IEEE Signal Processing Society Workshop on Multimedia Signal Processing
September 1999, Copenhagen, Denmark
© 1999 IEEE

FAST 3D MODELING FROM VIDEO

Pedro M. Q. Aguiar [*]
Instituto de Sistemas e Robótica
IST, Lisboa, Portugal
aguiar@isr.ist.utl.pt

José M. F. Moura
Electrical and Computer Eng. Dep.
CMU, Pittsburgh PA, USA
moura@ece.cmu.edu

Abstract - In this paper we build 3D models of rigid bodies from video sequences. The algorithm we use is simple and robust. It recovers the 3D shape parameters and the 3D motion parameters by first estimating the parameters of the induced optical flow representation. To estimate the 3D shape and 3D motion from the optical flow, we use a fast algorithm that is based on the factorization of a matrix that is rank 1 in a noiseless situation. We demonstrate our approach with a piecewise planar object shape built from a real life video clip. We highlight some of the potential applications of the 3D models obtained.

INTRODUCTION

This paper presents a fast reliable algorithm to recover three-dimensional (3D) models from monocular video sequences. Just like an image is worth ten thousand words, and video enhances tremendously our visual perception of the environment, 3D represents the next higher level in recreating a natural immersive multimedia environment for participants to interact collaboratively, and/or viewers to enjoy. The volume experience enables users to perceive differently the same scene from their own vantage point if view. Our challenge is to recover in an expedite way the 3D structure from a monocular video sequence.

Structure from Motion Obtaining 3D structure from video is a problem with a long tradition in computer vision, e.g., depth from stereo and depth from motion. Our fall under the usual heading of structure from motion problem. In this problem, usually authors extract a set of features, say corners of an object or of a building, and establish the correspondence of these features across the video sequence.

Early approaches to structure from motion processed a single pair of consecutive frames and provided existence and uniqueness results to the problem of estimating 3D motion and absolute depth from the 2D motion in the camera plane between two frames, see for example [1]. The two-frame algorithms are highly sensitive to image noise and, when the object is far from the camera, i.e., at a large distance when compared to the object depth, they fail

[*] The work of the first author was partially supported by INVOTAN.

even at low level image noise. More recent research has been oriented toward the use of longer image sequences.

Factorization Aided by constraints like the rigid body assumption, the inverse-problem of inferring 3D structure from the 2D motions of these features is then achieved. Among the existent approaches, the factorization method [2] is an elegant method to recover structure from motion without computing the absolute depth as an intermediate step. In [2] they treat orthographic projections. The object shape is represented by the 3D position of a set of feature points. The 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. This work was later extended to scaled-orthography and paraperspective projections [3]. Tracking and solving for the correspondence of these features is a computationally expensive problem. To reduce this cost, the number of features is usually small. In turn, this leads to sparse representations.

We avoid this quagmire by assuming parametric representations for the 3D surfaces to be reconstructed. These 3D parametric representations induce parametric time varying representations for the imaged surface across the video sequence. Rather than having a difficult correspondence problem to solve, our framework replaces this high order combinatorial problem by a low dimension parameter estimation problem. With our approach, we reduce the structure from motion problem to solving a rank one matrix factorization problem, for which we develop a fast algorithm.

Paper organization We start by summarizing the structure from motion approach. Then, we describe an experiment with real video. Finally, we illustrate some of the potential applications of the recovered 3D models.

APPROACH

The tracking of feature points may be unreliable when processing noisy video sequences. In [4] we extend the factorization method by using a robust region-based approach. We assume that the 3D shape of the rigid body is well described by a parametric representation and use orthographic projections. The parametric representation of the 3D rigid shape induces a parametric model for the optical flow. We use this optical flow parameterization to derive a two-stage algorithm to recover structure from motion: the first stage estimates the optical flow parameters; the second recovers the 3D shape and 3D motion parameters from the sequence of estimates of the optical flow parameters. We apply known techniques to estimate the optical flow parameters. To recover the 3D structure parameters from the optical flow parameters, we use Least Squares (LS): the 3D translation parameters are obtained in closed form, while the 3D rotation parameters and the 3D shape parameters are the solution of a nonlinear LS problem. Rather than attempting a direct nonlinear minimization, we show that the problem has a bilinear structure. Then, we solve the bilinear LS problem by factorizing a measurement matrix that is rank 1 in a noiseless situation, see [5]. Our approach handles general shaped structures. It is well suited to the analysis of scenes with polyhedral surfaces, where the optical flow model reduces to the well known affine motion model. We summarize the approach. For details, see [4, 5].

