Physics-Based Tracking of 3D Objects in 2D Image Sequences

Michael Chan[1], Dimitri Metaxas[1] and Sven Dickinson[2]
[1]Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.
[2]Dept. of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4.

Abstract

We present a new technique for tracking 3D objects in 2D image sequences. We assume that objects are constructed from a class of volumetric part primitives. The models are initially recovered using a qualitative shape recovery process. We subsequently track the objects using local forces computed from image potentials. Therefore we avoid the expensive computation of image features. By integrating measurements from stereo images, 3D positions (as well as other model parameters) of the objects can be continuously updated using an extended Kalman filter. Our model-based approach can handle occlusions in scenes with multiple moving objects by predicting their occurrences. To handle severe or unexpected occlusion we use a feedback mechanism between the quantitative and qualitative shape estimation systems. We demonstrate our technique in experiments involving image sequences from complex motions of objects.

1 Introduction

Various approaches in 3D model-based object tracking which attempt to recover the translation and rotation of an object have been studied [5, 10, 7]. These techniques require exact geometric specification of the objects. Recently, deformable models have been adopted for simultaneous estimation of the shape and motion of 3D objects from visual data [8, 13, 12]. The 2D problem has received similar attention [9, 1].

In this paper, we develop a new approach to tracking shapes and motions of objects in 3D from 2D image sequences. Our method makes use of both the framework for qualitative shape segmentation [4] and the physics-based framework for quantitative shape and motion estimation [12]. Assuming that objects are constructed from a finite set of volumetric part primitives, we initialize the models in the first frame of the sequence based on a shape recovery process that uses recovered qualitative shapes to constrain the fitting of deformable models to the data [11]. The qualitative shape recovery process can be avoided in successive frames, in favor of the faster physics-based model updating process requiring only a gradient computation in each frame. Assuming relatively small object motion between frames, local forces derived from stereo images are sufficient to update the positions, orientations, and shapes of the models in 3D. In fact, an advantage of our technique is that we do not need to perform costly feature correspondences during 3D tracking.

Kalman filtering techniques have been applied in the vision literature for the estimation of dynamic features [3] and rigid motion parameters [5, 2] of objects from image sequences. We use a Kalman filter for the estimation of the object's shape and motion, which consequently allows the prediction of possible edge occlusion and disocclusion. The occurrence of such situations may be due to changes of an object's aspect or due to motions of other independently moving objects. With our model-based approach, we can handle these situations by predicting their occurrence. Thus we can confidently determine which part of an object will be occluded and suppress their contributions to the net forces applied to the model. Furthermore, in case of severe or unexpected object occlusion we use a feedback mechanism between the quantitative and qualitative shape estimation techniques.

2 Dynamic deformable models

This section reviews the formulation of the deformable models we adopted for object modeling and the physics-based framework of visual estimation.

2.1 Geometry of deformable models

The positions of points on the model relative to a world coordinate frame of reference Φ are $\mathbf{x}(\mathrm{u},t) = (x(\mathrm{u},t), y(\mathrm{u},t), z(\mathrm{u},t))^T$, where u are the model's material coordinates. We express the position of a point as $\mathbf{x} = \mathbf{c} + \mathbf{R}\mathbf{p}$, where $\mathbf{c}(t)$ is the origin of a model reference frame ϕ located at the center of the model, $\mathbf{R}(t)$ is the rotation matrix that gives the orientation of ϕ relative to Φ, and $\mathbf{p}(\mathrm{u},t)$ gives the positions of points on the model relative to ϕ. We further write $\mathbf{p} = \mathbf{s} + \mathbf{d}$, as the sum of a reference shape $\mathbf{s}(\mathrm{u},t)$ and a displacement $\mathbf{d}(\mathrm{u},t)$. We express the reference shape as $\mathbf{s} = \mathbf{T}(\mathbf{e}(\mathrm{u};\, a_0, a_1, \ldots); b_0, b_1, \ldots)$, where $\mathbf{T}$ defines a *global deformation* (depending on the

parameters $b_i(t)$), which transforms a geometric primitive e defined parametrically in u and parameterized by the variables $a_i(t)$. We concatenate the global deformation parameters into the vector $\mathbf{q}_s = (a_0, a_1, \ldots, b_0, b_1, \ldots)^T$.

