3D Rigid Structure from Video: What Are "Easy" Shapes and "Good" Motions ?

Pedro M. Q. Aguiar
Institute for Systems and Robotics / IST, Lisboa, Portugal
E-mail: aguiar@isr.ist.utl.pt

José M. F. Moura
Carnegie Mellon University, ECE Dep., Pittsburgh PA, USA
E-mail: moura@ece.cmu.edu

Abstract—**Factorization algorithms are increasingly popular to recover 3D rigid structure from video. In this paper, we analyze the *rank 1 factorization* algorithm to determine what are the most suitable 3D shapes or the best 3D motions to recover the 3D structure from the 2D trajectories of the features. We show that the shape is best retrieved from orthogonal views aligned with the longest and smallest axes of inertia of the object.**

I. Introduction

Virtual reality systems demand efficient and cheap ways to obtain 3D models of real world objects. Human operators can build 3D models of artificial objects in a manual way but this isn't practical for applications requiring 3D photographies of real objects. Active sensing systems such as the laser range finder are accurate but expansive. This paper addresses the automatic recovery of 3D models from ordinary digital video.

The strongest cue to recover 3D information from the 2D video is the motion induced in the image plane, so that the task of recovering 3D structure is usually referred to as *structure from motion* (SFM). Among the approaches available, the *factorization method* introduced in the early nineties [1] is common because of its robustness. It captures the rigidity of the 3D shape in an algebraic way – it recovers the 3D structure by factoring a measurement matrix that is rank 3 in a noiseless situation. In reference [2], we exploit a degree of freedom not taken advantage of before – the choice of the reference view – and present a *rank 1 factorization* algorithm to solve the rigid SFM problem. Besides being computationally simpler than the original rank 3 factorization method, our approach also succeeds where the original one fails – the limiting case when the 3D shape is close to being a planar shape.

This paper is concerned with the following question: which 3D shapes are easier and what 3D motions are better suited in the SFM problem. We address this issue through analysis of the rank 1 factorization algorithm. Our conclusions are that the "easier" (normalized) 3D shapes are those with a large extension along an arbitrary direction in the 3D space, i.e., 3D shapes with an axis of small inertia, and the "best" 3D motions are the ones that provide: i) a reference view aligned with the axis of smallest inertia and ii) remaining views orthogonal to this axis. The orthogonality between the views of the optimal 3D motion sequence is in agreement with the optimal placement of the views in classic photogrammetry where 3D shape is recovered from calibrated cameras. Since computing 2D image motion (feature correspondences) between images with such a wide baseline is a complex task, to take advantage in practice of our conclusions, we move the camera from the reference view to the orthogonal direction in a smooth way. This enables simple feature tracking approaches.

Paper organization Section II overviews the rank 1 factorization approach. In section III, we analyze the performance of the algorithm. Section IV describes experiments that illustrate our theoretical analysis and section V concludes the paper.

II. SFM: Rank 1 Factorization

We describe the object shape by the positions $\{x_n, y_n, z_n\}$ in 3D of $n = 1 \dots N$ feature points. We show in [2] that the feature coordinates along the camera plane $\{x_n, y_n\}$ are known from their projection onto the reference image, usually taken as the first frame; the 3D shape is determined simply by the unknown depth $\{z_n, n = 1 \dots N\}$. The 3D motion of the camera is represented by a set of $f = 2 \dots F$ translation-rotation pairs that code the camera pose in each of $F-1$ frames with respect to its pose in the reference frame $f = 1$. The translational component of the 3D motion is easily obtained from the centroid of the projections of the features points [1]. After compensating for the 3D translation, the only unknowns in the 3D motions are the 3D rotations at each frame.

The relation between the 2D projections of the features and the 3D structure parameters is written in matrix format as [1]

$$\boldsymbol{R} = \boldsymbol{M}\boldsymbol{S}^T, \tag{1}$$

where: $\boldsymbol{R}$ is the $2(F-1) \times N$ observation matrix that collects the 2D trajectories of the projections of the features; $\boldsymbol{M}$ is the $2(F-1) \times 3$ motion matrix that collects the 3D rotation matrices that code the unknown camera poses; and $\boldsymbol{S}$ is the $N \times 3$ shape matrix that collects the 3D coordinates of the features. The two first columns of $\boldsymbol{S}$ are the known vectors $\boldsymbol{x} = [x_1 \dots x_N]^T$ and $\boldsymbol{y} = [y_1 \dots y_N]^T$, and the third column is the vector $\boldsymbol{z} = [z_1 \dots z_N]^T$ that defines the 3D shape. The *rank 1 factorization* algorithm estimates the the $\boldsymbol{M}$ and $\boldsymbol{z}$ from $\boldsymbol{R}$ in two steps: the *decomposition stage* solves the unconstrained bilinear problem in (1); and the *normalization stage* applies the constraints imposed by the pairwise orthonormality of the rows of the motion matrix $\boldsymbol{M}$ [2].

