COMPAQ

Recovery of 3D Articulated Motion from 2D Correspondences

David E. DiFranco *Tat-Jen Cham* *James M. Rehg*

Cambridge Research Laboratory
Technical Report Series

CRL 99/7
December 1999

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.
Vice President, Corporate Research

Recovery of 3D Articulated Motion from 2D Correspondences

David E. DiFranco
Massachusetts Institute of Technology
Cambridge, MA 02139
Tat-Jen Cham James M. Rehg
Cambridge Research Laboratory
Compaq Computer Corporation
Cambridge, MA 02139

December 1999

Abstract

We present a method for computing the 3D motion of articulated models from 2D correspondences. An iterative batch algorithm is proposed which estimates the maximum aposteriori trajectory based on the 2D measurements subject to a number of constraints. These include (i) kinematic constraints based on a 3D kinematic model, (ii) joint angle limits, (iii) dynamic smoothing and (iv) 3D key frames which can be specified the user. The framework handles any variation in the number of constraints as well as partial or missing data. This method is shown to obtain favorable reconstruction results on a movie dance sequence.

Authors email: ddif@mit.edu, tjc@crl.dec.com, rehg@crl.dec.com

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

Compaq Computer Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700
Cambridge, Massachusetts 02139 USA

1 Introduction

Video is the primary archival source for human movement, with examples ranging from sports coverage of Olympic events to dance routines in Hollywood movies. The ability to reliably track the human figure in movie footage would unlock a large, untapped repository of motion data. However, the recovery of 3D figure motion from a single sequence of uncalibrated video images is a challenging problem.

In contrast to the single-view case, 3D tracking of articulated objects from multiple camera views has been addressed by a number of authors [12, 14, 10, 1]. 3D kinematic models are used in these works to define the state space and constrain image measurements. Tracking proceeds by differentially adjusting the state using a nonlinear least squares (NLS) algorithm, until the projections of the kinematic model are aligned with the feature measurements in all of the images. The basic framework is described in [14].

In the multi-view case the tracker is solving two problems simultaneously: registering the model with all of the images, and reconstructing the 3D pose of the figure from 2D measurements. With an adequate set of viewpoints, the state space is fully observable and the state estimate will remain near the correct answer. In this case, the 3D kinematics provide a powerful constraint on image motion, simplifying the registration task.

When only a single camera viewpoint is available, however, ambiguities can arise in the reconstruction of 3D pose under orthographic projection. The standard reflective ambiguity results in a pair of solutions for the rotation of a single link out of the image plane [16]. In addition, kinematic singularities arise when the out-of-plane rotation is zero [11]. The resulting loss of rank in the kinematic Jacobian complicates the use of NLS tracking algorithms.

One solution, proposed in [11], is to decouple the registration and reconstruction stages. Motion of the figure in the image plane is described by a 2D Scaled-Prismatic model (described later in section 7.1). Tracking with the SPM model registers the figure with the image sequence, but defers the inference of 3D pose. The 3D reconstruction problem can then be formulated as a batch optimization over a series of SPM measurements. Within a batch formulation, it is easy to combine the entire set of image measurements with additional constraints in order to resolve ambiguities in 3D reconstruction.

The paper describes a batch framework for 3D reconstruction using SPM measurements. In addition to kinematic constraints, we explore the use of three other types of constraints: dynamic models, joint angle limits, and 3D key frames, in resolving ambiguities in the estimated 3D pose of the figure. This paper makes a number of contributions: (1) It characterizes the space of ambiguous 3D solutions associated with a set of 2D SPM measurements. (2) It contains a complete derivation of the equations for batch 3D reconstruction using SPM measurements and additional constraints. (3) It presents experimental results on the reconstruction of 3D motion from a Fred Astaire dance sequence.

1.1 Problems in 3D Tracking

There are two sources of ambiguity which make 3D figure tracking a difficult problem:

1. *Ambiguity in determining 2D model-to-image correspondences*.
2. *Ambiguity in reconstructing 3D pose*.

The 2D correspondence ambiguity arises from a variety of sources including background clutter and imperfect features in the image. For example, optic-flow estimation based on image gradients may fail when image motion is large, because of the nonlinearity of image structure.