To illustrate our approach in this paper, we will use as a reference shape a deformable superquadric ellipsoid that can also undergo parameterized tapering deformations, as defined in [12].

2.2 Model kinematics and dynamics

The velocity of a 3D point on the model is given by

$$\dot{\mathbf{x}} = \mathbf{L}\dot{\mathbf{q}}, \tag{1}$$

where $\mathbf{L}$ is the Jacobian matrix that converts q-dimensional vectors to 3D vectors [12]. The vector $\mathbf{q}(t)$ represents the generalized coordinates of the model consisting of the translation, rotation, global and local deformations. We make the model dynamic using the Lagrangian formulation and we obtain 2nd order equations of motion which take the form (see [14] for derivations):

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{g}_q + \mathbf{f}_q, \quad \mathbf{f}_q = \int \mathbf{L}^T \mathbf{f}\, d\mathbf{u}, \tag{2}$$

where $\mathbf{f}_q$ are generalized external forces associated with the components of $\mathbf{q}$, and $\mathbf{f}(\mathbf{u}, t)$ is the image force distribution applied to the model. Here $\mathbf{M}$ is the mass matrix, $\mathbf{D}$ is the damping matrix, $\mathbf{K}$ is the stiffness matrix and $\mathbf{g}_q$ is the vector of the generalized coriolis and centrifugal forces.

3 Physics-based tracking

In this section, we present our method for tracking in stereo image sequences.

3.1 Qualitative shape recovery and model initialization

We employ the methodology developed in [11] to initialize our deformable models. We start by assuming that objects are constructed from a finite set of volumetric part primitives [4]. The parts, in turn, are mapped to a set of viewer-centered aspects. During the qualitative shape recovery process, the system first segments the image into parts using an aspect matching paradigm. Each recovered qualitative part defines: 1) the relevant non-occluded contour data belonging to the part, 2) a mapping between the image faces in their projected aspects and the 3D surfaces on the quantitative models, and 3) a qualitative orientation that is exploited during model fitting. Based on these constraints, we assign forces from image data points to the corresponding points on the 3D model. The model is then fitted dynamically to the image data under the influence of the image forces.

3.2 Local forces from image potentials

For each successive frame in the image sequence, we create an image potential such that the "valleys" of this potential correspond to the locations in the image where there are sharp changes in intensity or edge features. If we denote the intensity image by $I(x, y)$, the image potential can be computed as follows [14]:

$$\Pi(x, y) = -\beta \left| \nabla (G_\sigma * I)(x, y) \right| \tag{3}$$

where σ determines the width of the Gaussian function G_σ, $*$ denotes the convolution operation, and β determines the "steepness" of the potential surface. This potential induces a 2D force field given by:

$$\mathbf{f}(x, y) = -\nabla \Pi(x, y). \tag{4}$$

The model's degrees of freedom respond to the 2D force field through a process which first projects the model's nodes into the image. As the projected nodes are attracted to the valleys of the potential surface, the model's degrees of freedom are updated to reflect this motion. The mapping of 2D image forces to generalized forces acting on the model requires the derivation a Jacobian matrix.

3.3 Jacobian computation

To allow shape and pose estimation in a world coordinate frame from images taken from a camera with a different frame of reference, the Jacobian matrix $\mathbf{L}$ used in (2) needs to be modified appropriately.

Let $\mathbf{x} = (x, y, z)^T$ denote the location of a point j w.r.t the world coordinate frame. Then we can write

$$\mathbf{x} = \mathbf{c}_c + \mathbf{R}_c \mathbf{x}_c, \tag{5}$$

where $\mathbf{c}_c$ and $\mathbf{R}_c$ are the translation and rotation of the camera frame w.r.t. the world coordinate frame, respectively, and $\mathbf{x}_c = (x_c, y_c, z_c)^T$ is the position of the point j w.r.t to the camera coordinate frame.