Decomposition stage We define

$$\boldsymbol{M} = [\boldsymbol{M}_0,\ \boldsymbol{m}_3] \quad \text{and} \quad \boldsymbol{S} = [\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z}] = [\boldsymbol{S}_0,\ \boldsymbol{z}], \tag{2}$$

where the matrices $\boldsymbol{M}_0$ and $\boldsymbol{S}_0$ contain the first two columns of $\boldsymbol{M}$ and $\boldsymbol{S}$, respectively, the vector $\boldsymbol{m}_3$ is the third column of $\boldsymbol{M}$, and the vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are the columns of $\boldsymbol{S}$. Now decompose the depth vector $\boldsymbol{z}$ into the component $\boldsymbol{S}_0\boldsymbol{b}$ that belongs to the space $\mathcal{S}_0$ spanned by the columns of $\boldsymbol{S}_0$ and the component $\boldsymbol{a}$ that belongs to the orthogonal complement $\mathcal{S}_0^{\perp}$,

$$\boldsymbol{z} = \boldsymbol{S}_0\boldsymbol{b} + \boldsymbol{a}, \quad \text{with} \quad \boldsymbol{a}^T\boldsymbol{S}_0 = [0,\ 0]. \tag{3}$$

0-7803-7713-3/02/$17.00 © 2002 IEEE

The linear *least squares* (LS) solution for M_0 is, see [2],

$$\widehat{M}_0 = RS_0 \left(S_0^T S_0\right)^{-1} - m_3 b^T. \quad (4)$$

We replace $\widehat{M}_0$ in (1) and define the matrix $\widetilde{R}$ as

$$\widetilde{R} = R\Pi_{\mathcal{S}_0^\perp} = m_3 a^T, \quad (5)$$

where $\Pi_{\mathcal{S}_0^\perp}$ is the orthogonal projector onto the space $\mathcal{S}_0^\perp$.

The solution for the vectors m_3 and a in (5) is obtained from the rank 1 matrix that best approximates $\widetilde{R}$. We get

$$\widetilde{R} \simeq u\sigma v^T, \qquad \widehat{m}_3 = \alpha u, \qquad \hat{a}^T = \frac{\sigma}{\alpha} v^T, \quad (6)$$

where σ is the largest singular value of $\widetilde{R}$, u and v are the corresponding left and right singular vectors, and α is a normalizing scalar different from zero.

Normalization stage We collect the parameters α and $b = [b_1\, b_2]^T$ into the parameter vector $\alpha = [\alpha\ b^T]^T$. The constraints that come from the structure of the matrix M can be written after some algebra in matrix format as

$$\Xi\epsilon(\alpha) = \xi, \quad (7)$$

where the 3×1 vector $\epsilon(\alpha)$ is

$$\epsilon(\alpha) = \left[\alpha b^T \ \ \alpha^2(1 + \|b\|^2)\right]^T, \quad (8)$$

and the $3(F-1) \times 1$ vector ξ and the $3(F-1) \times 3$ matrix Ξ depend on the known entities R, S_0, and u.

We compute α from the LS solution of the system of $3(F-1)$ equations (7), i.e., we minimize the cost function

$$C\left[\epsilon(\alpha)\right] = \left[\Xi\epsilon(\alpha) - \xi\right]^T \left[\Xi\epsilon(\alpha) - \xi\right]. \quad (9)$$

The LS solution for the parameters is obtained by inverting the defining relations for the entries of the vector ϵ leading to

$$|\widehat{\alpha}| = \sqrt{\epsilon_{3LS} - \epsilon_{1LS}^2 - \epsilon_{2LS}^2}, \quad \widehat{b}_1 = \epsilon_{1LS}/\widehat{\alpha}, \quad \widehat{b}_2 = \epsilon_{2LS}/\widehat{\alpha}. \quad (10)$$

Clearly these solutions exist and make sense if and only if $\epsilon_{3LS} > \epsilon_{1LS}^2 - \epsilon_{2LS}^2$. We discuss in section III when this fails.

