3D Reconstruction of a Smooth Articulated Trajectory from a Monocular Image Sequence *

Hyun Soo Park and Yaser Sheikh
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA, 15213
{hyunsoop,yaser}@cs.cmu.edu

Abstract

An articulated trajectory is defined as a trajectory that remains at a fixed distance with respect to a parent trajectory. In this paper, we present a method to reconstruct an articulated trajectory in three dimensions given the two dimensional projection of the articulated trajectory, the 3D parent trajectory, and the camera pose at each time instant. This is a core challenge in reconstructing the 3D motion of articulated structures such as the human body because endpoints of each limb form articulated trajectories. We simultaneously apply activity-independent spatial and temporal constraints, in the form of fixed 3D distance to the parent trajectory and smooth 3D motion. There exist two solutions that satisfy each instantaneous 2D projection and articulation constraint (a ray intersects a sphere at up to two locations) and we show that resolving this ambiguity by enforcing smoothness is equivalent to solving a binary quadratic programming problem. A geometric analysis of the reconstruction of articulated trajectories is also presented and a measure of the reconstructibility of an articulated trajectory is proposed.

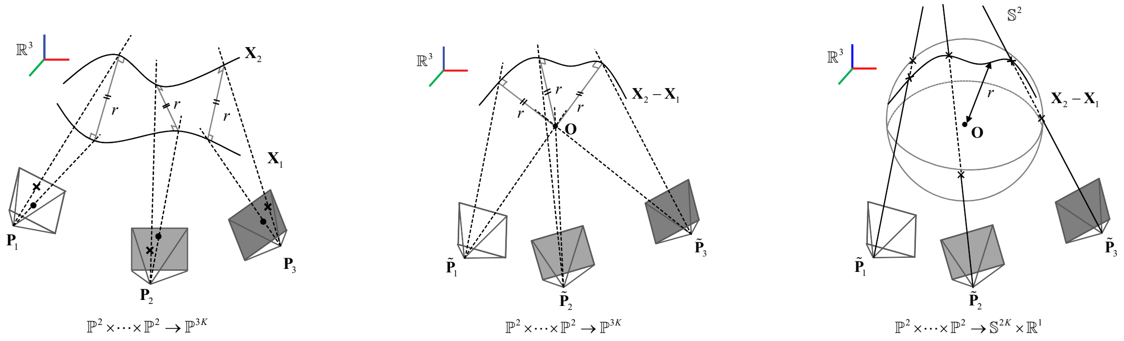
1. Introduction

Reconstructing a moving point in three dimensions from a sequence of two dimensional projections is an ill-posed problem; any point on the line of projection connecting the camera's optical center and an image measurement can be a solution. Yet, humans can effortlessly perceive depth if the 2D points correspond to articulations of a known skeleton [21]. We study the conjecture that if 3D points move smoothly with a known articulation structure, then it is possible to reconstruct their 3D locations from their 2D projections — without any activity-specific prior. The reconstruction of an *articulated* trajectory has a fundamental ambiguity because there are two intersecting points that satisfy an articulation constraint and an image measurement at each time instant [23]: for a 2D trajectory of F frames, there are 2^F 3D trajectories that remain at fixed distance to a parent trajectory[1]. The reconstruction of a *smooth* trajectory without spatial constraints is also known to be fundamentally ambiguous when the camera trajectory is smooth [26, 28]. We present an algorithm to reconstruct a smooth articulated trajectory in 3D by *simultaneously* applying articulation and smoothness constraints. The algorithm takes as input 2D projections of the trajectory, its parent trajectory in 3D, and the camera pose at each time instant. We present a measure of reconstructibility of an articulated trajectory which characterizes the stability of estimation under articulation and smoothness constraints.

Each trajectory is parameterized by coefficients of a trajectory basis in the spherical coordinate system to enforce smoothness and articulation constraints. We show that if a trajectory is embedded in the trajectory basis and articulation constraints are applied, the reconstruction problem is equivalent to a binary quadratic program which is known to be NP-hard [15]. A number of algorithms exist that produce an approximate solution [24, 25, 31] and we use a branch-and-bound method to produce an initialization. We refine the articulated trajectories by minimizing reprojection error. The results are smooth, length preserved 3D trajectories. We have applied our algorithm to recursively reconstruct the 3D motion of a human given the 3D motion of its root. Two general approaches have been explored in prior literature to reconstruct human articulated body motion. Data-driven approaches use repositories of exemplars to overcome the ambiguity [12, 19, 34, 35, 37] and physics-based approaches use dynamical models of the human body to fit to the image stream [11, 36, 38]. Unlike these approaches, our approach reconstructs human motion from *purely geometric constraints*. Thus, the target motion is not confined to pre-

*Project page: http://www.andrew.cmu.edu/user/hyunsoop/iccv2011/iccv_project_page.html

[1] The parent trajectory in a skeleton hierarchy is the proximal trajectory to the root trajectory and the child trajectory is the distal trajectory.



