Singularities in Articulated Object Tracking with 2-D and 3-D Models

James M. Rehg *Daniel D. Morris*

Cambridge Research Laboratory
Technical Report Series

CRL 97/8
October 1997

Cambridge Research Laboratory

The Cambridge Research Laboratory was founded in 1987 to advance the state of the art in both core computing and human-computer interaction, and to use the knowledge so gained to support the Company's corporate objectives. We believe this is best accomplished through interconnected pursuits in technology creation, advanced systems engineering, and business development. We are actively investigating scalable computing; mobile computing; vision-based human and scene sensing; speech interaction; computer-animated synthetic persona; intelligent information appliances; and the capture, coding, storage, indexing, retrieval, decoding, and rendering of multimedia data. We recognize and embrace a technology creation model which is characterized by three major phases:

Freedom: The life blood of the Laboratory comes from the observations and imaginations of our research staff. It is here that challenging research problems are uncovered (through discussions with customers, through interactions with others in the Corporation, through other professional interactions, through reading, and the like) or that new ideas are born. For any such problem or idea, this phase culminates in the nucleation of a project team around a well articulated central research question and the outlining of a research plan.

Focus: Once a team is formed, we aggressively pursue the creation of new technology based on the plan. This may involve direct collaboration with other technical professionals inside and outside the Corporation. This phase culminates in the demonstrable creation of new technology which may take any of a number of forms - a journal article, a technical talk, a working prototype, a patent application, or some combination of these. The research team is typically augmented with other resident professionals—engineering and business development—who work as integral members of the core team to prepare preliminary plans for how best to leverage this new knowledge, either through internal transfer of technology or through other means.

Follow-through: We actively pursue taking the best technologies to the marketplace. For those opportunities which are not immediately transferred internally and where the team has identified a significant opportunity, the business development and engineering staff will lead early-stage commercial development, often in conjunction with members of the research staff. While the value to the Corporation of taking these new ideas to the market is clear, it also has a significant positive impact on our future research work by providing the means to understand intimately the problems and opportunities in the market and to more fully exercise our ideas and concepts in real-world settings.

Throughout this process, communicating our understanding is a critical part of what we do, and participating in the larger technical community—through the publication of refereed journal articles and the presentation of our ideas at conferences–is essential. Our technical report series supports and facilitates broad and early dissemination of our work. We welcome your feedback on its effectiveness.

Robert A. Iannucci, Ph.D.
Director

Singularities in Articulated Object Tracking with 2-D and 3-D Models

James M. Rehg Daniel D. Morris

October 15, 1997

Abstract

Three dimensional kinematic models are widely-used in video-based figure tracking. We show that these models can suffer from singularities when motion is directed along the viewing axis of a single camera. The single camera case is important because it arises in many interesting applications, such as motion capture from movie footage.

We describe a novel 2-D Scaled Prismatic Model (SPM) for figure registration. In contrast to 3-D kinematic models, the SPM has fewer singularity problems and does not require detailed knowledge of the 3-D kinematics. We fully characterize the singularities in the SPM and demonstrate tracking through singularities using synthetic and real examples.

We demonstrate the application of our model to motion capture from movies. Fred Astaire is tracked in a clip from the film "Shall We Dance". Some simple video edits are presented. These results demonstrate the benefits of the SPM in tracking with a single source of video.

Rehg is with the Cambridge Research Laboratory.
Email: `rehg@crl.dec.com`

Address for Morris:
Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213, USA.
Email: `ddm@ri.cmu.edu`

The work described in this report was performed at the Cambridge Research Laboratory.

CRL Technical reports are available on the CRL's web page at
http://www.crl.research.digital.com.

Digital Equipment Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700
Cambridge, Massachusetts 02139 USA

1 Introduction

The kinematics of an articulated object provide the most fundamental constraint on its motion, and there has been a significant amount of research into the use of 3-D kinematic models for visual tracking of humans [8, 3, 13, 22, 5]. Kinematic models play two roles in tracking. First, they define the desired output—a state vector of joint angles that encodes the 3-D configuration of the model. Second, they specify the mapping between states and image features that makes registration possible.

Nonlinear least-squares tracking techniques that minimize a cost function over the state space have proven to be highly effective [8, 14, 12]. These techniques use the gradient of the residual error to obtain a locally linear model. There are two primary requirements for their success. First, to obtain a gradient the error function must be differentiable. Discontinuities can occur during occlusions and these have been addressed in [15, 8].

The second requirement is that the state space must be fully observable, ensuring that the constraints imposed by the kinematic model accurately reflect the motion of the object. Loss of observability occurs when some states have no instantaneous effect on the image features and the kinematic Jacobian loses rank and becomes singular. Kinematic singularities occur for particular configurations of the object relative to the camera, and can be reduced through the use of multiple camera viewpoints [13]. Unfortunately in certain tracking applications, such as motion capture from movie footage, there is only a single video source available.

An alternative to the direct 3-D tracking approach is to decompose figure tracking into separate *image registration* and *3-D reconstruction* stages, as is currently done in structure from motion problems [20]. This decomposition has two potential benefits. First, the registration stage can employ simple 2-D figure models which avoid most of the singularity problems associated with 3-D tracking in the case of a single video source. Second, the reconstruction stage can simultaneously estimate both dynamic state parameters, such as joint angles, and static parameters, such as link lengths. This would remove the need to specify an accurate figure model for 3-D tracking.

In this paper we introduce a novel class of 2-D kinematic models for figure registration, which we call *scaled prismatic models* (SPM). We show how to derive the SPM associated with an arbitrary 3-D kinematic model and demonstrate that SPM's have far fewer singularity problems than conventional 3-D models. We present a detailed discussion of the effect of singularities on tracking branched, open kinematic chains, along with experimental results for motion capture from movies. These results provide the first detailed analysis of singularities in articulated object tracking.

2 Singularities in Visual Tracking with 3-D Kinematic Models

We begin by analyzing the effect of singularities on gradient-based tracking algorithms for 3-D kinematic models. The standard approach is based on the direct registration of

3-D models with image features.[1] In this method, feature attributes such as the image position of an edge or a template are expressed as a function of the kinematic state variables, e.g. joint angles. State estimates are chosen by minimizing a residual error measure defined in the image. For example, the residual error for an SSD template feature,

$$R_j(\mathbf{q}) = I_j(\mathbf{q}) - \hat{I}_j, \quad I_j(\mathbf{q}) = I(\mathbf{PF}(\mathbf{q}, \mathbf{p}_j)), \tag{1}$$

measures the registration between template pixel $\hat{I}_j$ and the corresponding pixel $I_j(\mathbf{q})$ in input image I, given the state vector $\mathbf{q}$. Orthographic camera projection is modeled by $\mathbf{P}$ and the 3-D kinematics by the nonlinear function $\mathbf{F}(\mathbf{q}, \mathbf{p}_j)$, where $\mathbf{p}_j$ is the 3-D point corresponding to pixel $\hat{I}_j$. Figure (1) illustrates the geometric relationship between kinematic template model and image.