III. Analysis of the Factorization Algorithm

We study the behavior of the decomposition and normalization stages of the *rank 1 factorization* method. We start by analyzing the accuracy of the rank 1 approximation in the decomposition stage. Then, we discuss the situations that may cause the algorithm of the normalization stage to fail.

A. Influence of the 3D structure on the rank 1 approximation

The decomposition stage of section II estimates the vectors m_3 and a up to a scale parameter α by factoring the matrix $\widetilde{R}$, see expression (6). The decomposition stage determines thus the 1D linear subspaces where the estimates of the motion vector m_3 and shape vector a live. We study how the 3D rigid shape and 3D motion impact the estimation error of these subspaces. In particular, we ask the question "given a 3D shape, what is the best 3D motion, i.e., what is the sequence of 3D positions that leads to the smaller estimation error for the shape and motion subspaces?"

The error in estimating these subspaces depends inversely on the ratio between the singular value of the deterministic component of $\widetilde{R}$ and the singular value of its noise component. Assuming we have no control over the noise, which comes from the noisy output of the feature tracking algorithms, the error in estimating the subspaces decreases with the singular value of the deterministic component of $\widetilde{R} = m_3 a^T$, which equals the product $\|m_3\|\|a\|$. We now investigate the issue of maximizing $\|m_3\|$ and $\|a\|$ for a given 3D rigid shape by manipulating the relative motion between the camera and the object. We assume that the object is stationary, only the camera moves.

Maximizing $\|m_3\|$ The third column m_3 of the motion matrix M collects the entries i_{zf} and j_{zf} of the rotation matrices that orient the camera coordinate system relative to the object coordinate system, i.e., it collects the z-component of orthonormal pairs $\{i_f, j_f\}$,

$$m_3 = \left[\ i_{z2} \quad j_{z2} \quad i_{z3} \quad j_{z3} \quad \cdots \quad \cdots \quad i_{zF} \quad j_{zF}\ \right]^T. \quad (11)$$

Each pair of entries $\{i_{zf}, j_{zf}\}$ is constrained by

$$2 \le f \le F: \qquad i_{zf}^2 + j_{zf}^2 \le 1 \quad (12)$$

because the vectors i_f and j_f are orthogonal and have unit norm. The equality in (12) occurs when the z-component k_{zf} of the third vector k_f of the orthonormal reference system $\{i_f, j_f, k_f\}$ is zero, i.e., when the optical axis of the camera is perpendicular to the z-axis. Since we made the coordinate systems to coincide in the first frame, the condition $k_{zf} = 0$ means that the camera in frame f points in a direction that is perpendicular to the direction it pointed in frame 1. The magnitude of the vector m_3 in (11) is then maximized by making

$$2 \le f \le F: \qquad k_{zf} = 0, \quad (13)$$

i.e., by pointing the camera in instants $f = 2 \ldots F$ in directions perpendicular to the direction of the optical axis in the reference frame $f = 1$. This is intuitively pleasing: the unknown z-coordinates of the feature points are most accurately estimated from their projections onto planes that are parallel to the z-axis, i.e., planes that are orthogonal to the image plane in the reference view. Further, since the analysis did not restrict in any way the 3D shape of the object, we conclude that the optimal position of the camera for all frames after frame 1 does not depend on the particular object shape.

Maximizing $\|a\|$ The vector a, see (3), is the component of the relative depth vector z that belongs to the orthogonal complement $\mathcal{S}_0^\perp$ of the space spanned by the vectors x and y, i.e.,

$$a = \Pi_{\mathcal{S}_0^\perp} z. \quad (14)$$

The magnitude of a is

$$\|a\|^2 = \|\Pi_{\mathcal{S}_0^\perp} z\|^2 = \|z\|^2 - z^T S_0 \left(S_0^T S_0\right)^{-1} S_0^T z. \quad (15)$$

Since the last term of (15) is non-negative, the magnitude $\|\boldsymbol{a}\|$ increases with the magnitude $\|\boldsymbol{z}\|$ of the relative depth vector $\boldsymbol{z}$ and with the degree of orthogonality between the vector $\boldsymbol{z}$ and the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in matrix $\boldsymbol{S}_0$. The choice of the first view, the reference view, affects the magnitude $\|\boldsymbol{a}\|$ in (15) because it determines the object coordinate system and so, affects $\boldsymbol{S}_0$ and $\boldsymbol{z}$ in the shape matrix $\boldsymbol{S}$.

