\alpha}^{vk}_{f00} \tag{23}$$

Optical flow parameters matrix Replace the set of translation estimates $\left\{\widehat{t_{uf}}, \widehat{t_{vf}}\right\}$ given by (23) in the equation set (16-18). Define a set of parameters $\left\{\beta^{uk}_f, \beta^{vk}_f\right\}$ related to $\left\{\alpha^{uk}_{f00}, \alpha^{vk}_{f00}\right\}$ by

$$\beta^{uk}_f = \alpha^{uk}_{f00} - \frac{1}{K}\sum_{l=1}^{K} \alpha^{ul}_{f00}, \quad \beta^{vk}_f = \alpha^{vk}_{f00} - \frac{1}{K}\sum_{l=1}^{K} \alpha^{vl}_{f00} \tag{24}$$

Collect the parameters $\left\{\beta^{uk}_f, \alpha^{uk}_{fmn}, \beta^{vk}_f, \alpha^{vk}_{fmn}\right\}$ in the matrix $\boldsymbol{R}$, which we call the optical flow parameters matrix,

$$R = \left[\begin{array}{ccccccc} \beta^{u1}_1 & \alpha^{u1}_{110} & \alpha^{u1}_{101} & \cdots & \beta^{uK}_1 & \alpha^{uK}_{110} & \alpha^{uK}_{101} \\ \beta^{u1}_2 & \alpha^{u1}_{210} & \alpha^{u1}_{201} & \cdots & \beta^{uK}_2 & \alpha^{uK}_{210} & \alpha^{uK}_{201} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \beta^{u1}_F & \alpha^{u1}_{F10} & \alpha^{u1}_{F01} & \cdots & \beta^{uK}_F & \alpha^{uK}_{F10} & \alpha^{uK}_{F01} \\ \beta^{v1}_1 & \alpha^{v1}_{110} & \alpha^{v1}_{101} & \cdots & \beta^{vK}_1 & \alpha^{vK}_{110} & \alpha^{vK}_{101} \\ \beta^{v1}_2 & \alpha^{v1}_{210} & \alpha^{v1}_{201} & \cdots & \beta^{vK}_2 & \alpha^{vK}_{210} & \alpha^{vK}_{201} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \beta^{v1}_F & \alpha^{v1}_{F10} & \alpha^{v1}_{F01} & \cdots & \beta^{vK}_F & \alpha^{vK}_{F10} & \alpha^{vK}_{F01} \end{array}\right]$$

and the motion and shape parameters in matrices $\boldsymbol{M}$ and $\boldsymbol{S}$, as follows

$$M = \left[\begin{array}{cccccccc} i_{x1} & i_{x2} & \cdots & i_{xF} & j_{x1} & j_{x2} & \cdots & j_{xF} \\ i_{y1} & i_{y2} & \cdots & i_{yF} & j_{y1} & j_{y2} & \cdots & j_{yF} \\ i_{z1} & i_{z2} & \cdots & i_{zF} & j_{z1} & j_{z2} & \cdots & j_{zF} \end{array}\right]^T$$

$$S^T = \left[\begin{array}{ccccccc} 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 1 \\ a^1_{00} & a^1_{10} & a^1_{01} & \cdots & a^K_{00} & a^K_{10} & a^K_{01} \end{array}\right] \tag{27}$$

These enable us to factorize the optical flow matrix according to (16-18) and (24)

$$\boldsymbol{R} = \boldsymbol{M}\boldsymbol{S}^T \tag{28}$$

Matrix $\boldsymbol{R}$ is $2F \times 3K$. In a noiseless situation, $\boldsymbol{R}$ is rank deficient (rank 3) due to the redundancy of the observations imposed by the 3D rigid structure of the object. With noisy observations, $\boldsymbol{R}$ may be full rank but the 3 largest eigenvalues should contain most of the energy.