Under perspective projection, the point $\mathbf{x}_c$ projects into an image point $\mathbf{x}_p = (x_p, x_p)^T$ according to

$$x_p = \frac{x_c}{z_c} f, \quad y_p = \frac{y_c}{z_c} f, \tag{6}$$

where f is the focal length of the camera. By taking the time derivative of (6) we get $\dot{\mathbf{x}}_p = \mathbf{N}\dot{\mathbf{x}}_c$ where

$$\mathbf{N} = \begin{bmatrix} f/z_c & 0 & -x_c/z_c^2 f \\ 0 & f/z_c & -y_c/z_c^2 f \end{bmatrix}. \tag{7}$$

Based on (1), (5) and (7), we obtain

$$\mathbf{x}_p = \mathbf{N}(\mathbf{R}_c^{-1}\dot{\mathbf{x}}) = \mathbf{N}\mathbf{R}_c^{-1}(\mathbf{L}\dot{\mathbf{q}}) = \mathbf{L}_p\dot{\mathbf{q}}. \tag{8}$$

By replacing the Jacobian matrix $\mathbf{L}$ in (2) by $\mathbf{L}_p = \mathbf{N}\mathbf{R}_c^{-1}\mathbf{L}$, two dimensional image forces $\mathbf{f}$ can be appropriately converted into generalized forces $\mathbf{f}_q$ measured in the world coordinate frame.

3.4 Forces from stereo images

By computing generalized forces in the world coordinate frame, the 2D image forces in a pair of stereo images can be simultaneously transformed into generalized forces $\mathbf{f}_q$ measured in a common world coordinate frame. Measurements from two different views are sufficient to determine the scale and depth parameters of the model. If we denote the position of the jth active node on the model surface by $\mathbf{x}_j$, then the generalized force on the model can be computed by replacing the integral in (2) by the summation

$$\begin{aligned}\mathbf{f}_q &= \sum_{j\in\mathcal{A}_L} \mathbf{L}_{p_L}^T(\mathbf{f}_L(\mathbf{P}(\mathbf{R}_{c_L}^{-1}(\mathbf{x}_j - \mathbf{c}_{c_L})))) \\ &+ \sum_{j\in\mathcal{A}_R} \mathbf{L}_{p_R}^T(\mathbf{f}_R(\mathbf{P}(\mathbf{R}_{c_R}^{-1}(\mathbf{x}_j - \mathbf{c}_{c_R})))), \quad (9)\end{aligned}$$

where $\mathcal{A}$ is the set of indices of *active* nodes, those model nodes on which image forces are to be exerted. Here the subscripts L and R denote dependence on the left and right images respectively and $\mathbf{P}$ describes the perspective projection equation.

3.5 Determining active model nodes

When our measurements are 2D images, as opposed to 3D range data, only a subset of the nodes on the model surface are selected to respond to forces. >From a given viewpoint, we can compute this active subset of model nodes based on the model's shape and orientation. In particular, a model node is made active if at least one of the following conditions is true:

1. it lies on the occluding contour of the model from that viewpoint,

2. the local surface curvature at the node is sufficiently large and the node is visible.

Instead of calculating analytically the positions of the active nodes on the model surface, we "loop" over all the nodes on the discretized model surface and check if one of the above two conditions is true. Condition 1 is true if: $|\mathbf{i}_j \cdot \mathbf{n}_j| < \tau$, where $\mathbf{n}_j$ is the unit normal at the jth model node, $\mathbf{i}_j$ is the unit vector from the focal point to that node on the model, and τ is a small threshold. Condition 2 is true if

$$\exists k \in K_j \text{ s.t. } |\mathbf{n}_k \cdot \mathbf{n}_j| > \kappa \;\&\; \exists k \in K_j \text{ s.t. } \mathbf{n}_k \cdot \mathbf{i}_k < 0, \quad (10)$$

where K_j is a set of indices of the nodes adjacent to the jth nodes on the model surface. κ in (10) is a threshold to determine if the angle between adjacent normal vectors is sufficiently large.