(a) Fixed length trajectory

(b) Relative trajectory

(c) Trajectory on a sphere

Figure 1. (a) An articulated trajectory is defined as a trajectory $\mathbf{X}_2$ which preserves distance from its parent trajectory $\mathbf{X}_1$ across all time instances. (b) The articulated trajectory is transformed to the relative trajectory, $\mathbf{X}_2-\mathbf{X}_1$, by collapsing $\mathbf{X}_1$ to the origin. (c) The articulated trajectory lies on a sphere of radius r. There are two intersecting points at each time instant between the sphere and the ray connecting the camera's optical center and an image measurement, which allow 2^F possible 3D trajectories.

defined activities or view points. We defer a detailed discussion of related work to the end of the paper and present the geometry of the problem and our proposed algorithm in the sequel.

2. Geometry of an Articulated Trajectory

A point trajectory in 3D without any constraint can be represented by a series of points:

$$\mathbf{X}=\left(\begin{bmatrix}x_1\\ \vdots\\ x_F\end{bmatrix},\begin{bmatrix}y_1\\ \vdots\\ y_F\end{bmatrix},\begin{bmatrix}z_1\\ \vdots\\ z_F\end{bmatrix}\right), \tag{1}$$

where (x_i, y_i, z_i) is the Cartesian coordinate of a point at i^{th} time instant and F is the number of frames. If a trajectory is smooth, it is known that the trajectory can be expressed by a linear combination of a compact trajectory basis [2], i.e.,

$$\mathbf{X}=(\boldsymbol{\Theta}\mathbf{a}_x, \boldsymbol{\Theta}\mathbf{a}_y, \boldsymbol{\Theta}\mathbf{a}_z) \tag{2}$$

where $\boldsymbol{\Theta}$ is a $F \times K$ matrix composed of a collection of linear trajectory basis, $\mathbf{a}$ is the coefficients or the parameters of a trajectory, and K is the number of basis.

If two trajectories, $\mathbf{X}_1$ and $\mathbf{X}_2$, are articulated, the distance between trajectories remains constant across all time instances as shown in Figure 1(a), i.e.,

$$\Delta x_i^2+\Delta y_i^2+\Delta z_i^2=r^2, \quad i=1,\cdots,F, \tag{3}$$

where $\Delta\mathbf{X}=\mathbf{X}_2-\mathbf{X}_1$ is the relative trajectory.

When a perspective camera captures these two trajectories, points on the trajectories at the time instant are projected onto the camera plane. The camera representation in this paper is a 3×4 projection matrix, $\mathbf{P}_i = \mathbf{K}\mathbf{R}_i\left[\begin{array}{c|c}\mathbf{I}_3 & -C_i\end{array}\right]$ where $\mathbf{I}_3$, $\mathbf{K}$, $\mathbf{R}_i$, and C_i are a 3×3 identity matrix, the upper triangular intrinsic matrix, the camera rotation matrix, and the camera's optical center vector at the i^{th} time instant, respectively.

If we transform one of the trajectories, $\mathbf{X}_1$, to the origin, $\mathbf{O}$, the other trajectory, $\mathbf{X}_2$, maps to the relative trajectory, $\Delta\mathbf{X}$, and a camera, $\mathbf{P}_i$, maps to the relative camera pose, $\widetilde{\mathbf{P}}_i$ with respect to $\mathbf{X}_1$ as shown in Figure 1(b). The transformed relative trajectory lies on a sphere with radius r. There are two points intersecting the sphere and the ray connecting the camera's optical center and an image measurement at each time instant as shown in Figure 1(c). All intersecting points are candidate 3D points which the relative trajectory passes and thus, there are 2^F possible relative trajectories.

The representation of a relative trajectory between the articulated trajectories from Equation (2) (Cartesian coordinate representation) has to meet the additional quadratic equality constraints of Equation (3). Instead of the Cartesian coordinate representation, we introduce the spherical coordinate representation for a relative trajectory to control the distance between trajectories, explicitly, i.e.,

$$\Delta\mathbf{X}=(\boldsymbol{\Theta}\mathbf{a}_\theta, \boldsymbol{\Theta}\mathbf{a}_\phi, r), \tag{4}$$

where θ is inclination from the z axis, ϕ is azimuth from the x axis in the xy plane, and r is the radius . This representation enables us to describe an articulated trajectory precisely because it satisfies the temporal constraint and the length constraint simultaneously regardless of parameters by setting the radius constant explicitly. It also enforces that all imputed points between frames satisfy the articulation constraint while the Cartesian representation does not. For a topological point of view, the reconstruction from the spherical coordinate system is the mapping of $\mathbb{P}^{2F}\rightarrow\mathbb{S}^{2K}\times\mathbb{R}^1$ while the reconstruction from the Cartesian coordinate system is the mapping of $\mathbb{P}^{2F}\rightarrow\mathbb{P}^{3K}$ as shown in Figure 1(c).