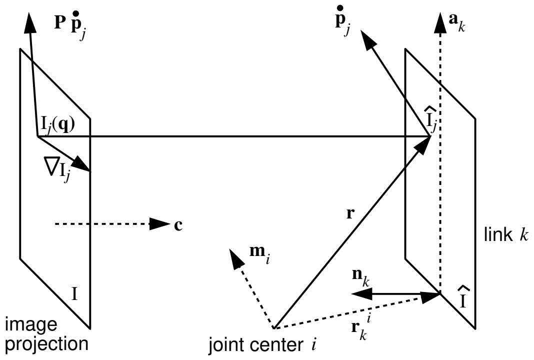


Figure 1: Schema of the projection along the camera axis $\mathbf{c}$ of the template attached to link k . Here template velocity is due to rotation around axis $\mathbf{m}_i$ of link i. Point $\hat{I}_j$ has position $\mathbf{p}_j$ and velocity $\dot{\mathbf{p}}_j$ in 3-D, and this is projected to pixel I_j whose image gradient is ∇I_j.

Given a template model for each link in the object, tracking proceeds by minimizing the squared residual error, $E(\mathbf{q}) = \frac{1}{2}\mathbf{R}^T\mathbf{R}$, where the vector $\mathbf{R}$ holds a rasterization of the residual from Equation 1 over all of the template pixels. Algorithms such as Levenberg-Marquardt [2] use the linearized residual at an operating point $\mathbf{q}_0$ to compute a step towards the local minima.[2] The residual gradient for pixel I_j can be expressed

$$\dot{R}_j = [\frac{\partial R_j}{\partial \mathbf{q}}(\mathbf{q}_0)]\dot{\mathbf{q}} \equiv \mathbf{J}_j \dot{\mathbf{q}} \tag{2}$$

$$J_{ji} \equiv \frac{\partial R_j}{\partial q_i} = (\nabla I_j)^T \mathbf{P} \frac{\partial \mathbf{F}}{\partial \mathbf{q}}(\mathbf{q}_0, \mathbf{p}_j) \tag{3}$$

As Equation 3 demonstrates, the residual Jacobian is made up of three terms: the standard 3-D kinematic Jacobian, $\mathbf{J}_j^k \equiv \partial\mathbf{F}/\partial\mathbf{q}$ [19], the camera projection model $\mathbf{P}$,

[1] We use the term “feature” to describe a wide variety of measurements, including flow [22, 1], templates [15, 17], and edges [14, 8].

[2] In a tracking application, $\mathbf{q}_0$ is given by the estimate from the previous frame.

and the image feature gradient, which in this case is ∇I_j. By definition, we also have $\dot{\mathbf{p}}_j = \mathbf{J}_j^k \dot{\mathbf{q}}$. We see that the residual velocity is the result of projecting a 3-D point velocity through the camera model and along the image feature gradient, as illustrated in Figure (1).

From the figure it is clear that pixel I_j will provide no information about q_i if $\dot{\mathbf{p}}_j$ is directed along the optical axis or if $\mathbf{P}\dot{\mathbf{p}}_j$ acts perpendicular to the gradient. Other possible feature gradients in equation (3) include the curve normal for a contour feature, or the identity matrix in the case of a point feature.

The complete residual Jacobian $J(\mathbf{q})$ is formed by stacking up $\mathbf{J}_j$'s from equation (2) for all points I_j, resulting in a linear map from state space to residual space. The nullspaces of this mapping provide fundamental insight into its properties. The left nullspace of the Jacobian, $\mathcal{N}(\mathbf{J}^T)$, defines the constraints inherent in the kinematic model. Residual velocities in the left nullspace, $\dot{R} \subseteq \mathcal{N}(\mathbf{J}^T)$, are excluded by equation (2). An empty left nullspace indicates that the kinematics do not constrain the motion. In tracking there will typically be more image measurements than parameters, $m > n$, resulting in a non-empty left nullspace and $\mathrm{rank}(\mathbf{J}) = n$.

The right nullspace of the residual Jacobian defines the *observability singularities* of the articulated object. State velocities in the right nullspace, $\dot{\mathbf{q}} \subseteq \mathcal{N}(\mathbf{J})$, do not effect the residual $\dot{R}$ and so are termed *singular directions*. The right nullspace is non-zero only when the Jacobian has lost rank, i.e. $\mathrm{rank}(\mathbf{J}) < n$.

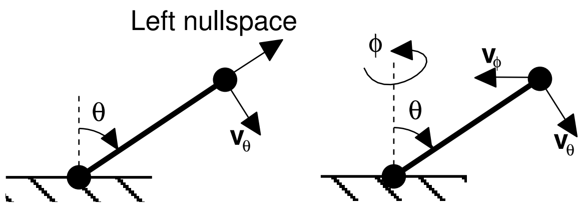


Figure 2: Examples of (a) 1, and (b) 2 degree of freedom manipulators.

2.1 Examples of 3-D Singularities

We illustrate the Jacobian's properties with two simple examples. Figure (2a) shows a one-link revolute planar manipulator with a single degree of freedom (DOF) θ. The residual Jacobian for the end-point feature is defined by:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \cos(\theta) \\ -\sin(\theta) \end{bmatrix} \begin{bmatrix} \dot{\theta} \end{bmatrix}, \tag{4}$$

assuming that the camera and joint axes are parallel. The kinematic constraint is given by the left nullspace: $\dot{R}_c = \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T$. The right nullspace is empty and there are no observability singularities.

Next consider the manipulator in Figure (2b), formed by adding an additional DOF, ϕ, to Figure (a), which allows the link plane to tilt out of the page. With the same point feature and camera viewpoint we have

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\sin(\theta)\sin(\phi) & \cos(\theta)\cos(\phi) \\ 0 & -\sin(\theta) \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \end{bmatrix} \tag{5}$$

Singularities now occur when $sin(\phi) = 0$ and also when $sin(\theta) = 0$. In both cases the singular direction is $\dot{\mathbf{q}} = [1\ 0]^T$, implying that changes in ϕ cannot be recovered in these two configurations.