To make explicit the influence of the reference view on the shape matrix $\boldsymbol{S}$, we start with the SVD of $\boldsymbol{S}$

$$\boldsymbol{S} = \boldsymbol{U}_S \boldsymbol{\Sigma}_S \boldsymbol{V}_S^T. \tag{16}$$

The freedom of choice of the reference view enables us to work with a shape matrix $\boldsymbol{S}^*$ that is obtained from $\boldsymbol{S}$ by rigidly rotating all the 3D feature points,

$$\boldsymbol{S}^* = \boldsymbol{S}\boldsymbol{\Theta}, \tag{17}$$

where the rotation matrix $\boldsymbol{\Theta}$ appears on the right side of $\boldsymbol{S}$ because $\boldsymbol{\Theta}$ acts on the rows of $\boldsymbol{S}$. Since $\boldsymbol{\Theta}$ is a unitary matrix, we use (16) to express the SVD of $\boldsymbol{S}^*$ as

$$\boldsymbol{S}^* = \boldsymbol{U}_S \boldsymbol{\Sigma}_S \boldsymbol{V}_S^T \boldsymbol{\Theta}. \tag{18}$$

The magnitude $\|\boldsymbol{a}\|$ in (15) is maximized when the third column $\boldsymbol{z}$ of $\boldsymbol{S}^*$ is orthogonal to the first two and its norm $\|\boldsymbol{z}\|$ is the largest possible. Since the columns of $\boldsymbol{U}_S$ in (16) and (18) are orthonormal vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$, the choice of $\boldsymbol{\Theta}$ must be such that $\boldsymbol{z} = \sigma_{i\max} \boldsymbol{u}_{i\max}$ where $\sigma_{i\max}$ is the largest singular value in $\boldsymbol{\Sigma}_S$ in (16) and (18), and $\boldsymbol{u}_{i\max}$ the corresponding singular vector. Assuming that the singular values in $\boldsymbol{\Sigma}_S$ are nondecreasingly ordered, an optimal $\boldsymbol{\Theta}$ is such that $\boldsymbol{V}_S^T \boldsymbol{\Theta} = \boldsymbol{I}$. In this case $\sigma_{i\max} = \sigma_3$. Since $\boldsymbol{V}_S$ is unitary, an optimal solution for the rotation matrix $\boldsymbol{\Theta}$ is then

$$\boldsymbol{\Theta} = \boldsymbol{V}_S. \tag{19}$$

This solution is not unique. The condition $\boldsymbol{z} = \sigma_{i\max} \boldsymbol{u}_{i\max} = \sigma_3 \boldsymbol{u}_3$ restricts only two of the three degrees of freedom of $\boldsymbol{\Theta}$. The third degree of freedom, a rotation between the vectors $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ and $\{\boldsymbol{x}, \boldsymbol{y}\}$ does not affect the magnitude $\|\boldsymbol{a}\|$ in (15), $\|\boldsymbol{a}\| = \|\boldsymbol{z}\| = \sigma_3^2$.

With the optimal rotation matrix $\boldsymbol{\Theta}$ in (19), the shape matrix $\boldsymbol{S}^*$ is simply given by

$$\boldsymbol{S}^* = \boldsymbol{U}_S \boldsymbol{\Sigma}_S = \begin{bmatrix} \sigma_1 \boldsymbol{u}_1 & \sigma_2 \boldsymbol{u}_2 & \sigma_3 \boldsymbol{u}_3 \end{bmatrix}, \tag{20}$$

i.e., the optimal choice for the reference view corresponds to aligning the camera optical axis, the z-axis, with the object axis of smallest inertia (in this case, the inertial moment wrt the z-axis is given by $\sigma_1^2 + \sigma_2^2$).

In what respects to the object shape, we conclude that the larger is the largest singular value $\sigma_{i\max}$ of the 3D shape matrix $\boldsymbol{S}$, the more accurate is the rank 1 subspace estimation, i.e., the "easier" 3D shapes are those with a large extension along an arbitrary direction in the 3D space.

B. *Normalization failure*

If $\boldsymbol{\epsilon}_{\text{LS}} = [\epsilon_{1\text{LS}}, \epsilon_{2\text{LS}}, \epsilon_{3\text{LS}}]^T$ determined by the LS solution of the system (7), see section II, is such that

$$\epsilon_{3\text{LS}} < \epsilon_{1\text{LS}}^2 + \epsilon_{2\text{LS}}^2, \tag{21}$$

we cannot determine the scalar α and the vector $\boldsymbol{b} = [b_1, b_2]^T$ from (10), and the gradient of the cost function $C[\boldsymbol{\epsilon}(\boldsymbol{\alpha})]$ in (9) is nonzero over the whole space where $\boldsymbol{\alpha}$ lives. This is a *normalization failure*, see also [1] where the normalization stage computes a normalization matrix $\boldsymbol{A}$ by factoring the estimate $\widetilde{\boldsymbol{B}}$ of an intermediate matrix $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^T$ that may fail to be nonnegative definite.