3. Method

In this section, we present an algorithm for recovering a trajectory which satisfies spatial and temporal constraints using the spherical coordinate representation of a relative trajectory presented in the previous section.

3.1. Objective Function of 3D Reconstruction

From the spherical coordinate representation, we reconstruct smooth articulated trajectories which minimize the reprojection errors:

$$\operatorname*{argmin}_{\Delta\mathbf{X}_1,\cdots,\Delta\mathbf{X}_P} \sum_{i,j}^{F,P} d\left(\mathbf{x}_{ij}, \hat{\mathbf{x}}_{ij}\right), \tag{5}$$

where $\Delta\mathbf{X}_j$ is the j^{th} articulated (or relative) trajectory parameterized by $(\mathbf{\Theta}\mathbf{a}_{\theta,j}, \mathbf{\Theta}\mathbf{a}_{\phi,j}, r_j)$, $d(\cdot,\cdot)$ is the L_2 distance between two arguments, P is the number of articulated points, and $\mathbf{x}_{ij}$ and $\hat{\mathbf{x}}_{ij}$ are a 2D image measurement and a reprojection of the j^{th} point trajectory at the i^{th} time instant, respectively.

If articulated trajectories are sequentially linked, the trajectories are

$$\mathbf{X}_j = f(\mathbf{X}_R; \Delta\mathbf{X}_1, \cdots, \Delta\mathbf{X}_{j-1}), \tag{6}$$

where $f(\cdot)$ is the forward kinematic function that takes the root trajectory, $\mathbf{X}_R$, and all parent relative trajectories, $\Delta\mathbf{X}_1, \cdots, \Delta\mathbf{X}_{j-1}$, and outputs the j^{th} trajectory, $\mathbf{X}_j$, in the Cartesian coordinate system. The reprojection, $\hat{\mathbf{x}}_{ij}$ is

$$\hat{\mathbf{x}}_{ij} = \left(\frac{\mathbf{P}_i^1\widetilde{\mathbf{X}}_j(i)}{\mathbf{P}_i^3\widetilde{\mathbf{X}}_j(i)}, \frac{\mathbf{P}_i^2\widetilde{\mathbf{X}}_j(i)}{\mathbf{P}_i^3\widetilde{\mathbf{X}}_j(i)}\right), \tag{7}$$

where $\mathbf{P}_i^l$ is the l^{th} row of the camera projection matrix at the i^{th} time instant and $\widetilde{\mathbf{X}}_j(i)$ is the homogeneous representation of the i^{th} point in the j^{th} trajectory, $\mathbf{X}_j(i)$.

3.2. Initialization of Equation (5)

The objective function of Equation (5) is highly nonlinear and direct optimization falls into a local minimum. Therefore, a good initialization of trajectory parameters is necessary. When the parent joint position and the length between trajectories are known, there are two intersecting points between a sphere whose origin is the parent joint position, X_p, and a line connecting an image measurement and camera optical center, C, at each time instant as shown in Figure 2(a). A point lying on the line is $C + s\mathbf{v}$ where s is an unknown scalar and $\mathbf{v}$ is the direction of the projection, i.e.,$\mathbf{v} = \mathbf{R}^{\mathsf{T}}\mathbf{K}^{-1}\begin{bmatrix} \mathbf{x}^{\mathsf{T}} 1 \end{bmatrix}^{\mathsf{T}}$. Then, the intersecting points are

$${}^1X = C + s_1\mathbf{v}, \ \ {}^2X = C + s_2\mathbf{v}, \tag{8}$$

where

$$s_{1,2} = \frac{-\mathbf{v}^{\mathsf{T}}\Delta C \pm \sqrt{\left(\mathbf{v}^{\mathsf{T}}\Delta C\right)^2 - \|\mathbf{v}\|^2\left(\|\Delta C\|^2 - r^2\right)}}{\|\mathbf{v}\|^2} \tag{9}$$

and $\Delta C = C - X_p$. For each time instant, we have two candidate 3D points through which the reconstructed trajectory must pass. Across all time instances, there are 2^F possible trajectories which satisfy the image measurements. Among those trajectories, we look for the trajectory best described by the trajectory basis.