Singularities impact visual tracking through their effect on error minimization. Consider tracking the model of Figure (2b) using the Levenberg-Marquardt update step:

$$\mathbf{q}_k = \mathbf{q}_{k-1} + \mathbf{dq}_k = \mathbf{q}_{k-1} - (\mathbf{J}^T\mathbf{J} + \mathbf{\Lambda})^{-1}\mathbf{J}^T\mathbf{R} \tag{6}$$

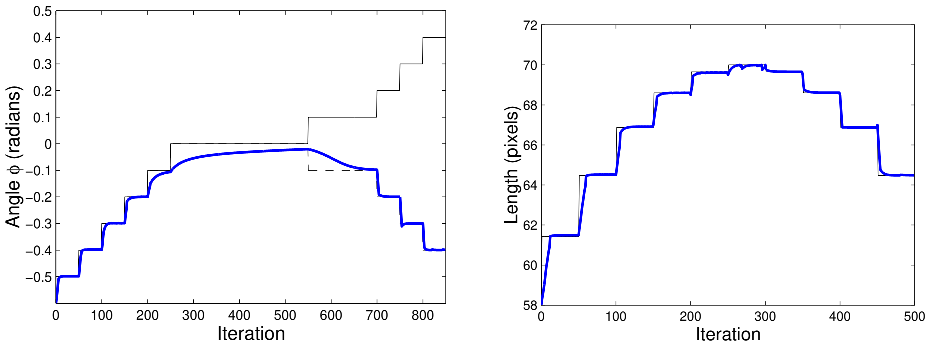
where Λ is a diagonal stabilizing matrix. At the singularity $sin(\phi) = 0$, the update step for all trajectories has the form $\mathbf{dq}_k = [0\ C]$, implying that no updates to ϕ will occur regardless of the measured feature motion. This singularity arises physically when the link rotates through the plane parallel to the image plane, resulting in a point velocity in the direction of the camera axis.

Figure (3a) illustrates the practical implications of singularities for tracker performance. The stair-stepped curve corresponds to discrete steps in ϕ in a simulation of the two DOF example model. In this example, the arm is planar with a randomly textured template model. The solid curve shows the state estimates produced by Equation (6) as a function of the number of iterations of the solver. The loss of useful gradient information and resulting slow convergence of the estimator as the trajectory approaches the point $\phi = 0$ is symptomatic of tracking near singularities. In this example, the singular state was never reached and the tracker continued in a direction opposite the true motion, as a consequence of the usual reflective ambiguity under orthographic projection (shown by the dashed line). Perspective projection also suffers from this ambiguity for small motions.

These examples illustrate the significant implications of singularities for visual tracking. If the search for feature measurements is driven by prediction from the state estimates, singularities could result in losing track of the target altogether. Even when feature correspondences are always available, such as when markers are attached to the object, the solver will slow down dramatically near singularities, since each step has only a small effect on the residual. This is analogous to the effect of classical kinematic singularities in robotic manipulators [9]: manipulator control near singularities may require arbitrarily large forces; here tracking near singularities may require arbitrarily large numbers of iterations!

2.2 Conditions for 3-D Singularities

It would be useful to obtain general conditions under which singularities can arise in tracking with 3-D kinematic models. This is a challenging task due to the high dimensionality and nonlinearity of kinematic models. Attempts have been made to classify the singularities in robot manipulators from the standpoint of both manipulator design [11] and visual-servoing control [18]. In this section we derive some local conditions for 3-D singularities and characterize some important special cases.



(a) (b)

Figure 3: Singularity example. (a) Tracking the 3-D manipulator in example 2 through a singular point along the singular direction. While the true angle ϕ continues to increase, the tracker loses track near the singularity and then picks up an ambiguous path. (b) 2-D tracker from Section 3 is applied to the same motion as in (a), but here extension length rather than angle is recovered, and this correctly increases and decreases without change in damping.

From the preceding discussion and Equation (3), it is clear that singularities depend fundamentally on the interaction between projected 3-D point velocities and their associated image feature gradients. Jacobian analysis and computation can be simplified by assuming that the shape of each link is locally planar or cylindrical. This is a good approximation when the depth variation along the link is small relative the camera distance.

In this analysis we employ the planar link model introduced in [13]. The image appearance of the *kth* link is modeled by the projection of a view-dependent texture-mapped plane which is attached to the link's coordinate frame. The normal vector, $\mathbf{n}_k$, of the plane is adjusted as the link moves to satisfy the following two constraints: $\mathbf{n}_k \perp \mathbf{a}_k$ and $\mathbf{n}_k \perp (\mathbf{a}_k \times \mathbf{c})$ where $\mathbf{a}_k$ defines the centerline of the link and $\mathbf{c}$ is the camera axis (see Figure (1)). These constraints keep the plane "facing" the camera. The plane is considered to be rigidly attached to the link for the purpose of Jacobian computation.

The planar link model can be initialized from a single image and then applied across an image sequence if the link appearance doesn't change dramatically with the viewing direction. When this assumption fails, the template contents can be allowed to change with the viewing direction.

In the point feature case we have the following condition for a 3-D point $\mathbf{p}$ to contribute *no information* about a revolute state q_i: $\mathbf{c} \parallel \dot{\mathbf{p}} \Rightarrow \mathbf{c} \parallel \dot{q}_i \mathbf{m}_i \times \mathbf{r}$ where $\mathbf{r}$ is the vector from the joint center of link i to $\mathbf{p}$, $\mathbf{m}_i$ is the joint axis, and $\mathbf{c}$ is the camera axis, all expressed in world coordinates.

Now let $\mathbf{p}$ be located on the template plane for link k. Then $\mathbf{r} = \rho\mathbf{r}_k^i + u\mathbf{a}_k + v(\mathbf{a}_k \times \mathbf{n}_k)$ where $\mathbf{a}_k$ and $\mathbf{n}_k$ define the template plane as above, u, v give the position of $\mathbf{p}$ in template coordinates, and $\rho\mathbf{r}_k^i$ is the vector connecting joint i to the base of link k. All vectors are unit vectors.

We can now derive conditions on the template plane such that all of its point velocities project along the camera axis. We find three conditions:

$$\mathbf{n}_k \parallel \mathbf{c}, \quad \mathbf{r}_k^i \perp \mathbf{c}, \quad \mathbf{m}_i \perp \mathbf{c}, \tag{7}$$

which together are sufficient. We can make three observations about these conditions. First, use of the planar appearance model makes it possible to greatly simplify the singularity analysis for a link. Furthermore, when $\mathbf{a}_k$ is directed along the line connecting the joint centers, the link can be modeled as a 3-D line segment for analysis purposes. Second, this analysis does not rely upon any particular parameterization of the model's kinematic DOF's and should apply quite broadly to models of the figure. Third, additional singular configurations can arise in cases, such as along contours, where the feature model does not fully constrain image motion.

Finally, it is worth pointing out that in spite of the potential problems in using 3-D kinematic models for tracking, they can be extremely reliable in practice if a sufficient number of camera views are available. This observation is the basis for the optical motion capture industry, for example.

3 A 2-D Scaled Prismatic Model for Registration

The previous section outlined the conditions under which singularities can occur in 3-D kinematic models. Singularities have two implications for 3-D tracking with a single video source. First, some additional source of information will have to be used to estimate the unobservable parts of the state space for singular configurations. For example, assumptions about object dynamics could be used in a Kalman filter framework to extrapolate an estimated state trajectory across a singular point. Second, the utility of the kinematic model for image registration is reduced, since it will not always supply a useful constraint on pixel motion.