Since the cost function C in (9) is strictly nondecreasing and grows unbounded with $\|\boldsymbol{\alpha}\|$ (see the definition of $\boldsymbol{\epsilon}$ in (8)), the minimum of C with respect to $\boldsymbol{\alpha}$ occurs at the boundary, i.e., in the limit when α goes to zero. At the boundary, $\alpha = 0$, we have $\boldsymbol{\epsilon} = \boldsymbol{0}$, see (8), thus, from (9), the minimum value of the cost function C approaches $\boldsymbol{\xi}^T \boldsymbol{\xi}$. This is much larger than the small value for the minimum of C that we expect to obtain at the true value of the normalization parameter vector $\boldsymbol{\alpha}$. This indicates that the two-stage algorithm *decomposition-normalization* does not work and the matrix $\widetilde{\boldsymbol{R}}$ in (5,6) is not well approximated by a rank 1 matrix.

The matrix $\widetilde{\boldsymbol{R}}$ is not well approximated by a rank 1 matrix if its rank is either larger or smaller than 1. The first situation arises when the scene contains dramatic perspective effects that can be resolved only by taking into account in the analysis the perspective projection. The second situation occurs when the 3D shape of the object or its 3D motion cause $\widetilde{\boldsymbol{R}} = \boldsymbol{0}$ in a noiseless situation. From expression (5), this happens in either of the two degenerate cases: the 3D motion is such that the third column of the matrix $\boldsymbol{M}$ is $\boldsymbol{m}_3 = \boldsymbol{0}$; or $\boldsymbol{a} = \boldsymbol{0}$ as when the 3D shape is planar, see expression (3), and the relative depth vector $\boldsymbol{z}$ lives in the space $\mathcal{S}_0$, i.e., $\boldsymbol{z} = \boldsymbol{S}_0 \boldsymbol{b}$. In either case, there is not enough information in the feature trajectories to recover the 3D structure. In spite of this, the images in the sequence can still be aligned by computing $\widehat{\boldsymbol{M}}_0$ according to expression (4), for example, by making $\widehat{\boldsymbol{M}}_0 = \boldsymbol{R}\boldsymbol{S}_0(\boldsymbol{S}_0^T \boldsymbol{S}_0)^{-1}$.

IV. Experiments

We describe three experiments that illustrate the predictions of section III with respect to the camera trajectory and the reference frame viewing angle. The first two experiments use a set of 50 features sampled from the surface of the synthetic object shown in Figure 1. The third experiment uses a real video clip.

Camera trajectory Maintaining a fixed reference frame, we created several trajectories by moving the camera around the object. To make clearer the comparison, the angle θ between the reference and the successive views was chosen to be constant for each trajectory. The left plot in Figure 2, computed from the ground truth, represents the norm $\|\boldsymbol{m}_3\|$ of the motion vector $\boldsymbol{m}_3$ as a function of the angle θ. As expected from the analysis of section III, $\|\boldsymbol{m}_3\|$ is maximum when each other view is

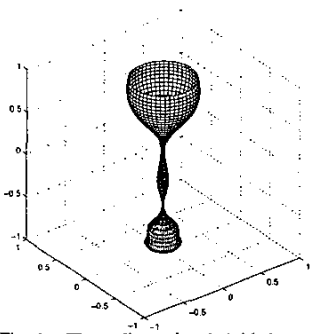

Fig. 1. Three-dimensional rigid shape.

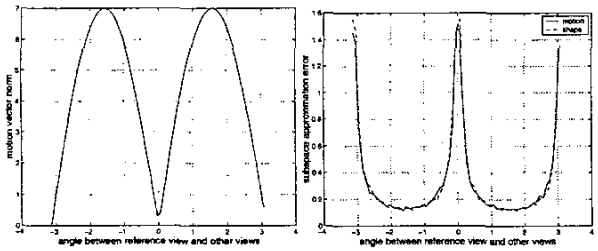

Fig. 2. $\|m_3\|$ and subspace estimation errors as functions of the angle θ of the camera pose.

orthogonal to the reference view, i.e., when $\theta = \pi/2 + k\pi$, and $\|m_3\| = 0$ when they are parallel, i.e., when $\theta = k\pi$.