Additionally, correspondence data alone from a single view is insufficient for 3D reconstruction in that it does not encode 3D depth or orientation. Consider the problem of inferring the pose of a 3D articulated chain in a calibrated orthographic view. The projection of each link is consistent with two possible 3D poses because of reflective ambiguity. This problem is not easily solved by simply maintaining multiple solutions [16] as the number of possible solutions grows exponentially with additional links.

Note that these two sources of ambiguity are significantly nonlinear. In particular, solving 2D correspondence ambiguity has to be cast as a *search problem*, i.e. there is no known formulaic function which computes correspondences from a vector of image pixels. For the problem of solving 3D reconstruction, such a direct computation may well exist but would contain numerous ambiguities; the problem in this case is selecting the correct solution for a list of candidates.

Most of the previous work for tracking a figure in a video sequence involves an online linear framework for tracking a 3D kinematic model directly from image data. However, it is clear from the list of ambiguities above that the problem is significantly nonlinear, and the performance of such tracking methods is therefore highly dependent on the how well the problem can be linearized. We believe that the interaction of the nonlinearities in the two ambiguities causes the problem to be difficult to linearize directly in 3D state-space. Furthermore, the search efficiency for image features from 3D state-space becomes increasingly poor in the vicinity of kinematic singularities.

This paper then takes the same view as in [11] that tracking should be separated into the two processes of (i) 2D registration, and (ii) 3D motion recovery to properly cope with the nonlinear ambiguities. In this paper, only the 3D motion recovery problem is considered, and 2D correspondences are assumed to be available (section 7.1 discusses a method for obtaining these automatically).

2 3D Motion Recovery

The problem of 3D motion recovery can be approached from a signal reconstruction perspective. Suppose the 3D state for an articulated structure is represented by the concatenation of all 3D pose parameters for all links in the structure. Next consider a set of such states representing the 'true' 3D motion sequence of the structure. The observed states are the result of each true state undergoing successive degradation through one to three *lossy channels* (see figure 1):

1. *Noise*. Noise is added to the true 3D states.
2. *Projection*. Additionally, depth information is removed from some states through perspective projection.
3. *Deletion*. Partial or full data of some states are deleted, e.g. in the case of partial or full occlusion or dropped frames.

The goal is therefore to recover the true states from the set of available observed states.

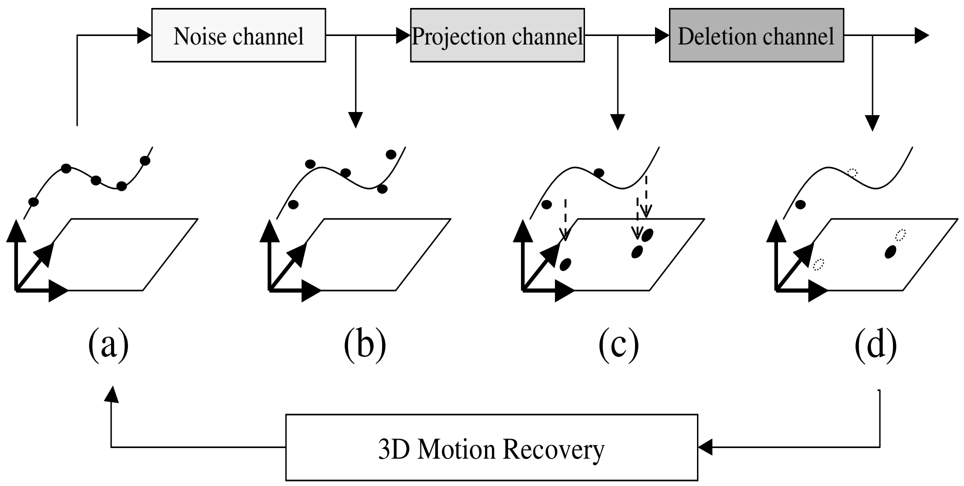

Figure 1: A channel-based model of the data degradation process. (a) shows the true 3D trajectory with discrete states. (b) noise is added to the states. (c) some states are projected onto the image plane, losing depth information. (d) shows the final set of observed states after deletion of more states. The goal is to recover the true states from the set of available observed states.