3D Shape The shape $\mathcal{S}$ of the rigid object is a parametric description of the object surface. We consider objects whose shape is given by a piecewise planar surface with K patches. The shape parameter vector $\boldsymbol{a}$ collects the coefficients of these polynomials, i.e., $\boldsymbol{a} = \left\{a_{00}^k, a_{10}^k, a_{01}^k, 1 \leq k \leq K\right\}$ where $z = a_{00}^k + a_{10}^k x + a_{01}^k y$ describes the shape of the patch k in the object coordinate system.

3D Motion We define the 3D motion of the object by specifying the position of the object coordinate system relative to the camera coordinate system. The parameters $\left(t_{u\,f}, t_{v\,f}, t_{w\,f}\right)$ are the coordinates of the origin of the object coordinate system (3D translation) and $(\theta_f, \phi_f, \psi_f)$ are the Euler angles that determine the orientation of the object coordinate system (3D rotation).

Optical flow In [4], we show that the optical flow between the frames $\boldsymbol{I}_1$ and $\boldsymbol{I}_f$ in the region corresponding to surface patch k is expressed in terms of a set of optical flow parameters. For planar patches, we get the affine motion model for the optical flow. We use known numerical techniques to estimate the optical flow parameters, see [6].

3D Structure from optical flow The optical flow parameters are directly related to the 3D shape and 3D motion parameters. This relation leads to a set of equations that define an overconstrained system with respect to the 3D shape parameters $\left\{a_{00}^k, a_{10}^k, a_{01}^k, 1 \leq k \leq K\right\}$ and to the 3D positions $\left\{t_{u\,f}, t_{v\,f}, \theta_f, \phi_f, \psi_f, 1, \leq f, \leq F\right\}$ (under orthography, the component of the translation along the camera axis, $t_{w\,f}$, can not be recovered). The problem of inferring structure from motion is formulated as the LS solution of the system. First, we solve for the translation parameters which leads to a closed-form solution. Then, replace the translation estimates and solve for the remaining motion parameters and shape parameters by factorizing a measurement matrix.

Rank 1 factorization After replacing the translation estimates, the relation between the optical flow parameters and the 3D structure parameters is written in matrix format as

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

where $\boldsymbol{R}$, $(2(F-1) \times 3K)$, collects the optical flow parameters and $\boldsymbol{M}$, $(2(F-1) \times 3)$, and $\boldsymbol{S}$, $(3K \times 3)$, collect the 3D motion and 3D shape parameters, respectively, see [4]. The matrix of optical flow parameters $\boldsymbol{R}$ is $2(F-1) \times 3K$

but it is rank deficient. In a noiseless situation, $\boldsymbol{R}$ is rank 3 reflecting the high redundancy in the data, due to the rigidity of the object. In [5] we show how to compute the 3D shape and 3D motion parameters from the optical flow parameters by factorizing a modified matrix $\widetilde{\boldsymbol{R}}$. The matrix $\widetilde{\boldsymbol{R}}$ is $\boldsymbol{R}$ multiplied by the orthogonal projector onto the orthogonal complement of the space spanned by the first two columns of $\boldsymbol{S}$. We get

$$\widetilde{\boldsymbol{R}} = \boldsymbol{m}_3 \boldsymbol{a}_1^T \tag{2}$$

where $\boldsymbol{m}_3$ is the third column of $\boldsymbol{M}$, and $\boldsymbol{a}_1$ is a vector that collects the 3D shape parameters. The 3D shape and 3D motion parameters are computed by factorizing the rank 1 matrix $\widetilde{\boldsymbol{R}}$. See [5] for the details. In reference [7] we illustrate the properties of the rank 1 matrix when the original matrix $\boldsymbol{R}$ collects trajectories of feature points.

EXPERIMENT

We used a hand hold taped video sequence of 50 frames showing a box over a carpet. The camera motion was approximately a rotation around the box. We processed the box video sequence by using the method described above. Figure 1 shows two perspective views of the reconstructed 3D shape and texture. The 3D model is described in terms of four planar patches. One corresponds to the floor, and the other three correspond to the three visible faces of the box. We see that the angles between the planar patches are correctly recovered.

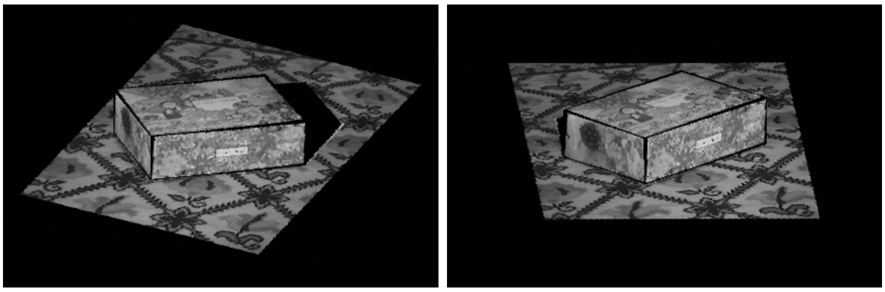

Figure 1: Two perspective views of the reconstructed 3D shape and texture.