Using these trajectories, we synthesized noisy feature projections and applied the rank 1 factorization. The right plot in Figure 2 plots the estimation error for the shape and motion subspaces as function of the angle θ. These errors are the angles[1] between the ground truth subspaces and the subspaces recovered by the rank 1 factorization We see that, as predicted by the analysis in section III, the errors are smaller when the views are close to orthogonal ($\theta = \pi/2 + k\pi$) and larger when they are close to parallel ($\theta = k\pi$).

Reference view We then fixed the camera trajectory and used several reference frame viewing angles. Again to make simpler the analysis, we chose the optical axis of the reference frame to be always vertical with respect to the orientation of the object in Figure 1 and varied the elevation angle ϕ, $\phi = 0$ corresponding to the top view. The plots in Figure 3 represent, respectively, the norm $\|a\|$ of the shape vector a, and the subspace estimation errors, as functions of the angle ϕ. Again, these plots confirm the predictions of section III – $\|a\|$ is larger, and the errors smaller, when the reference view is aligned with the axis of smallest inertia, i.e., when $\phi = k\pi$.

Real video Figure 4 shows frames 1 and 50 from the CMU's hotel video sequence. On the left image, taken as the reference frame, we marked with white squares the 50 feature points used by the *rank 1 factorization* algorithm. In this video sequence, the camera undergoes a slow rotation around the object.

To illustrate the relevance of the view selection, we run our algorithm with two distinct sets of 12 frames. In the first experiment, we used consecutive frames, thus all the views had very similar orientation. The reconstructed 3D shape is shown on the left side of Figure 5. In the second experiment, we selected one frame each eight video frames, thus the orientation of the last view was very distinct from the first one (although not orthogonal). The better quality of the 3D reconstruction obtained with this sparse view selection, shown on the right side plot of Figure 5, confirms our theoretical analysis.

[1]The angle between the 1D subspaces spanned by vectors s_1 and s_2 is the angle between those vectors, $\arccos\{s_1^T s_2/(\|s_1\|\,\|s_2\|)\}$.

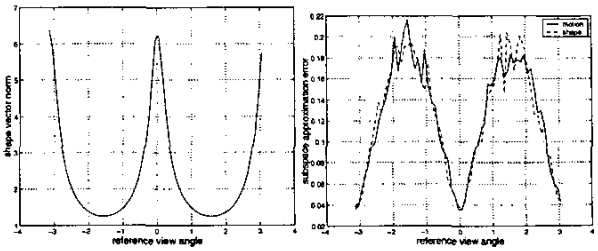

Fig. 3. $\|a\|$ and subspace estimation errors as functions of the reference view angle ϕ.

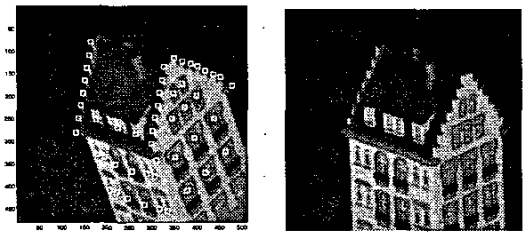

Fig. 4. Two frames from the hotel sequence.

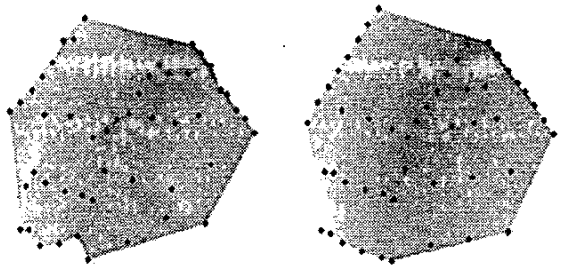

Fig. 5. 3D shape recovered from the hotel sequence.

V. Conclusion

We analyzed the rank 1 factorization algorithm in terms of the influence of the 3D motion and the 3D shape on the accuracy of the 3D recontruction. We concluded that the 3D structure is best retrieved when the reference view is aligned with the smallest axis of inertia and the other views are orthogonal to this direction.

References

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int. Journal of Computer Vision*, vol. 9, no. 2, 1992.

[2] P. M. Q. Aguiar and J. M. F. Moura, "Factorization as a rank 1 problem," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Fort Collins, CO, USA, 1999.