The separation of these degradation channels is useful for formulating a *unified framework* for seamlessly handling a large range of scenarios with different data degradation – e.g. from smoothing of noisy motion capture data with dropped frames, to estimating 3D figure motion from a video sequence with multiple occlusion events.

3 Constraints for 3D Inference

The difficult problem of interest is that of inferring the 3D states from 2D measurements (projection channel degradation), which is inherently ill-posed. In order to regularize this problem, we need to utilize a number of constraints, which are

- Kinematic constraints
- Joint angle limits
- Dynamic smoothing

- 3D key frames

The most important constraints are the **kinematic constraints**. Kinematic constraints enforce connectivity between adjacent links, link length constancy as well as restrict joint motion to rotation about a fixed local axes in the case of revolute joints. These constraints are hard constraints and automatically enforced when estimation is done in the state-space of a 3D kinematic model.

Of particular note is that simply applying a 3D kinematic model to 2D measurements restricts the solution to a number of isolated candidate regions in the kinematic state-space (modulo depth of base link under orthography). These correspond to the discrete combinations of 3D reflective ambiguities at each link (see figure 2) mentioned in section 1.1. A candidate solution can be computed in each region using an iterative linear algorithm.

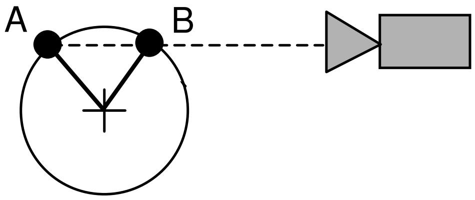

Figure 2: 3D reflective ambiguity. The figure shows a revolute link which can rotate in a full circle. From the camera position shown, it is impossible to distinguish pose A from pose B based only on the link projection.

However as mentioned earlier, a multiple-hypothesis or Viterbi-decoding scheme is generally infeasible because the number of such regions increases exponentially with the number of links. Hence the constraints listed below are also required to bias the solution to the correct candidate region using an iterative scheme.

Joint angle constraints specify the limits to which joints can rotate about their corresponding axes. **Dynamic smoothing models** describe the probability of a particular state based on past and future states, and are generally used to bias continuous rather than abrupt 3D motion. **3D key frames** are 3D states which are interactively established by the user, and are equivalent to observed states undergoing only noise channel degradation. The formulation of these constraints is further described below. In order to take advantage of key frames and smoothing dynamics within the sequence, the 3D estimation is done in a *batch framework* described in section 5.

Note that these constraints are not limited to 3D inference from 2D measurements. They can also be applied to 3D measurements and partial data (noise and deletion channel degradation). Further note that while these constraints can be applied as intrinsic hard constraints which must be satisfied, it is more flexible to set them as soft constraints. These would affect the estimation by adding components to the overall residual equation.

4 Measurements and Constraints on Kinematic States

In the following sections we describe the measurements and constraints used in our batch estimation framework. In particular, we will derive residual equations which can be expressed in the form of

$$\boldsymbol{G}(\boldsymbol{q}) = \boldsymbol{w} + \boldsymbol{\epsilon}$$

where $\boldsymbol{q}$ is the kinematic state to be computed, $\boldsymbol{G}(\cdot)$ is a linearizable function, $\boldsymbol{w}$ is a measurement vector, and $\boldsymbol{\epsilon}$ is the residual vector which is to be minimized. These equations can then be then be combined together and solved simultaneously as described in section 5.

4.1 2D Measurements

As mentioned in section 1, the image plane projections of links in an articulated structure can be recovered from image data using a 2D SPM tracker. This allows us to express 2D measurements at a more abstract level than raw pixel data.

In our framework, the image positions of the joints are used. The 2D measurement can then be expressed in the estimation framework as

$$\boldsymbol{P}\,\boldsymbol{X}_j(\boldsymbol{q}_t) = \boldsymbol{x}_{jt} + \boldsymbol{\epsilon}_{jt} \tag{1}$$

where $\boldsymbol{q}_t$ is the kinematic state for the tth time frame, $\boldsymbol{X}_j(\cdot)$ is the forward kinematic function for computing the 3D position of the jth joint, $\boldsymbol{P}$ is the image plane projection matrix, $\boldsymbol{x}_{jt}$ is the observed image position of the joint, and $\boldsymbol{\epsilon}_{jt}$ is the measurement noise. Since $\boldsymbol{X}_j(\cdot)$ is nonlinear, (1) has to be relinearized at each iteration.