Let χ be the relative direction vector with respect to the parent point as shown in Figure 2(a). For each time instant, χ_i takes either ${}^1\chi_i$ or ${}^2\chi_i$, i.e.,

$$\begin{aligned} \chi_i &= {}^1\chi_i b_i + {}^2\chi_i(1 - b_i), \\ &= ({}^1\chi_i - {}^2\chi_i)b_i + {}^2\chi_i, \quad \text{where } b_i \in \{0, 1\}. \end{aligned} \tag{10}$$

Then, all possible trajectories can be represented as:

$$\begin{aligned} \begin{bmatrix} \chi_1 \\ \vdots \\ \chi_F \end{bmatrix} &= \begin{bmatrix} \Delta\chi_1 & & \\ & \ddots & \\ & & \Delta\chi_F \end{bmatrix} \mathbf{b} + \begin{bmatrix} {}^2\chi_1 \\ \vdots \\ {}^2\chi_F \end{bmatrix} \\ \text{or} \quad \chi &= \mathbf{E}\mathbf{b} + \mathbf{F}, \end{aligned} \tag{11}$$

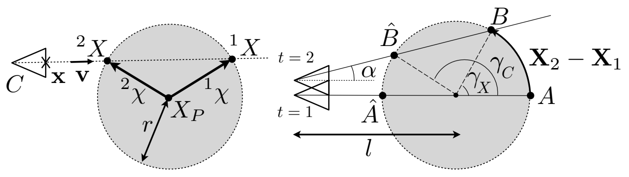
where $\mathbf{b}$ is a binary variable vector, ${}^1\chi_i$ and ${}^2\chi_i$ are two relative direction vectors, and $\Delta\chi_i = {}^1\chi_i - {}^2\chi_i$. Finding the best trajectory is equivalent to finding the binary vector, $\mathbf{b}$, which minimizes the following cost,

$$\begin{aligned} \mathbf{b}^* &= \operatorname*{argmin}_{\mathbf{b}} \left\| \left(\mathbf{\Theta}\mathbf{\Theta}^{\mathsf{T}} - \mathbf{I}\right)\left(\mathbf{E}\mathbf{b} + \mathbf{F}\right)\right\|^2, \\ & \text{subject to} \quad \mathbf{b} \in \{0, 1\}^F. \end{aligned} \tag{12}$$

Note that $\mathbf{\Theta}\mathbf{\Theta}^{\mathsf{T}} - \mathbf{I}$ is the projection operation onto the null space of the trajectory basis, $\mathbf{\Theta}$. Equation (12) is a quadratic problem over binary variables.

A binary quadratic programming problem is NP-hard in general. The structure of our problem does not fall into one of the solvable cases; our quadratic matrix has positive off-diagonal elements [30], is a non-singular matrix [3, 14], and cannot be represented by a tri-/five-diagonal matrix [16]. Also, the underlying graph structure is not series parallel [6]. Thus, in theory, this is an intractable problem. However, a number of approaches have been proposed to approximate a solution of the problem efficiently using spectral or semidefinite relaxation. A branch-and-bound routine[2] with binary relaxation is one technique for global optimization. Since our quadratic matrix is positive definite, the objective function behaves convexly in a branched rectangle, which enables us to define a tight lower bound of the rectangle in polynomial time.

[2] http://www.dii.unisi.it/~hybrid/tools/miqp/



(a) Intersection (b) Relative trajectory

Figure 2. (a) There are two solutions, ${}^1X$ and ${}^2X$ which satisfy the articulation constraint and an image measurement. (b) Articulated trajectory and the camera pose are transformed with respect to the parent trajectory.

Once $\mathbf{b}^*$ is recovered, we project $\boldsymbol{\chi} = \mathbf{E}\mathbf{b}^* + \mathbf{F}$ onto the trajectory basis space of the spherical coordinate system to produce low dimensional parameters, i.e., $\Delta\mathbf{X} = (\boldsymbol{\Theta}\mathbf{a}_\theta, \boldsymbol{\Theta}\mathbf{a}_\phi, r)$. This yields an accurate initialization which can be refined by nonlinear optimization of Equation (5).

When the relative trajectory, $\Delta\mathbf{X}$, passes a singular point in the spherical coordinate system in the process of projecting $\boldsymbol{\chi}$ onto the spherical coordinate system, a discontinuity of angular trajectory occurs. For example, when ϕ passes from $\epsilon > 0$ to $2\pi - \epsilon$, this results in a discontinuity of the angular trajectory because ϕ is defined in the interval $[0, 2\pi)$. To deal with discontinuous trajectories, we find the best angular representation among all spherical representations of χ which preserves local continuity by allowing the domains of θ and ϕ to be $(-\infty, \infty)$.