Factorization To estimate the shape and motion parameters we use a two-step iterative method. The steps are: **i)** solve for shape with known positions, and **ii)** solve for positions with known shape. We will see that step **i)** leads to a linear LS problem and step **ii)**, although nonlinear, decouples the estimation of the positions at different instants. The initialization is done by computing the *Singular Value Decomposition* (SVD) of the matrix $\boldsymbol{R}$ and selecting the 3 largest eigenvalues. We get $\boldsymbol{R} \simeq \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U}$ is $2F \times 3$, $\boldsymbol{\Sigma}$ is diagonal 3×3, and $\boldsymbol{V}^T$ is $3 \times 3K$. We initialize $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}$ and $\boldsymbol{S}^T = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}V^T$ where $\boldsymbol{A}$ is a non-singular 3×3 matrix determined by the constraints imposed by the structure of the matrix $\boldsymbol{M}$ (expressions (9-14) and (26)). This step is similar to the procedure in reference [2], see [4] for the details.

Shape estimate for known motion The shape is described by the third row of the matrix $\boldsymbol{S}^T$ which we denote by $\boldsymbol{s}_3^T$. Given the matrix $\boldsymbol{M}$, we denote by $\tilde{\boldsymbol{R}}$ a matrix equal to $\boldsymbol{R}$ at all entries except for the set $\left\{\alpha^{uk}_{f10}, \alpha^{uk}_{f01}, \alpha^{vk}_{f10}, \alpha^{vk}_{f01}\right\}$ which is replaced by $\left\{\gamma^{uk}_{f10}, \gamma^{uk}_{f01}, \gamma^{vk}_{f10}, \gamma^{vk}_{f10}\right\}$ defined as

$$\gamma^{uk}_{f10} = \alpha^{uk}_{f10} - i_{xf}, \qquad \gamma^{uk}_{f01} = \alpha^{uk}_{f01} - i_{yf} \tag{29}$$

$$\gamma^{vk}_{f10} = \alpha^{vk}_{f10} - j_{xf}, \qquad \gamma^{vk}_{f01} = \alpha^{vk}_{f01} - j_{yf} \tag{30}$$

According to expressions (25-30), we have

$$\tilde{\boldsymbol{R}} = \boldsymbol{m}_3\boldsymbol{s}_3^T \tag{31}$$

where $\boldsymbol{m}_3$ is the third column of the matrix $\boldsymbol{M}$. The estimation of $\boldsymbol{s}_3$ from $\tilde{\boldsymbol{R}}$ and $\boldsymbol{m}_3$, according to expression (31), is a linear problem. The LS solution is

$$\hat{\boldsymbol{s}}_3^T = \left(\boldsymbol{m}_3^T\boldsymbol{m}_3\right)^{-1}\boldsymbol{m}_3^T\tilde{\boldsymbol{R}} = \frac{\boldsymbol{m}_3^T\tilde{\boldsymbol{R}}}{\boldsymbol{m}_3^T\boldsymbol{m}_3} \tag{32}$$

Motion estimate for known shape From (25-27), note that the object position at instant f only affects the rows f and $f+F$ of matrix $\boldsymbol{R}$. Given the object shape $\boldsymbol{S}$, the LS estimate of the object rotation

$\{\theta_f, \phi_f, \psi_f\}$ for each frame f is given by

$$\boldsymbol{c}_f(\theta,\phi,\psi) = \left[\begin{array}{c} \boldsymbol{r}_f \\ \boldsymbol{r}_{f+F} \end{array}\right] - \left[\begin{array}{c} \boldsymbol{S}[\ i_x \quad i_y \quad i_z\ ]^T \\ \boldsymbol{S}[\ j_x \quad j_y \quad j_z\ ]^T \end{array}\right]$$

$$\left\{\hat{\theta}_f\ \hat{\phi}_f, \hat{\psi}_f\right\} = \arg\min_{\{\theta,\phi,\psi\}} \boldsymbol{c}_f^T(\theta,\phi,\psi)\boldsymbol{c}_f(\theta,\phi,\psi) \quad (34)$$