4.2 Joint Angle Constraints

One way to limit our solution for 3D motion to a physically valid result is to incorporate limits on the range of joint angles for our kinematic model. For example, a human elbow can only rotate through about 135 degrees; it is advantageous to use this knowledge to obtain a plausible solution for 3D motion. For example as shown in figure 3, knowing the forbidden interval for the joint angle of the link allows unambiguous selection of pose B.

To incorporate limits on the range of revolute joint angles, we introduce inequality constraints such as $q_j \geq \theta_j$ where q_j is the jth revolute angle parameter and θ_j is the fixed lower limit for q_j. A method of implementing this is to use a slack variable λ_j to obtain an equation $q_j - \theta_j - \lambda_j^2 = 0$ as described in [5, 6]. This method is preferred to applying traditional differentiable barrier functions in the region of the limit, as it does not discourage the joint angle from reaching and staying at the limit. This typically occurs in human motion, e.g. when limbs are straightened.

However, for each limit specified, an additional variable is introduced into the estimation. This causes a significant rise in computation time as the state-space dimension will increase approximately threefold. Further analysis shows that an almost identical effect can be created by using a first-order discontinuous residual function $\epsilon(q_j)$ of the

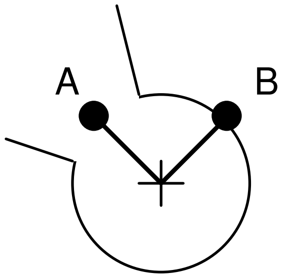

Figure 3: Disambiguation from joint angle limits. The joint angle limits prevents the selection of pose A leaving pose B as the only possibility.

form

$$\epsilon(q_j) = \begin{cases} 0, & q_j \leq \theta_j \\ k(q_j - \theta_j) & q_j > \theta_j \end{cases} \tag{2}$$

where k is a constant factor determining the strength of the inequality constraint. $\epsilon(q_j)$ is added to the overall residual equation for minimization, and the regime in effect is determined according to the value of q_j at each iteration. Because this barrier function remains zeroth-order continuous, there is negligible solution instability at the limits.

4.3 Dynamics

Dynamics are used to express the greater probability for a particular form of motion, e.g. the typical preference for smooth continuous motion compared to abrupt motion (see figure 4). There are many different variants of dynamic models, ranging from simple hand-constructed constant velocity models to complex switching models automatically learned from data [13]. The typical application of dynamic models in tracking is for forward prediction in the context of the Fokker-Planck drift-diffusion. However, dynamic models also can be expressed in an interpolating, or *smoothing* manner. This is particularly useful in a batch framework where the estimation of states in all time frames is done simultaneously.

For general linear models, we can express the dynamic constraints

$$\boldsymbol{Q} - \boldsymbol{A}\boldsymbol{Q} = \boldsymbol{0} + \boldsymbol{\epsilon}_d \tag{3}$$

where

$$|\boldsymbol{I} - \boldsymbol{A}| = 0, \quad \boldsymbol{A} \neq \boldsymbol{I}$$

Here $\boldsymbol{Q}$ is the vector of concatenated states for all time frames and $\boldsymbol{\epsilon}_d$ is the dynamics process noise; $\boldsymbol{A}$ is the matrix of dynamic coefficients which will satisfy the stated conditions. Equation (3) represents the dynamics component to be added to the overall residual equation.

In our experiments, a second order constant velocity model was used. In this case, the predicted current state $\boldsymbol{q}_t = (\boldsymbol{q}_{t-1} + \boldsymbol{q}_{t+1})/2 + \boldsymbol{\epsilon}_{dt}$, is the mean of immediate past and future states.