4. Reconstructibility

We now explore the reconstruction ambiguity of an articulated trajectory and analyze configurations in which the reconstruction is accurate. Let $\mathbf{X}_1$ be a known parent trajectory and $\mathbf{X}_2$ be an articulated child trajectory which are observed at two time instances as shown in Figure 2(b). The ground truth relative trajectory between $\mathbf{X}_1$ and $\mathbf{X}_2$ moves from A to B. $\hat{A}$ and $\hat{B}$ are impostor points that satisfy the image measurements as well as the articulation constraint. In this section, we show that the relationship between the true trajectory and the impostor trajectory inherently determines the reconstruction accuracy.

We define a measure of *reconstructibility of an articulated trajectory*, η_a, as a criterion to characterize reconstruction accuracy where

$$\eta_a = \frac{\left\|\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{C}}\right\|}{\left\|\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{X}}\right\|}, \tag{13}$$

$\boldsymbol{\gamma}_\mathbf{X} = \boldsymbol{\Theta}\mathbf{a}_{\boldsymbol{\gamma}_\mathbf{X}} + \boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{X}}$, $\boldsymbol{\gamma}_\mathbf{C} = \boldsymbol{\Theta}\mathbf{a}_{\boldsymbol{\gamma}_\mathbf{C}} + \boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{C}}$, and $\boldsymbol{\Theta}^\perp$ is the null space of the trajectory basis. If the reconstructibility of an articulated trajectory goes to infinity, there exists a unique solution and it corresponds to the ground truth trajectory. This can be proven by the following. For each time instant, there are two intersecting points and an estimation should be one of them:

$$\hat{\gamma} = (1-b)\gamma_X + b\gamma_C, \quad b = 1 \text{ or } 0 \tag{14}$$

where $\hat{\gamma}$ is an estimated angle. For an estimated angular trajectory,

$$\hat{\boldsymbol{\gamma}} = (\mathbf{I} - \mathbf{B})\,\boldsymbol{\gamma}_\mathbf{X} + \mathbf{B}\boldsymbol{\gamma}_\mathbf{C}, \tag{15}$$

where $\mathbf{B}$ is a diagonal matrix whose entries takes either 1 or 0. The best trajectory represented by the trajectory basis minimizes following:

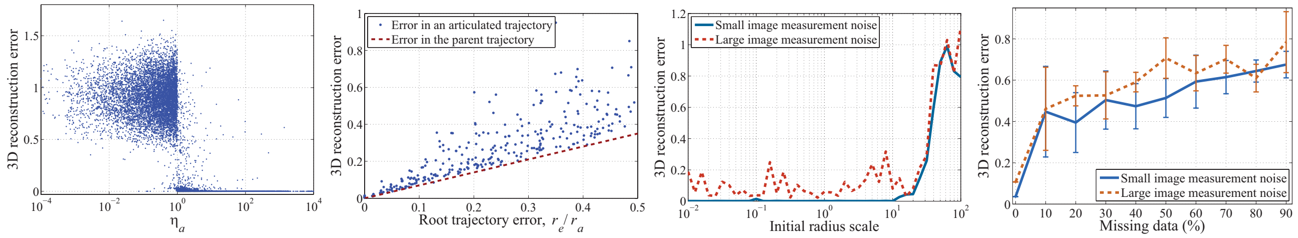
$$\begin{aligned}
& \underset{\hat{\mathbf{a}},\mathbf{B}}{\operatorname{argmin}} \left\|\boldsymbol{\Theta}\hat{\mathbf{a}} - \hat{\boldsymbol{\gamma}}\right\|^2 & (16)\\
= \; & \underset{\hat{\mathbf{a}},\mathbf{B}}{\operatorname{argmin}} \left\|\boldsymbol{\Theta}\hat{\mathbf{a}} - (\mathbf{I} - \mathbf{B})\,\boldsymbol{\gamma}_\mathbf{X} - \mathbf{B}\boldsymbol{\gamma}_\mathbf{C}\right\|^2 & (17)\\
= \; & \underset{\hat{\mathbf{a}},\mathbf{B}}{\operatorname{argmin}} \Big\|\boldsymbol{\Theta}\hat{\mathbf{a}} - (\mathbf{I} - \mathbf{B})\,\boldsymbol{\Theta}\mathbf{a}_{\boldsymbol{\gamma}_\mathbf{X}} - \mathbf{B}\boldsymbol{\Theta}\mathbf{a}_{\boldsymbol{\gamma}_\mathbf{C}} & \\
& \quad - (\mathbf{I} - \mathbf{B})\,\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{X}} - \mathbf{B}\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{C}}\Big\|^2. & (18)
\end{aligned}$$