It is important here to distinguish two separate goals: a registration objective in which the model projections are aligned with image pixels, and a reconstruction goal in which the state trajectory for a 3-D kinematic model is estimated. For some applications, such as gesture recognition, registration may be all that is required. In other applications such as motion capture it is desirable to reconstruct the 3-D motion along with the kinematic model parameters. Once the images have been registered, 3-D reconstruction can be cast as a batch estimation problem, since the registration step gives the complete correspondence between model points and image coordinates in each frame. The batch nature of the formulation is well-suited to our application of motion capture from movies, and makes it possible to enforce smoothness constraints in both time directions, improving the quality of the estimates.

The remainder of this section focuses on the registration step. Although we do not want to employ the full 3-D model, we would like to employ the strongest possible kinematic constraints so as to improve robustness to image noise. We will see that a

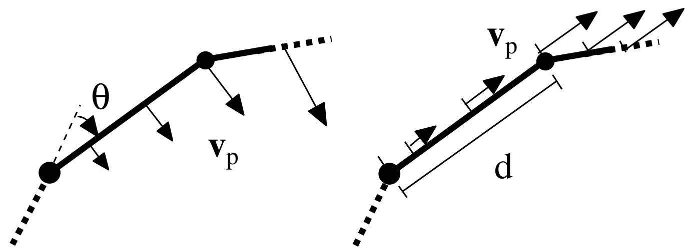


Figure 4: 2-D SPM chain showing residual velocities due to state velocities: (a) $\dot{q}_i = \dot{\theta}$, and (b) $\dot{q}_i = \dot{d}$.

novel 2-D Scaled Prismatic Model (SPM) formed by "projecting" the 3-D model into the image plane provides a useful constraint for registration.

3.1 Kinematics of the 2-D SPM Class

The 2-D SPM acts in a plane parallel to the plane of the camera and simulates the image motion of the 3-D model. Links have the same connectivity as in the 3-D model, rotate around their base joint on an axis perpendicular to the plane, and scale uniformly along their length. Each link is thus represented as a line segment having two parameters; its angle of rotation θ_i and length d_i along its direction $\mathbf{n}_i$. As in the 3-D case a template is attached to each link which rotates and scales with the link. Figure (4) shows both of these parameters for a link in the 2-D SPM.

In this section we briefly derive the kinematics of this model class, show that it can capture the projected motion of a 3-D figure, and then show that it is precisely in the cases where the 3-D model suffers from singularities that the 2-D SPM behaves well.

The residual velocity can be expressed as the sum of the Jacobian columns times their corresponding state parameter velocity: $\dot{\mathbf{R}} = \sum_i \mathbf{J}_i \dot{q}_i$. Hence we can calculate individual Jacobian columns for each state variable, q_i independently and then combine them. Since a column of the Jacobian, $\mathbf{J}_i$, maps the state velocity $\dot{q}_i$ to a residual velocity, by finding the residual velocity in terms of this state we can obtain an expression for $\mathbf{J}_i$. If $q_i = \theta$ is the angle of a revolute joint shown in Figure (4a), it will contribute an angular velocity component to links further along the chain given by $\omega = \dot{q}\mathbf{a}$. Here $\mathbf{a}$ is the axis of rotation which for the 2-D model is just the z axis. The image velocity, $\mathbf{v}_p$, of a point at location $\mathbf{r}$ on the manipulator chain resulting from this rotation is given by:

$$\mathbf{v}_p = \mathbf{P}\omega \times \mathbf{r} = \mathbf{P}\mathbf{a} \times \mathbf{r}\dot{q} = \mathbf{r}_{2d}\dot{q} \tag{8}$$

where the orthographic projection $\mathbf{P}$ selects the x and y components. This equation expresses the desired mapping from state velocities of axis i to image velocities of point j on link k giving the components of the column Jacobian, $\mathbf{J}_i$:

$$J_{ji} = \begin{cases} 0 & \text{links } k\text{, where } k < i \\ \mathbf{r}_{2d} & \text{links } k\text{, where } k \geq i \end{cases} \tag{9}$$

If $q_i = d$ refers to the extension of the scaled prismatic link shown in Figure (4b), its derivative will contribute a velocity component to points on link i proportional to their position on the link: bq_i, where b is the fractional position of the point over the total extension q_i. The velocity component for a point, p, on the link is thus $\mathbf{v}_p = bq_i\dot{q}_i\mathbf{n}_i$. Subsequent links, $k > i$, will be affected only by the end-point extension of the link, and so have a velocity component from this joint given by: $\mathbf{v}_p = q_i\dot{q}_i\mathbf{n}_i$. Hence the Jacobian element at point j on link k for an extension parameter, q_i, is given by:

$$J_{ji} = \begin{cases} 0 & \text{links } k\text{, where } k < i \\ bq_i\mathbf{n}_i & \text{link } i \\ q_i\mathbf{n}_i & \text{links } k\text{, where } k > i \end{cases} \tag{10}$$

We show that given certain modeling assumptions, the 2-D SPM with the above specifications is flexible enough to represent the projected image of any 3-D model in any legal configuration. We assume that a model consists of a branched chain of links connected at their end-points by revolute joints. We use the template plane model from Section (2.2) to describe link appearance. We identify the *link segment* for each link as the 3-D line segment connecting the link's joint centers and oriented in the direction of $\mathbf{a}_k$. The 3-D model specifies the link lengths and the orientation of each revolute joint axis, while in the SPM the link lengths vary and the axis of each revolute joint is perpendicular to the image plane. The state of a 3-D model is thus a vector of joint angles, $\mathbf{q}_m = [\ \phi_1 \quad \phi_2 \quad \ldots\ ]^T$, and the state of a 2-D SPM is a vector of angles and joint lengths, $\mathbf{q}_n = [\ \theta_1 \quad d_1 \quad \theta_2 \quad d_2 \quad \ldots\ ]^T$. Then more formally:

Proposition 1 *The linear projection of the link segments of a 3-D kinematic model onto a plane and the assignment of revolute joints with axes perpendicular to the plane between each pair of connected links defines a many to one mapping $\mathcal{F}_M : \mathbf{M}^3 \to \mathbf{M}^2$ from the space of all 3-D models $\mathbf{M}^3$ to 2-D models $\mathbf{M}^2$. Furthermore for each pair of models, $m \in \mathbf{M}^3$ and $n = \mathcal{F}_M(m)$, it defines another mapping $\mathcal{F}_S : \mathbf{Q}^3_m \to \mathbf{Q}^2_n$ that maps every state of the 3-D model $\mathbf{q}_m \in \mathbf{Q}^3_m$ to a state of the 2-D SPM $\mathbf{q}_n \in \mathbf{Q}^2_n$.*

Proof: Consider the graph, G, of a 3-D model with each joint represented by a vertex and each link by an edge. There may be many 3-D models with the same graph G since 3-D joints may have multiple revolute axes. When a 3-D model in any state is projected onto a plane under a linear projection, the new graph G' will have the same topology of vertices and edges, and the projected edges will remain linear. Now interpret the graph, G', drawn in the plane as each straight edge representing an extensible link, and the intersection point of each connected pair of edges as a revolute joint. This defines a unique 2-D model, and thus the mapping $\mathcal{F}_M$. The state of a 2-D SPM is specified by the distances in the plane between connected joints (i.e. the link lengths d_i's), and the angles between links that share joints (i.e. θ_i's) as illustrated in Figure (4). Now the state of the 3-D model determines, through the projection, the relative positions of the vertices in 2-D and thus the 2-D state. For any distribution of vertices in the plane here must exist a 2-D state $\mathbf{q}_n$ that captures it since line segments can join any two connected vertices, and any relative orientation between two line segments can be described by a single angle. There thus must exist a mapping $\mathcal{F}_S$ for all 3-D states.