3.6 Tracking and prediction

We incorporate into our dynamic deformable model formulation a Kalman filter by treating the differential equations of motion (2) as the system model. Based on the corresponding extended Kalman filter, we perform tracking by updating the model's generalized coordinates $\mathbf{q}$ according to the following equation

$$\dot{\hat{\mathbf{u}}} = \mathbf{F}\hat{\mathbf{u}} + \mathbf{g} + \mathbf{P}\mathbf{H}^T\mathbf{V}^{-1}(\mathbf{z} - \mathbf{h}(\hat{\mathbf{u}})), \quad (11)$$

where $\mathbf{u} = (\dot{\mathbf{q}}^T, \mathbf{q}^T)^T$ and matrices $\mathbf{F}, \mathbf{H}, \mathbf{g}, \mathbf{P}, \mathbf{V}$ are associated with the model dynamics, the error in the given data and the measurement noise statistics [12]. Since we are measuring local forces directly from the image potential we compute, the term $\mathbf{z} - \mathbf{h}(\hat{\mathbf{u}})$ represents the 2D image forces. Using the above Kalman filter, we can predict at every step the expected location of the data in the next image frame, based on the magnitude of the estimated parameter derivatives $\dot{\mathbf{q}}$.

3.7 Self occlusion and disocclusion

As an object rotates in space, or as the viewpoint of the observer changes substantially, certain faces of the object will become occluded or disoccluded (a *visual event*). Hence, the corresponding line segment or edge feature in the image will appear or disappear over time. By using the Kalman filter to predict the position and orientation of the model in the next time frame, we can quantitatively predict the occurrence of a visual event. In other words, we can determine by using our active node determination approach, which subset of the model nodes will be active in the next image frame, and suppress their contributions to the net forces applied to the model. For stereo images, this prediction can be performed independently to the left and right images. In this case, two sets of active model nodes are maintained at any particular moment.

3.8 Tracking multiple objects with feedback

Our framework for object tracking can be extended to deal with multiple independently moving objects and multi-part objects. The complication here is that object parts may occlude one another in different ways. By tracking objects using stereo images, we can predict the 3D positions of the nodes on each model based on the current estimates of their respective model parameters and their rate of change. Active nodes on each model will be made "inactive" if they are predicted to be occluded by surfaces of other models. This visibility checking is performed for each model node against every surface of the other models in the scene. In practice, much of this checking can be avoided based on