Reconstructibility of an articulated trajectory goes to infinity when $||\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{C}}|| \rightarrow \infty$ or $||\boldsymbol{\Theta}^\perp \mathbf{a}^\perp_{\boldsymbol{\gamma}_\mathbf{X}}|| \rightarrow 0$. For either case, $\mathbf{B}$ has to approach $\mathbf{0}$ to eliminate the residual of the null components in Equation (18), which leads to $\hat{\mathbf{a}} \rightarrow \mathbf{a}_{\boldsymbol{\gamma}_\mathbf{X}}$.

From the method of Park *et al.* [28], if the camera motion is slow or stationary, there is no way to reconstruct an accurate trajectory using the trajectory basis because it spans the camera trajectory well. The reconstructibility of an articulation states that if the parent trajectory is independent of the camera trajectory, the trajectory reconstruction is still possible because mixed motion between the camera and the parent motions induces $\boldsymbol{\alpha}$ motion where $\boldsymbol{\alpha}$ is the trajectory of viewing angles from a camera, α, as shown in Figure 2(b). Even when camera and parent motions are stationary, the reconstruction is possible if $\boldsymbol{\gamma}_\mathbf{X} \in \boldsymbol{\Theta}$ because each α is a nonlinear function of γ_X, i.e., $\alpha = \tan^{-1}(\sin\gamma_X/(l + \cos\gamma_X))$ where l is the distance between the parent trajectory and camera trajectory, and thus $\boldsymbol{\alpha} \notin \boldsymbol{\Theta}$ and $\boldsymbol{\gamma}_\mathbf{C} \notin \boldsymbol{\Theta}$ unless $l = 0$ or $l = \infty$ (i.e., orthographic projection) as shown in Figure 2(b).

Figure 3(a) shows the distribution of 3D reconstruction error with respect to reconstructibility of an articulated trajectory, η_a, from the CMU motion capture data[3]. A trajectory initialized by binary quadratic programming is the best fitted trajectory by the trajectory basis. When η_a is high ($\gg 1$), 3D reconstruction error of an articulated trajectory is low because the ground truth trajectory is well described by the trajectory basis and the ground truth trajectory and the impostor trajectory are well separable. In contrast, when η_a is low ($\ll 1$), our solution converges to the impostor trajectory because the trajectory basis spans the impostor trajectory better.

[3] http://mocap.cs.cmu.edu/



(a) Reconstructibility (b) Error in the root trajectory (c) Initialization error of radius (d) Effect of missing data

Figure 3. (a) The accuracy of the reconstruction is high when η_a is greater than 1 where the trajectory basis spans the ground truth trajectory better than the impostor trajectory. (b) Performance of our algorithm against error in the root trajectory, (c) the initialization error of the radius, (d) amount of missing data are illustrated.

5. Results

To validate our method, we tested it with the HumanEva-II dataset, synthesized trajectories, and the CMU motion capture data quantitatively and with real human motion examples taken by video cameras qualitatively. We use the first K Discrete Cosine Transform (DCT) basis[4] in order of increasing frequency and the number of basis is chosen manually to span the trajectory well.

5.1. Quantitative Evaluation

We compare our method with the state-of-art human pose estimation [4, 8, 20, 29] using the HumanEva-II dataset[5]. Subject S2 with camera C1 is used to reconstruct the articulated trajectories. Our method results in 128.8mm of 3D mean error with 17.75mm standard deviation. This error is comparable to the error of the state-of-art pose estimation algorithms (82mm∼211.4mm). It should be noted that while all methods rely on activity specific training data to reconstruct motions, our approach uses only activity independent geometric constraints.

We generate synthetic 2D perspective projections from synthetic data and the CMU motion capture data and evaluate for three aspects: error in the root trajectory, error in radius of an articulated trajectory, and missing data. For evaluation of errors in the root trajectory and radius, we set the camera stationary and vary error of the root trajectory and radius error while the root position is moving. For the evaluation of missing data, we artificially remove 2D projections randomly.

We measure 3D reconstruction error of an articulated trajectory by varying the ratio between the average distance error of the root trajectory, r_e, and the radius of the articulated trajectory, r_a, as shown in Figure 3(b). The error in the parent trajectory is a lower bound on the reconstruction error of the articulated trajectory. While the variance of the distribution for small root trajectory error (< 0.2) is low, i.e., the reconstruction can be done reliably, the reconstruction from high root trajectory error (> 0.3) causes high error in the child trajectory as well.