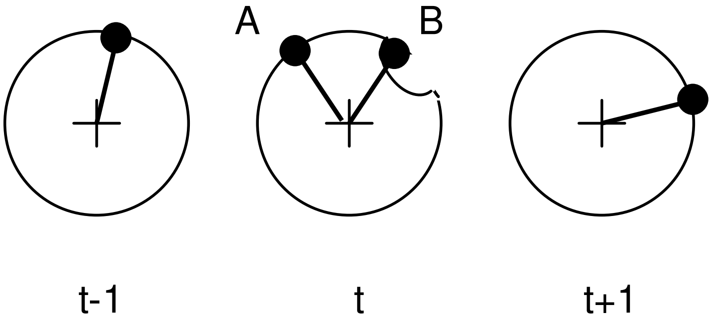

Figure 4: Disambiguation from dynamics. Knowing the approximate poses of the joint at frames $t-1$ and $t+1$ preferentially selects pose B at frame t when dynamics is used to bias towards smooth motion.

The use of even simple dynamic prediction significantly helps in eliminating incorrect sets of hypotheses due to 3D reflective ambiguities. While more accurate learned models are preferred if available, they unfortunately require vast amounts of training data for modeling such that intra-class and inter-class variations are captured. This poses a problem for learning 3D human motion models due to a difficulty of obtaining a large volume of data. In [8], a probabilistic model was derived from limited motion-capture data. However, it is difficult to imagine such a model will generalize well, especially considering that most motion-capture systems currently require individuals to don cumbersome apparatus which hinder the naturalness of the motion.

4.4 3D Key Frames

Despite the application of kinematics, joint angle limits and dynamic smoothing, 3D motion recovery is generally still underconstrained. Instead of attempting to solve the hard problem of using additional constraints such as shading cues, an intermediate solution is to access manual aid and allow an interactive user to set 3D key frames. See figure 5.

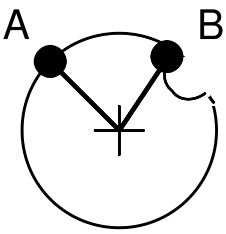

Figure 5: Disambiguation from key frames. A key frame would specify the approximate pose of the joint, which in this case is located near pose B. Hence pose B is selected. Note that the key frame does not need to be exact.

Since these 3D key frames are inherently noisy, we treat these as noise channel degraded observations:

$$\boldsymbol{q}_t = \boldsymbol{q}_t^* + \boldsymbol{\epsilon}_{kt} \tag{4}$$

where $\boldsymbol{q}_t^*$ is a key frame observed at frame t, and $\boldsymbol{\epsilon}_{kt}$ is the measurement noise of the key frame.

For greater generality, we also allow the specification of *partial* key frames, in which only some state parameters are established. For example this may be used to disambiguate the angles of one joint in a human figure model if this is the only ambiguous limb. In the context of (4), the unestablished state parameters will have infinite variance.

In an interactive setting, the user will initially apply the solver with a minimal number of key frames, e.g. at the start and the end of the sequence and potentially problematic frames with departure from the expected dynamics. Any resulting gross estimation errors may be corrected by introducing additional key frames and reapplying the solver.

5 3D Batch Framework

Our 3D batch framework involves iterative least squares and solving for the 2D measurements and other constraints simultaneously in all time frames. Assuming that the noise in the measurements and constraints are Gaussian distributed, the framework computes the maximum aposteriori (MAP) estimate

$$\hat{\boldsymbol{Q}} = \arg\max_{\boldsymbol{Q}} \{p(\boldsymbol{Q}|\boldsymbol{Z})\}$$

for the full trajectory state $\boldsymbol{Q}$ (consisting of the states in all time frames) given the 2D measurements $\boldsymbol{Z}$. In this case, the priors are the constraints which are applied to the estimation.

Note that our batch framework applies dynamic models differently than the standard Kalman and forward-backward smoothing filters. Therefore the results obtained are different – the filters compute a sequence of marginal MAP states based on $p(\boldsymbol{q}_t|\boldsymbol{Z})$, which is not the same as the t state in the MAP trajectory $\hat{\boldsymbol{Q}}$ computed by our framework.