We conclude that the 2-D SPM class can capture any projected 3-D model in any configuration.

3.2 Singularity Analysis of the 2-D SPM

An important advantage of the SPM is the location of its singularities. In the 3-D model the singularities occur in the frequently traversed region of configuration space in which links pass through the image plane. The 2-D SPM has all of its rotation axes parallel to the camera axis and so never fulfils the 3-D singularity condition: $\mathbf{m}_i \perp \mathbf{c}$ from Equation (7). Here we show that the SPM only has singularities when $d_i = 0$, corresponding to a 3-D link pointing towards the camera, and that the singular direction is perpendicular to the entering velocity and so usually does not affect tracking.

Proposition 2 *Given x and y measurements of endpoints of each joint in a linear chain scaled-prismatic manipulator, observability singularities occur if and only if at least one of the joint lengths is zero.*

Proof: We define a state vector made of pairs of components for each link: $\mathbf{q} = \begin{bmatrix} \theta_1 & d_1 & \dots & \theta_n & d_n \end{bmatrix}^T$, and the residual vector to be the error in x and y end-point positions of each link. We assume the proposition holds for a $n-1$ link manipulator with Jacobian $\mathbf{J}^{(n-1)}$ whose elements are defined as in Equations (9) and (10). The Jacobian for the n length manipulator is given by:

$$\mathbf{J}^{(n)} = \begin{bmatrix} \mathbf{J}^{(n-1)} & A \\ B & C \end{bmatrix} \tag{11}$$

where $\mathbf{J}^{(n-1)}$ is a square matrix of size $2n-2$. Matrix A is of size $2n-2 \times 2$ and expresses the dependence of the n'th link's parameters on the position of the other links positions and so is zero. Matrix C and its square are given as:

$$C = \begin{bmatrix} \cos(\theta_T) & -d_n \sin(\theta_T) \\ \sin(\theta_T) & d_n \cos(\theta_T) \end{bmatrix}, \tag{12}$$

$$C^T C = \begin{bmatrix} 1 & 0 \\ 0 & d_n^2 \end{bmatrix} \tag{13}$$

where $\theta_T = \sum_{i=1}^{n} \theta_i$. From this we see that C has rank two if and only if $d_n \neq 0$. If C has rank two, then the bottom two rows of $\mathbf{J}^{(n)}$ are linearly independent of all other rows and if $\mathbf{J}^{(n-1)}$ is full rank then $\mathbf{J}^{(n)}$ must have rank $2n$. If C or if $\mathbf{J}^{(n-1)}$ do not have full rank then $\mathbf{J}^{(n)}$ will not have rank $2n$, and there will be an observability singularity. To complete the proof we need only demonstrate that the proposition applies to the base case, $n = 1$. Here the whole Jacobian is given by C which has full rank only when $d_1 \neq 0$. Thus the proposition is proven.

A further mitigating property of the 2-D singularities is that unlike in the 3-D observability singularities where the singular direction is along the motion trajectory, the singular direction in the 2-D case is always perpendicular to the direction in which the singularity was entered. We can see this for the single arm manipulator described by a Jacobian equal to C in equation (12). When $d = 0$ the velocity direction is: $\dot{R} = \begin{bmatrix} \cos(\theta) & sin(\theta) \end{bmatrix}^T$, but the left nullspace is orthogonal to this by definition. Hence a manipulator will typically pass through a 2-D singularity without the increased damping caused by moving along a singular direction. Only if the link enters in one direction and leaves orthogonally does the singularity obstruct tracking.

The assumption that we have information on endpoints is equivalent to assuming there is sufficient texture or edge information on the link to obtain length and direction estimates. When this assumption fails there may be more singularities for both the 3-D and 2-D models.

While both 2-D and 3-D model classes can represent articulated motion, the 2-D SPM provides weaker constraints. It has the two advantages of avoiding the singularities of the 3-D model and relaxing the need for accurate knowledge of link lengths and joint axes which is required by the 3-D model. Moreover, the 2-D and 3-D models are complementary in that their singularities occur in different parts of the state space.

4 Previous Work

There have been numerous papers on 3-D and 2-D tracking of articulated objects. However, none of them have addressed the question of singularities and their implications for tracking with a single video source. The 3-D kinematic analysis in this paper is based primarily on our earlier work [14, 15]. We believe it applies quite broadly.

The first works on articulated 3-D tracking were by O'Rourke and Badler [10] and David Hogg [5]. They employed the classical AI techniques of constraint propagation and tree search, respectively, to estimate the state of the figure. Hogg was the first to show results for a real image sequence. A modern version of the discrete search strategy is employed by Gavrila and Davis [3], who use a hierarchical decomposition of the state space to search for the pose of a full body 3-D model.

Yamamoto and Koshikawa [22] were the first to apply modern kinematic models and gradient-based optimization techniques to figure tracking. Their experimental results were limited to 2-D motion and they did not address kinematic modeling issues in depth. The gradient-based tracking framework was extended by Rehg and Kanade to handle self-occlusions [15] and applied to hand tracking [13].

The objective function used by Yamamoto et. al. is still popular. It compares measured image flow to the image flow predicted by the kinematic Jacobian of the figure. The same approach was explored by Pentland and Horowitz for nonrigid motion analysis in [12], which includes an example of figure tracking. More recently, Bregler and Malik have used the same cost function in their analysis of the Muybridge plates [1].

A number of 3-D tracking systems have used explicit shape models for the limbs, usually some form of superquadric or generalized cylinder. The system of Kakadiaris and Metaxas is a more complete example, and addresses model acquisition [7] and self-occlusion handling [8]. Related work by Goncalves and Perona is described in [4].

Gradient-based search strategies have the crucial performance advantage of exploiting information about the local shape of the objective function to select good search directions. This leads to extremely fast search performance in high dimensional state spaces. For example, Rehg and Kanade demonstrated tracking of a 26 DOF hand model using live video [14]. In addition, the field of robot control is based on an analogous use of the kinematic Jacobian. However, as in robotics, singularities often arise in 3-D tracking and can cause significant problems, as we have demonstrated.