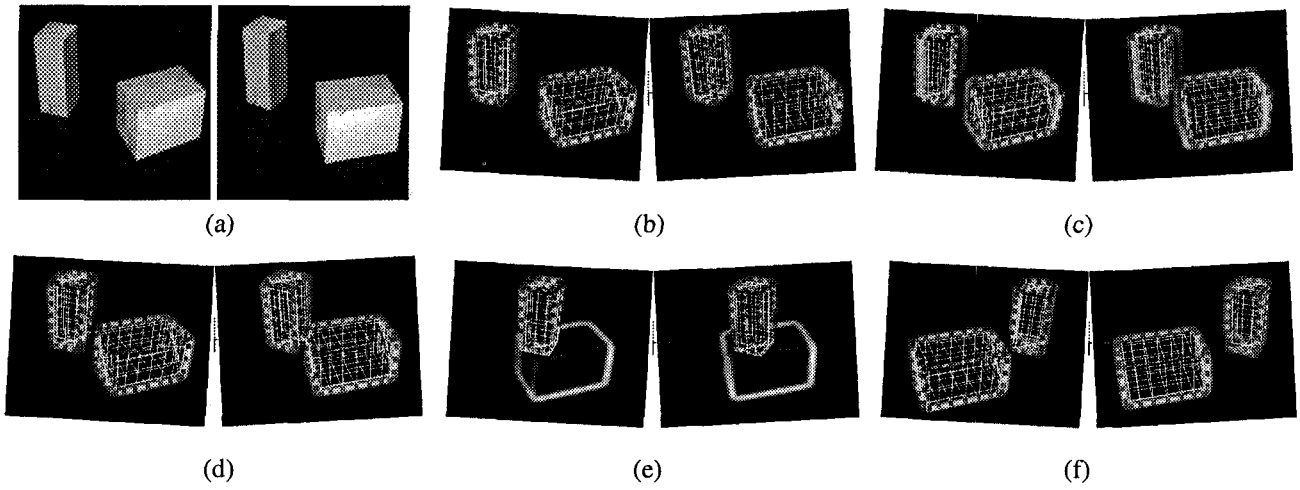

Figure 1: Tracking two independently moving blocks in a sequence of stereo images: (a) initialized models, (b) coming of a new frame, (c) beginning of the occlusion, (d) taller block partially occluded, (e) taller block becomes disoccluded, (f) no more occlusion. Note that the active model nodes are highlighted.

approximate estimates of each object's size and 3D location. We demonstrate in the next section that our approach can handle partial object occlusion.

There are also two more cases of object occlusion in case of multiple independently moving objects. The first case occurs when another moving object that was not previously present in the scene occludes the object being tracked. The second is due to an error from the qualitative segmentation system which did not detect an object during the model initialization step. By monitoring the exerted forces on the model's active nodes, if their magnitude exceeds a threshold or new unexpected forces are sensed, a feedback mechanism is invoked which triggers the application of the qualitative segmentation system to resolve the ambiguity. After proper re-initialization of our models, we continue the quantitative tracking using local image forces based on our physics-based technique.

4 Experiments

We demonstrate our approach in a series of tracking experiments involving real stereo image sequences. In the first experiment, we consider a sequence of stereo images of two independently moving objects. The objects move towards each other along 2 different linear paths, the relative angle between which is about 20 degrees. Fig. 1(a) shows the first pair of stereo images. The initial pose and shape of the objects are recovered using qualitative techniques mentioned before and they are subsequently tracked based on local image forces only. Figs. 1(b-g) show snapshots of the two objects being tracked with the wire-frame models overlaid on the image potential. They demonstrate that our technique is able to continue the tracking even when one of the blocks becomes partially occluded and then disoccluded. Note that those active model nodes which are temporarily occluded are automatically identified and made inactive.

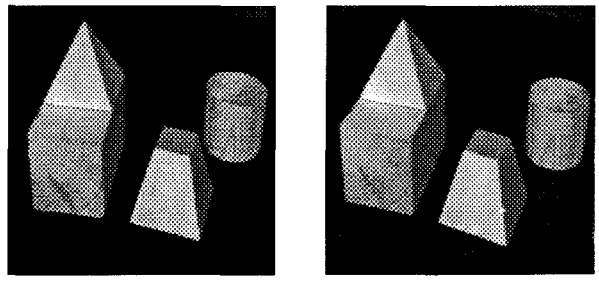

Figure 2: Initial pair of stereo images of a multi-object scene.

In the second experiment, we consider a sequence of stereo images of a scene containing multiple objects, including a two-part object. Fig. 2 shows the initial stereo images of the scene. The cameras are rotated around the scene at a constant rate. Fig. 3(a) shows the initialized models recovered using the same technique as before. Fig. 3(b) shows image potentials at an instant which the aspects of some parts have changed and some parts have become partially occluded. Each object is successfully tracked under these circumstances. Figs. 3(c-f) show the individual part models overlaid on the potentials in Fig. 3(b).