Measurements and constraints are incorporated in the batch estimation by merging all residual equations expressed in (1), (2), (3), and (4) for all time frames in the form of

$$\boldsymbol{J}^{\boldsymbol{T}}\Sigma^{-1}\boldsymbol{J}\, d\boldsymbol{Q} = \boldsymbol{J}^{\boldsymbol{T}}\Sigma^{-1}\boldsymbol{R} \tag{5}$$

where $\boldsymbol{J}$ is the overall Jacobian, $\boldsymbol{R}$ is the overall measurement vector, and Σ is the measurement covariance. The matrix $\boldsymbol{J}^{\boldsymbol{T}}\Sigma^{-1}\boldsymbol{J}$ is block-diagonal and grouped according to time frames. We further add a stabilization term kI to the $\boldsymbol{J}^{\boldsymbol{T}}\Sigma^{-1}\boldsymbol{J}$, where k is a constant.

Note that the 2D measurements, joint angle limits, dynamics and 3D key frames are represented as rows in $\boldsymbol{J}$ and treated in the same unified manner by the framework. The allows great flexibility, e.g. for including as many 3D key frames as required, or even changing constraints on the fly. Handling partial missing data simply involves zeroing some of the entries in Σ^{-1}.

Finally, (5) is solved iteratively using the Gauss-Newton least-squares method with a sparse-matrix inversion routine.

6 Results

One of our key goals is to recover 3D human motion from video footage, and in this paper we present results from a Fred Astaire dance sequence. The top row of figure 6 shows four frames from the test sequence, superimposed with 2D SPM measurements. These 2D SPM measurements have to be manually specified as none of the current trackers are able to track successfully from video when there is significant 3D body rotation, which occurs in our test sequence. These 2D measurements for the 14-frame sequence are used as input into our 3D estimation framework. Two 3D key frames are manually specified at the start and end of the sequence. The estimation involved running the algorithm for 20 iterations taking a total of 27 seconds. The final output was imported into 3D Studio Max and rendered.

The middle row of figure 6 shows the corresponding reconstruction of the frames rendered from approximately the original camera viewpoint. The bottom row show the same reconstruction but from a viewpoint which is up and left of the original camera position.

The results reflect a highly credible reconstruction of the original 3D motion. However, one artifact is the small inter-penetration of the legs in the middle of the sequence. This is because the thickness of the body parts were unaccounted for in the estimation framework, and we currently do not place any constraints on limbs inter-penetrating. One other noticeable artifact is that head rotation about the spinal axis is not recovered and hence the facial direction cannot be reconstructed.

7 Previous Work

Many researchers have tackled the problem of tracking 3D articulated models with multiple cameras. Rehg [15] tracked hands using an extended Kalman filter framework. O'Rourke and Badler [12] and Gavrila and Davis [3] used multiple cameras to obtain 3D positions of the human body, while Bregler and Malik [1] used Kalman filtering to exploit dynamic constraints.

To obtain a less complete, but still useful, interpretation of motion, many researchers have attempted tracking in 2D from a single camera. Hogg [7] and Bregler and Malik [1] studied the case of the human walking parallel to the image plane, which limits the solution to two dimensions. Hel-Or and Werman [6] applied joint constraints to find 2D in-plane motion, in both Kalman filter and batch solutions. Other papers allow motion out of the plane of view, but only attempt to fit a 2D model to the image stream [9]. Morris and Rehg [11] both used 2D models with prismatic joints to do this. Such tracking data may be useful for classification of 3D motion, but it is inadequate for true 3D motion analysis.