The work of Ju et. al. [17] is perhaps the closest to our 2-D Scaled Prismatic Model. They model the image motion of rigid links as affine flow patches with imposed attachment constraints. Their model can be viewed as an extension of classical patch tracking techniques, that incorporates constraints between patches. In contrast, our model is derived from an explicit consideration of the effects of 3-D kinematics on image motion. As a consequence, it has a more direct connection to the underlying 3-D motion. We believe this property will become important in reconstructing the 3-D motion of the figure from SPM measurements. The SPM also has fewer parameters than a patch-based description of flow.

5 Experimental Results

We present two sets of experimental results that demonstrate the differences between 3-D and 2-D tracking for real image sequences and give some preliminary results for our motion capture from movies application.

5.1 Comparison between 2-D and 3-D Models

Figure (5a) and (5b) show the starting and ending frames of a 30 frame sequence of an arm moving through a singularity. In this example the arm remains rigid, approximating the model of Figure (2b), but with the addition of a base link capable of translation in the image plane. The trajectory of the arm was similar to the simulation in Figure (3), but with the addition of a nonzero θ component. Overlaid on the images are the positions of the 2-D SPM resulting from the state estimates. The longer part of the "T" shape on the arm is the prismatic joint axis. The second link superimposed on the torso has X and Y translation DOF's, which were negligible.

We conducted three experiments in which the sequence in Figure (5) was tracked with an SPM and two 3-D kinematic models with different damping factors Λ. In each case, the tracker was given a budget of twenty iterations with which to follow the motion in a given frame. By analogy to the simulation example, we would expect the 3-D models to lose ground in the vicinity of a singularity. Figure (6a–c) compares the relative performance of the 2-D and 3-D models. Plot (a) shows the length of the arm link projected into the image plane for the three trackers. As expected, the 2-D SPM tracker is unaffected by the singularity and exhibits uniform convergence rates throughout the trajectory. The extension of the arm corresponds to the prismatic state d_1 in the SPM model, which is plotted with large dots in the figure.

The under-damped 3-D tracker drawn with dashed lines in Plot (a) performs well until it approaches the singularity, upon which it begins oscillating wildly. These oscillations in projected arm length are the result of fluctuations in the out-of-plane rotation angle, which is shown in Plot (b). Once the arm leaves the singular configuration the under-damped tracker recovers and tracks the remainder of the sequence. In contrast, the well-damped tracker plotted with a solid line in Plot (a) does not oscillate at the singularity. It does, however, have more difficulty escaping from it and lags the SPM tracker by several pixels over several frames of measurements.

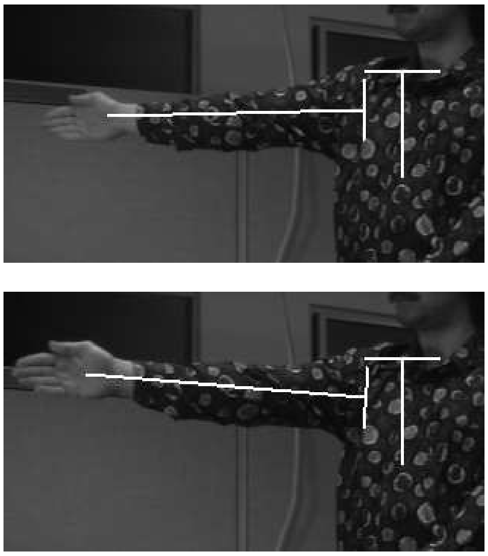


Figure 5: Test sequence. Frames 15 (a) and 36 (b) from the test sequence for singularity comparison, showing the 2-D SPM estimates.

In a real application, an algorithm such as Levenberg-Marquardt would be used to automatically adapt the amount of damping. It is clear, however, that any 3-D tracker will be forced to do a significant amount of work in the vicinity of the singularity to avoid poor performance. Unfortunately, in spite of this effort the 3-D tracker will be quite sensitive to both image noise and errors in the kinematic model parameters, such as link lengths, during this part of the trajectory.

Figure (6b) shows the out-of-plane rotation angle, ϕ, for the two 3-D models. The divergence of the two curves following the singularity is a consequence of the usual orthographic ambiguity. Plot (c) shows the in-plane rotation angle, θ, which is essentially the same for all of the models. In summary, the 2-D SPM exhibits more consistent and uniform registration performance, as expected. Performance of the 3-D model depends critically on determining the correct amount of damping.

5.2 Motion Capture from Movies

For the second experiment, we developed a 2-D SPM for the human figure and applied it to a short dance sequence by Fred Astaire. Figure (7) shows stills from the movie "Shall We Dance", overlaid with their associated state estimates. The overall quality of the registration is fairly good, especially considering the low contrast between figure and background. The tracker slipped off the right leg around the third frame, due probably to low contrast, but managed to get back on over the next frame. Finally, the tracker fails in the last frame due to self-occlusion.

As a preprocessing step, we used standard image stabilization techniques [6] to

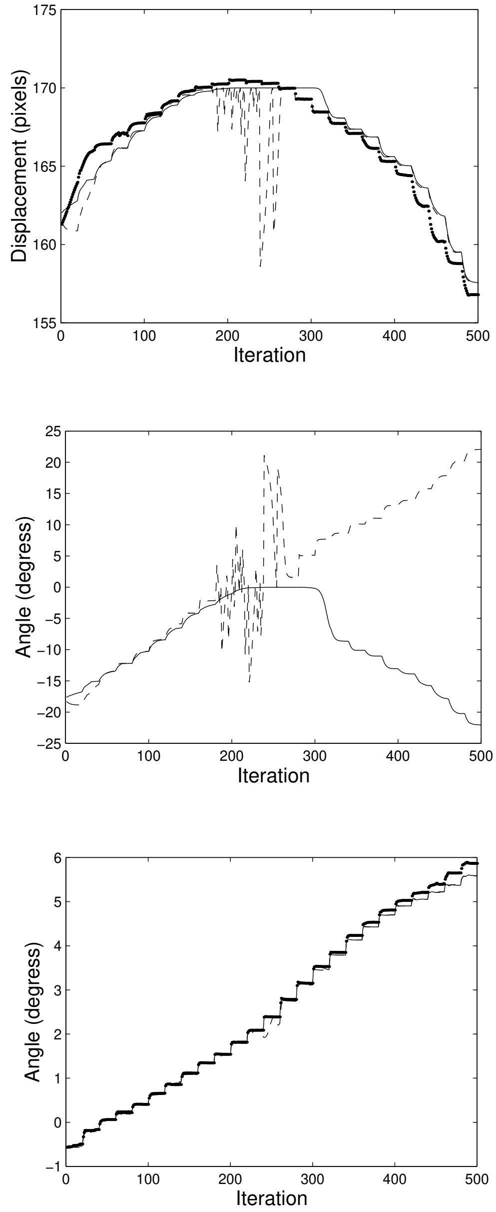


Figure 6: Tracking results for 2-D SPM and 3-D kinematic models using the motion sequence in Figure (5). 2-D SPM data is shown by large dots, while 3-D model data is shown by a solid curve in the well-damped case and a dashed line in the under-damped case. (a) (Top) Displacement in pixels corresponding to the length of the arm link after projection into image plane using the estimated state. (b) (Middle) angle ϕ of 3-D trackers, and (c) (Bottom) in-plane rotation θ of each model.

compensate for camera motion and generate a background image that does not contain the moving figure. Both the initial state of the kinematic model and the SSD template model were initialized by hand on the first frame of the sequence.