For the evaluation of the error in radius, we measure 3D reconstruction error for erroneous radii multiplied by scale[6]. Figure 3(c) illustrates robustness to erroneous initialization. Even though the initial scale is small (i.e., $10^{-2} \sim 10^0$), the 3D reconstruction can be done reliably because before solving the binary quadratic programming, we adjust the radius of the sphere to intersect with the line of projection at one point at least. When the initial scale is high ($> 10^1$), however, the reconstruction becomes unreliable because the ray intersects with the sphere at all time instances and the optimization falls into a local minimum around a mis-estimated trajectory.

We also test with the CMU motion capture data for the evaluation of missing data caused by occlusion or measurement failure. When there are missing data, our spatial and temporal constraints enable us to impute missing points. For this experiment, we artificially introduce length errors, image measurement noise, and root trajectory error while the camera is stationary. Our algorithm produces an average error[7] of 13% for 5% missing data as shown in Figure 3(d).

5.2. Qualitative Evaluation

We apply our algorithm to reconstruct human body motion in 3D from 2D perspective projections. Reconstruction from a stationary camera and a moving camera are tested and the statistical anthropometric length ratio of the human body is used for the initialization of length ratio with some modifications for accurate skeleton estimation purpose. The scale of the skeleton is roughly initialized and we manually label image measurements for articulated points.

Figure 4(a) and Figure 4(b) show the reconstruction of the *juggling motion* and the *motion in front of a webcam*,

[4] Hamidi and Pearl [17] have shown that the DCT provides the optimal performance to encode the signal under the first order Markov processes. Ahkter *et al.* [2] have empirically justified its optimality on motion capture data.

[5] http://vision.cs.brown.edu/humaneva/

[6] Initial radius scale error 1 means the ground truth.

[7] error = $||\mathbf{X} - \hat{\mathbf{X}}||/||\mathbf{X}||$, where $\mathbf{X}$ is the ground truth trajectory and $\hat{\mathbf{X}}$ is the estimated trajectory.

respectively, in 3D from a stationary video camera. We project the 2D root trajectory to the unit depth plane and use it as the 3D root trajectory because the depth of the root trajectory is underdetermined from a stationary camera. For both experiments, we use the torso as the root. From the root trajectory, all articulated trajectories are reconstructed recursively.

We also apply our method to data captured from a moving camera to recover the *playing card motion* and the *yoga motion* as shown in Figure 4(c) and Figure 4(d), respectively. Both camera trajectories are smooth and well spanned by the trajectory basis. For the reconstruction of the root trajectory, we choose a relatively rigid part of human body through a sequence and reconstruct them using the structure from motion algorithm. Once relative camera poses are estimated from the rigid part of the human body, we estimate the similarity transform between the relative camera poses and the original camera poses estimated by 3D static structure. Head and torso are used as the root for playing card motion and yoga motion, respectively.

6. Related Work

Points on objects involve spatial and temporal constraints in general. Bregler *et al.* [10] introduced a spatial constraint by assuming that a shape undergoes a deformation which can be well expressed by a linear combination of a compact shape basis. Subsequent work has discussed ways to select the basis and bilinear optimization schemes that separate information of camera motion and structure [9, 33, 39]. For multiple rigid bodies, Costeira and Kanade [13] proposed a factorization method using 2D trajectories in an image stream. Yan and Pollefeys [40] extended this noting that there is rank loss when two rigid bodies are articulated. In the temporal domain, Akhter *et al.* [2] proposed a linear trajectory basis using the Discrete Cosine Transform (DCT) by assuming that an underlying 3D trajectory moves continuously and smoothly. With a calibrated camera, an algebraic solution of trajectory reconstruction in 3D for specific algebraic curves (lines and conics) has been proposed by Avidan and Shashua [5] and later it was generalized by Kaminski and Teicher [22]. Park *et al.* [28] proposed a linear solution for the trajectory reconstruction. Previous methods exploit either spatial constraints or temporal constraints. In our method, we use the spatial constraint in the form of articulation and the temporal constraint in the form of smoothness.