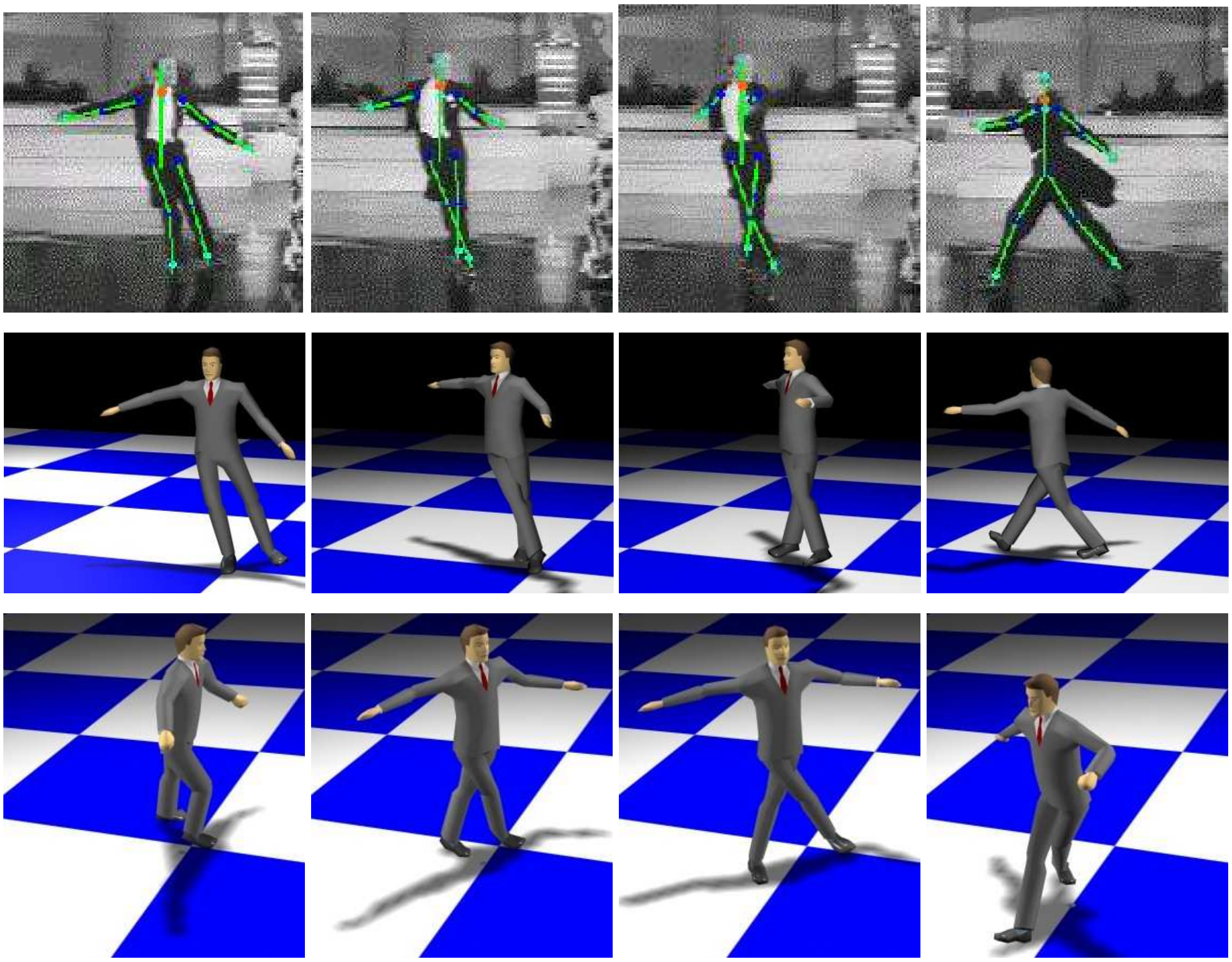

Figure 6: Top: Manually specified 2D SPM measurements for Fred Astaire in four frames. Middle: 3D state estimates produced by batch estimation algorithm, shown from original camera viewpoint. Bottom: 3D state estimates shown from a different camera viewpoint.

Few attempts have been made to capture 3D motion from a single image stream. Goncalves et al. [4] tracked a human arm in a very constrained environment with minimal reflective ambiguity. Shimada et al. [16] capture hand motion from one camera, using Kalman filtering and sampling the solution probability space. They exploit join constraints by truncating the probability space. The strength of joint constraints in the hand model helped make this possible (e.g. finger joints can only rotate approximately 90 degrees). Howe et al. [8] also recover 3D position of a human figure, but with limited movement out of the plane of vision and no body rotation.

7.1 2D SPM Tracking

In order to obtain 2D joint positions from image data, we describe the figure projection using a *scaled prismatic model* (SPM), introduced in [11]. The model enforces 2D constraints on figure motion that are consistent with the underlying 3D kinematic model.