The combination of figure tracking and recovery of a background image enables some simple video edits. The first is to remove Fred from the background of the original sequence and replay his motion against a new background. Stills from the resulting video clip are displayed in Figure 8. In this example, a photograph of the second author's office at CRL provides a new backdrop for the original dance sequence.

The compositing technique used to generate the new sequence is a straightforward application of the forward kinematic model, using the estimated state of the figure in each frame. The forward model contains a set of templates initialized on the first frame of the sequence. Each state estimate produces a specific configuration of these templates in the image plane. Our composition algorithm simply replaces pixels in the background image with template pixels to synthesize a new frame.

In this example, the segmentation of the templates pixels is quite crude. We use a rectangular bounding box to identify the pixels associated with each link in the kinematic model. While this approximation is adequate for tracking, it leads to artifacts in the synthesized images. There are two forms of artifacts. The first are segmentation errors in which the template contains pixels from the background of the initial frame in the sequence. The second artifact occurs at the joints, where the template model fails to capture the visual continuity of the figure across links. Both of these artificts can be eliminated through the use of image compositing and blending techniques. This will be addressed in future work.

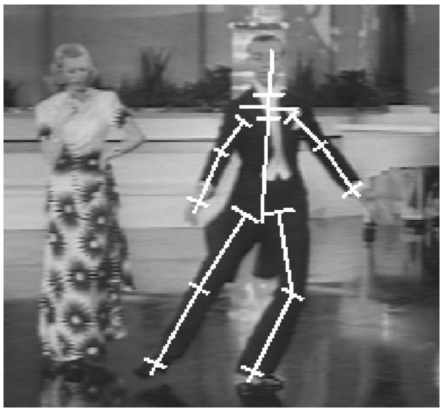
The second edit we performed was to replace the pixels in Fred's template model with pixels from the image of another person. This allows us to animate the image of a different person with the original motion against the background image from the video. The result is the illusion of a different person performing the dance. This is illustrated in Figure 9, which shows stills from a sequence in which the director of CRL, Dr. Robert Iannucci, dances under the watchful eye of Ginger Rogers.

This last edit is an example of *motion transfer*, using the pixel motion from one video sequence to directly animate the pixel content of another image (or set of image). Here we are making an analogy to view transfer, in which a set of images is "transferred" to a new camera viewpoint. If 3-D motion information is available, more complex motion transformations are possible. See [21] for an example.

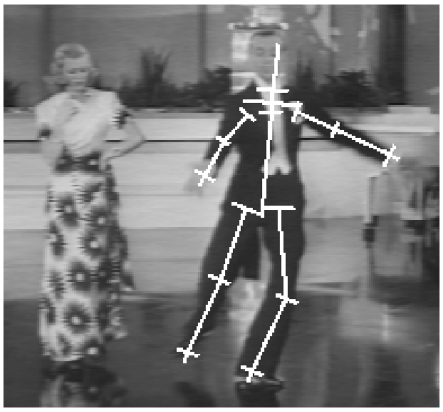
From the standpoint of video editing, we can view the tracking process as an automatic extrapolation across the image sequence of the segmentation provided by the user in the first frame.

6 Future Work

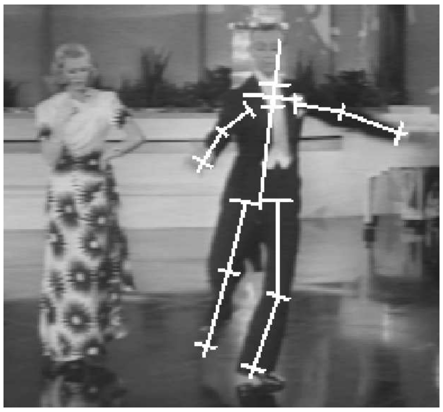
In future work, we plan to address the 3-D reconstruction of fixed and variable kinematic parameters from a sequence of SPM estimates. We are also interested in using more sophisticated estimation techniques such as multiple hypothesis tracking [16] to compensate for self-occlusions, background clutter, and photometric variation.



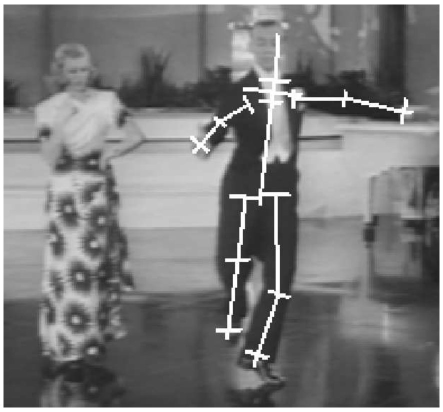
(a)



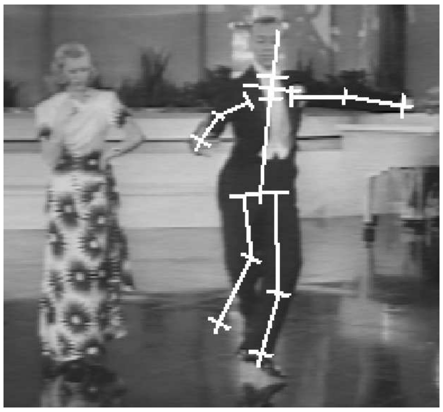
(b)



(c)



(d)



(e)

Figure 7: Fred Astaire tracked in an image sequence using the SPM-based tracker. Images (a)–(e) correspond to frames 0,4,5,6 and 7 from the input sequence.