To solve the non-linear minimization (34), we a Gauss-Newton method. Starting from an initial point $\{\theta_0, \phi_0, \psi_0\}$, the increments $\{\delta_\theta, \delta_\phi, \delta_\psi\}$ are found by minimizing (34) after truncating the Taylor series of $\boldsymbol{c}_f(\theta_0+\delta_\theta, \phi_0+\delta_\phi, \psi_0+\delta_\psi)$. Neglecting second and higher order terms, we have

$$\boldsymbol{c}_f(\theta_0+\delta_\theta, \phi_0+\delta_\phi, \psi_0+\delta_\psi) \simeq \boldsymbol{c}_f(\theta_0,\phi_0,\psi_0) + \nabla_{\boldsymbol{c}_f}(\theta_0,\phi_0,\psi_0)\left[\begin{array}{c} \delta_\theta \\ \delta_\phi \\ \delta_\psi \end{array}\right] \quad (35)$$

where ∇ is the gradient operator. Equating to zero the partial derivates of the cost function $\boldsymbol{c}_f(\theta_0+\delta_\theta, \phi_0+\delta_\phi, \psi_0+\delta_\psi)^T\boldsymbol{c}_f(\theta_0+\delta_\theta, \phi_0+\delta_\phi, \psi_0+\delta_\psi)$ with respect to the increments $\delta_\theta, \delta_\phi, \delta_\psi$, we obtain linear estimates $\hat{\delta}_\theta, \hat{\delta}_\phi, \hat{\delta}_\psi$ from the solution of

$$\nabla^T_{\boldsymbol{c}_f}(\theta_0,\phi_0,\psi_0)\nabla_{\boldsymbol{c}_f}(\theta_0,\phi_0,\psi_0)\left[\begin{array}{c} \hat{\delta}_\theta \\ \hat{\delta}_\phi \\ \hat{\delta}_\psi \end{array}\right] = -\nabla^T_{\boldsymbol{c}_f}(\theta_0,\phi_0,\psi_0)\boldsymbol{c}_f(\theta_0,\phi_0,\psi_0) \quad (36)$$

When computing $\hat{\theta}_f, \hat{\phi}_f, \hat{\psi}_f$ we start the iterative process with the initial guess $\theta_0 = \hat{\theta}_{f-1}, \phi_0 = \hat{\phi}_{f-1}, \psi_0 = \hat{\psi}_{f-1}$. Initially, we have $\theta_1 = \phi_1 = \psi_1 = 0$ by definition. See [4] for the details.

4 Experiment

We used a real life video showing a static corner taped by a moving camera. Figure 2 shows two consecutive frames from the sequence of 10 images.

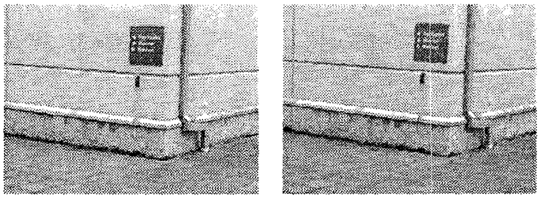

Figure 2: Image sequence.

Figure 3 shows a perspective view of the reconstructed 3D shape. It contains 3 planar patches: the floor and two walls. The angle between the walls is clearly seen. The angle of the walls with the floor can also be perceived.

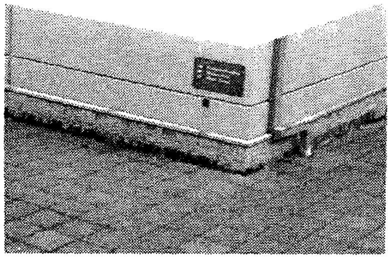

Figure 3: Reconstructed 3D shape and texture.

5 Conclusions

We recover 3D structure from 2D video. The results obtained so far show that our method can be used in content-based video analysis tasks.

References

[1] Radu S. Jasinschi and José M. F. Moura. Content-based video sequence representation. In *Proceedings of the IEEE International Conference on Image Processing*, Washington DC, U.S.A., October 1995.

[2] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[3] Pedro M. Q. Aguiar and José M. F. Moura. Robust 3D structure from motion under ortography. In *Proceedings of the Tenth IEEE Image and Multidimensional Digital Signal Processing Workshop*, Alpbach, Austria, July 1998.

[4] Pedro M. Q. Aguiar and José M. F. Moura. 3D shape and motion recovery from induced parametric optical flow models. To be submitted.

[5] James R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the Second European Conference on Computer Vision*, Santa Margherita Ligure, Italy, May 1992. Springer-Verlag.

[6] Serge Ayer. *Sequential and Competitive Methods for Estimation of Multiple Motions*. PhD thesis, École Polytechnique Féderale de Lausanne, 1995.