Each link in a scaled prismatic model describes the image plane appearance of an associated rigid link in an underlying 3D kinematic chain. Each SPM link can rotate and translate in the image plane. The rotation degree of freedom (dof) captures the projected link orientation of revolute joints in the 3D model. The translation dof captures the foreshortening that occurs when 3D links rotate into and out of the image plane. The figure can then be modeled in 2D as a branched chain with arms, legs, and head modeled as SPM links. A complete discussion of SPM models, including a derivation of the SPM Jacobian and an analysis of its singularities, can be found in [11].

Each SPM link is associated with a template representation of its appearance. Tracking then involves minimizing the difference between the projected template and the corresponding pixels in the image frame. The multiple-hypothesis statistical framework proposed in [2] generates a probabilistic representation for the SPM state which can eventually be integrated with the current 3D estimation framework. However, as the template representation for a torso does not cope well with out-of-plane rotation, automated tracking is generally limited to motion without significant torso rotation.

8 Summary and Future Work

We presented a method for recovering the 3D motion of articulated models from a sequence of 2D SPM measurements. It exploits a number of constraints including kinematic constraints, joint angle limits, dynamic smoothing and 3D key frames. The equations for these constraints were derived and integrated into a 3D batch estimation framework. The estimation framework is flexible and can easily cope with variation in the number of constraints applied, and also with partial or missing data. The goal of this work is to be able to apply the method to reconstructing figure motion from video footage.

The favorable reconstruction results shown for a Fred Astaire dance sequence illustrate the capability of using multiple constraints to reduce 3D ambiguity. However one problem encountered is the partial inter-penetration of limbs in the sequence.

For the future, we intend to add further constraints to our framework. This includes volume exclusion constraints to avoid inter-penetration of links, as well as making use of self-occlusion cues to further help disambiguate 3D pose. We also plan to enhance the estimation framework to cope with remaining unfiltered ambiguities, possibility using a multiple hypothesis statistical framework. Finally, we will explore ways to fully automate the process of video to 3D figure motion recovery. This will include the interleaving of the 3D estimation framework with 2D tracking to improve both the robustness of 2D registration and the quality of 3D reconstruction.

References

[1] C. Bregler and J. Malik. Estimating and tracking kinematic chains. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 8–15, Santa Barbara, CA, 1998.

[2] T.-J. Cham and J.M. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume II, pages 239–245, Fort Collins, Colorado, 1999.

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

[4] L. Goncalves, E.D. Bernado, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *Proc. Int. Conf. on Computer Vision*, pages 764–770, Cambridge, MA, 1995.

[5] W.E. Grimson. On the recognition of curved objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):632–643, 1989.

[6] Yacov Hel-Or and Michael Werman. Constraint fusion for recognition and localization of articulated objects. *Int. Journal of Computer Vision*, 19(1):5–28, 1996.

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

[8] Nicholas Howe, Michael Leventon, and William Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *Neural Information Processing Systems*, Denver, Colorado, Nov 1999.

[9] Shannon X. Ju, Michael J. Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Intl. Conf. Automatic Face and Gesture Recognition*, pages 38–44, Killington, VT, 1996.

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

[11] Daniel D. Morris and James M. Rehg. Singularity analysis for articulated object tracking. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 289–296, Santa Barbara, CA, June 23-25 1998.

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

[13] V. Pavlović, J.M. Rehg, T.J. Cham, and K.P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *Proc. Int. Conf. on Computer Vision*, volume I, pages 94–101, Corfu, Greece, 1999.

[14] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *Proc. European Conference on Computer Vision*, pages II: 35–46, Stockholm, Sweden, 1994.

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

[16] Nobutaka Shimada, Yoshiaki Shirai, Yoshinori Kuno, and Jun Miura. Hand gesture estimation and model refinement using monocular camera— ambiguity limitation by inequality constraints. In *Proc. 3rd Int. Conf. Automatic Face and Gesture Recognition*, pages 268–273, Nara, Japan, 1998.

Recovery of 3D Articulated Motion from 2D Correspondences

David E. DiFranco Tat-Jen Cham
James M. Rehg

CRL 99/7
December 1999