(a)

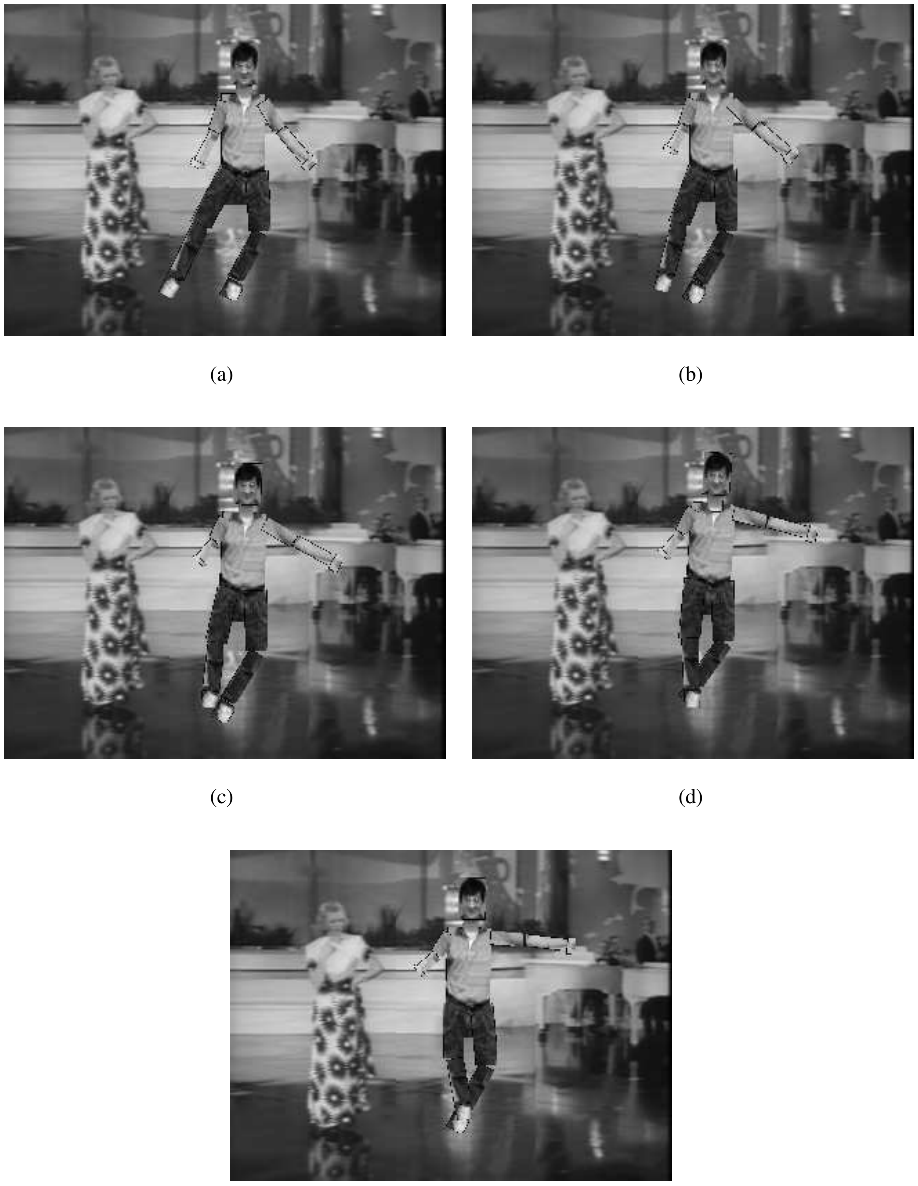
(b)

(c)

(d)

(e)

Figure 8: First video edit. Obtained by synthesizing a sequence of Fred Astaire's motion against an office background. The figure poses in (a)–(e) correspond to input frames 0,2,4,6 and 7.



(a) (b) (c) (d) (e)

Figure 9: Second video edit. Obtained by replaying the motion estimates using template pixels from a second person. The background was obtained from the original sequence. The figure poses in (a)–(e) correspond to input frames 0,2,4,6 and 7.

7 Conclusions

While kinematic models provide powerful constraints for gradient-based tracking algorithms, we have shown that trackers utilizing 3-D kinematic models suffer from singularities when motion is directed along the viewing axis of a single camera. This results in greater sensitivity to noise and possible loss of registration.

We have introduced a 2-D Scaled Prismatic Model which captures the image plane motion of a large class of 3-D kinematic models. The SPM has the following three advantages. First, it has fewer singularity problems than 3-D kinematic models. In addition, unlike the general 3-D model, its singularities can be fully characterized enabling it to be used only in appropriate situations. Second, the SPM does not require the specification of link lengths and joint axes, which can sometimes be difficult. In cases where 3-D information is unnecessary the SPM alone may provide sufficient motion estimation. Third, when 3-D motion estimates are desired, they can be obtained from SPM motion estimates using a batch estimation approach.

We have used the SPM to track Fred Astaire in a video clip taken from the movie "Shall We Dance". We have demonstrated some simple applications of this tracking technology to video editing.

References

[1] C. Bregler and J. Malik. Estimating and tracking kinematic chains. In *Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998. Accepted for publication.

[2] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[3] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multiview approach. In *Computer Vision and Pattern Recognition*, pages 73–80, San Fransisco, CA, 1996.

[4] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. In *Proceedings of International Conference on Computer Vision*, pages 764–770, Cambridge MA, June 20–23 1995.

[5] D. Hogg. Model-based vision: a program to see a walking person. *Image Vision Computing*, 1(1):5–20, 1983.

[6] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, 1994.

[7] I. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. In *Proceedings of International Conference on Computer Vision*, pages 618–623, Cambridge, MA, June 1995.

[8] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Computer Vision and Pattern Recognition*, pages 81–87, San Fran., CA, June 18–20 1996.

[9] Y. Nakamura. *Advanced Robotics: Redundancy and Optimization*. Addison-Wesley, 1991.

[10] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):522–536, 1980.

[11] D. K. Pai. *Singularity, Uncertainty, and Compliance of Robot Manipulators*. PhD thesis, Cornell U., Ithaca, NY, May 1988.

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

[13] J. M. Rehg. *Visual Analysis of High DOF Articulated Objects with Application to Hand Tracking*. PhD thesis, Carnegie Mellon University, April 1995. Available as School of Computer Science tech report CMU-CS-95-138.

[14] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *ECCV*, volume 2, pages 35–46, Stockholm, Sweden, 1994.

[15] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of Fifth Intl. Conf. on Computer Vision*, pages 612–617, Boston, MA, 1995.

[16] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, AC-24(6):843–854, December 1979.

[17] M. J. B. S. X. Ju and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Intl. Conf. Automatic Face and Gesture Recognition*, pages 38–44, Killington, VT, 1996.

[18] R. Sharma and S. Hutchinson. Motion perceptibility and its application to active vision-based visual servoing. *IEEE Transactions on Robotics and Automation*, 13(4):607–617, Aug. 1997.

[19] M. Spong. *Robot Dynamics and Control*. John Wiley and Sons, 1989.

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

[21] A. Witkin and Z. Popovic. Motion warping. In *Proc. of SIGGRAPH 95*, pages 105–108, Los Angeles CA, August 6–11 1995.

[22] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *Computer Vision and Pattern Recognition*, pages 664–665, 1991. Also see ETL TR 90-46.

digital™

Singularities in Articulated Object Tracking with 2-D and 3-D Models

James M. Rehg Daniel D. Morris

CRL 97/8
October 1997