APPLICATIONS

Virtualized reality The 3D models obtained from the video data can be used to build a synthetic image sequence. This synthesis is achieved by specifying the sequence of viewing positions along time. The viewing positions are specified by the user, either in a interactive way or from an automatic procedure. For example, the images in figure 1 were obtained by rotating the recovered 3D model. Other views are generated in a similar way. Synthetic images are obtained by selecting from these views a rectangular window, corresponding to the camera field of view. This is an example of virtual manipulation of real objects. More complex scenes are obtained by merging real objects with virtual entities.

Video coding Model-based video representations enable very low bit rate compression. Basically, instead of representing a video sequence in terms of frames and pixels, we use the recovered 3D structure. A video sequence is represented by the 3D shape and texture of the object, and its 3D motion. Since the 3D motion and 3D shape are coded with a few parameters, the number of bytes necessary to code the entire sequence is governed by the size of the object texture representation. The texture is coded as a set of ordinary images, one for each planar patch. By using this model-based representation, we reduce dramatically the storage space because we code only once the brightness values, as opposed to the redundancy of coding the brightness values at each of the frames of the original sequence.

In figure 2 we illustrate this compression scheme with the box video sequence. The original sequence has $50 \times 320 \times 240 = 3840000$ bytes. The representation based on the 3D model needs $\sum_i T_i + \sum_i S_i + 50 \times M = \sum_i T_i + 2248$ bytes, where T_i is the storage size of the texture of patch i, S_i is the storage size of the shape of patch i, and M is the storage size of each camera position. We used the JPEG standard to compress texture of each surface patch. Since the temporal redundancy was eliminated, the compression ratio chosen for the spatial conversion governs the overall video compression ratio. The first frame of the original box video sequence is on the left side of figure 2. The center and right images show the first frame of the synthesized sequence for two different spatial compression ratios. In the center image, we used an higher spatial compression ratio, leading to the overall video compression ratio of 575:1. The right image corresponds to an overall video compression ratio of 317:1. We can see that the overall quality is good but there are small artifacts in the boundaries of the surface patches.

Video content addressing Content-based addressing is an important application of the 3D model-based video representation. Current systems that provide content-based access work by first segmenting the video in a sequence of shots and then labeling each shot with a distinctive indexing feature. The most common features used are image-based features, such as color histograms or image moments. By using 3D models we improve both the temporal segmentation and the indexing. The temporal segmentation can account for the 3D content of the scene. Indexing by 3D features, directly related to the 3D shape, enable queries by object similarity. See [8] for illustrative examples of the use of 3D models in video processing.

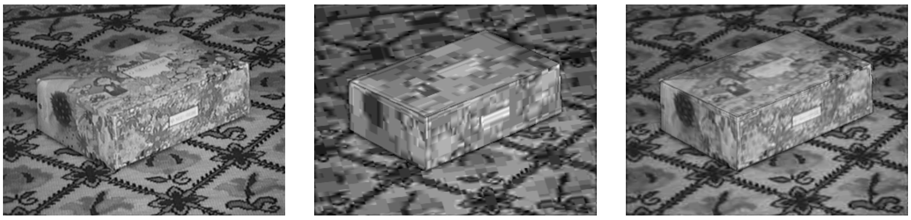

Figure 2: Video compression. Left: frame 1 of the box video sequence, Center: frame 1 of the synthesized sequence for a compression ratio of 575:1, Center: frame 1 of the synthesized sequence coded for a compression ratio of 317:1.

CONCLUSION

We recover 3D rigid models from 2D video. The experimental results obtained so far illustrate the performance of the method used. We highlight potential applications of the 3D models.

References

[1] Roger Y. Tsai and Thomas S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Transactions on PAMI*, 6(1):13–27, 1984.

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

[3] Conrad Poelman and Takeo Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE PAMI*, 19(3), 1997.

[4] Pedro M. Q. Aguiar and José M. F. Moura. Video representation via 3D shaped mosaics. In *IEEE ICIP*, Chicago, USA, October 1998.

[5] Pedro M. Q. Aguiar and José M. F. Moura. A fast algrithm for rigid structure from motion. In *IEEE ICIP*, Kobe, Japan, October 1999.

[6] J. Bergen et al. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision*, Italy, 1992.

[7] Pedro M. Q. Aguiar and José M. F. Moura. Factorization as a rank 1 problem. In *IEEE CVPR*, Fort Collins, USA, June 1999.

[8] Fernando C. M. Martins and José M. F. Moura. Video representation with three-dimensional entities. *IEEE Journal on Selected Areas in Communications*, 16(1), 1998.