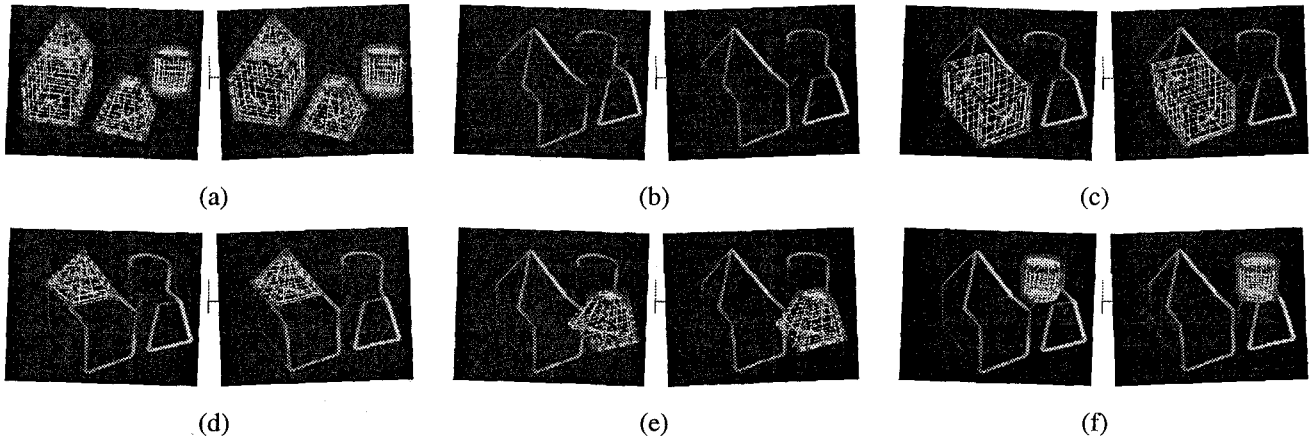

Figure 3: Tracking multiple objects in a sequence of stereo images (a) initialized models, (b) image potentials of an intermediate frame (both occlusions and visual events have occurred) (c-f) each object part correctly tracked with part models overlaid on the image potentials in (b). Note that the active model nodes are highlighted.

5 Conclusion

We have presented a new technique to object tracking in 3D from 2D stereo image sequences. After initializing our deformable models based on a part-based qualitative shape recovery process, we subsequently track the objects using only local image forces based on our physics-based approach. Costly feature correspondences are avoided. By integrating measurements from stereo images, we can continuously update the 3D positions (as well as other model parameters) of the objects using an extended Kalman filter. We also demonstrated that our model-based approach can deal with visual events and occlusions in scenes with multiple moving objects by predicting their occurrences. Only geometric information has been used here. We plan to study the integration of other low-level tracking techniques into our framework.

References

[1] A. Blake, R. Curwen, and A. Zisserman. Affine-invariant contour tracking with automatic control of spatiotemporal scale. In *Proc. IEEE 4th International Conference on Computer Vision*, pages 502–507, 1993.

[2] T. J. Broida, S. Chandrashekhar, and R. Chellappa. Recursive 3-D motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26(4):639–656, 1990.

[3] R. Deriche and O. Faugeras. Tracking line segments. *Image and Vision Computing*, 8(4):261–270, 1990.

[4] S. Dickinson, A. Pentland, and A. Rosenfeld. Shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.

[5] E. D. Dickmanns and Volker Graefe. Applications of dynamic monocular machine vision. *Machine Vision and Applications*, 1:241–261, 1988.

[6] J. S. Duncan, R. L. Owen, and P. Anandan. Measurement of nonrigid motion using contour shape descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 318–324, 1991.

[7] D. Gennery. Visual tracking of known three-dimensional objects. *International Journal of Computer Vision*, 7(3):243–270, 1992.

[8] T. S. Huang. Modeling, analysis and visualization of nonrigid object motion. In *Proc. IEEE 10th International Conference on Pattern Recognition*, volume 1, pages 361–364, 1990.

[9] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[10] D. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.

[11] D. Metaxas and S. Dickinson. Integration of quantitative and qualitative techniques for deformable model fitting from orthographic, perspective, and stereo projections. In *Proc. IEEE 4th International Conference on Computer Vision*, pages 641–649, 1993.

[12] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, 1993.

[13] A. Pentland and B. Horowitz. Recovery of non-rigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730–742, 1991.

[14] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.