For human motion in particular, human pose estimation from a single image by applying a spatial constraint (skeletal structure) was proposed by Taylor [32] (parameterization of limb lengths by a scalar), by Barron and Kakadiaris [7] (joint motion constraint from the anthropometric statistics), by Parameswaran and Chellappa [27] (camera pose estimation from head orientation and rigidity of torso), and by Agarwal and Triggs [1] (silhouette based regression). Human motion estimation from an image sequence of a monocular camera has been studied as an extension of human pose estimation. Two popular approaches have been explored: data driven approach and physics based approach. Data driven approach is to learn low dimensional subspace or latent variables that control underlying human skeletal motion fully using motion capture data or annotated video data. Sidenbladh *et al.* [18] applied a Bayesian framework for 3D human pose tracking using a generative model of the human body and a prior distribution defined by a temporal dynamics model. Howe *et al.* [19] showed Baysian learning, Choo and Fleet [12] sampled high dimensional training space from hybrid Monte Carlo method, and Urtasun *et al.* [34] used Principle Coordinate Analysis (PCA) for learning of specific motion (e.g. walking and golfing). Like Taylor's work [32], Wei and Chai [37] introduced a geometric solution of motion reconstruction using the bone symmetric constraint from biomechanical data. Valmadre and Lucey [35] discussed the validity of Wei and Chai [37]'s work and extended their algorithm using a structure from motion scheme. Recently, physics based approaches have received attention. Brubaker *et al.* [11] have shown reconstruction of a bipedal locomotion from a dynamical model and Vondrak *et al.* [36] have applied multibody dynamics simulation to infer the most plausible human motion in 3D. Wei and Chai [38] have built an interactive system that integrates a dynamical model to capture motion from a video.

Unlike previous methods, our approach relies purely on a geometric interpretation of the articulation constraint by parameterizing a trajectory in a way that satisfies both spatial and temporal constraints simultaneously.

7. Discussion

In this paper, we study an articulated trajectory which remains at a constant distance with respect to the parent trajectory. The relative trajectory is a trajectory on a sphere and there are 2^F trajectories that meet the spatial constraint and image measurements. Among those trajectories, we look for the best trajectory spanned by the trajectory basis and we identify that this is equivalent to solving a binary quadratic programming problem. The relative trajectory obtained by the binary quadratic program is parameterized by a compact trajectory basis in the spherical coordinate system, which satisfies spatial and temporal constraints, simultaneously. We optimize the trajectory by minimizing reprojection error. Reconstruction of the articulated trajectory is fundamentally limited by the motion induced by the camera and the parent trajectory and we propose a measure of reconstructibility of an articulated trajectory, which characterizes the reconstruction accuracy. Our results show that we are able to reconstruct highly articulated human motions from a stationary camera and a moving camera.

Acknowledgements

The authors wish to thank Tomas Simon and the anonymous reviewers for their invaluable comments. This work was supported by NSF grant IIS-0916272.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 2006. 6

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *PAMI*, 2011. 2, 5, 6

[3] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner. A polynomial case of unconstrained zero-one quadratic optimization. *Mathematical Programming*, 2001. 3

[4] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010. 5

[5] S. Avidan and A. Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *PAMI*, 2000. 6

[6] F. Barahona. A solvable case of quadratic 0-1 programming. *Discrete Applied Mathematics*, 1986. 3

[7] C. Barron and I. A. Kakadiaris. Estimating anthropometry and pose from a single image. In *CVPR*, 2000. 6

[8] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *IJCV*, 2010. 5

[9] M. Brand. A direct method for 3D factorization of nonrigid motion observed in 2D. In *CVPR*, 2005. 6

[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 1999. 6

[11] M. A. Brubaker and D. J. Fleet. The kneed walker for human pose tracking. In *CVPR*, 2008. 1, 6

[12] K. Choo and D. J. Fleet. People tracking using hybrid monte carlo filtering. In *ICCV*, 2005. 1, 6

[13] J. P. Costeira and T. Kanade. A multibody factorization method from independently moving objects. *IJCV*, 1998. 6

[14] J.-A. Ferrez, K. Fukuda, and T. M. Liebling. Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm. *European Journal of Operations Research*, 2004. 3

[15] M. R. Garey and D. S. Johnson. *Computer and Interactability: A guide to the theory of NP-Completeness*. Freeman, 1979. 1

[16] S. Gu. Polynomial time solvable algorithms to binary quadratic programming problems with q being a tri-diagonal or five-diagonal matrix. In *WCSP*, 2010. 3

[17] M. Hamidi and J. Pearl. Comparison of the cosine and fourier transforms of markov-1 signal. *IEEE Trans. Acoust. Speech, Signal Process*, 1976. 5

[18] M. J. B. Hedvig Sidenbladh and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, 2000. 6

[19] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. In *NIPS*, 1999. 1, 6

[20] Z. L. Husz, A. M. Wallace, and P. R. Green. Evaluation of a hierarchical partitioned particle filter with action primitives. In *CVPR Workshop*, 2007. 5

[21] G. Jansson, S. S. Bergstorm, and W. Epstein. *Perceiving Events and Objects*. Lawrence Erlbaum, 1994. 1

[22] J. Y. Kaminski and M. Teicher. A general framework for trajectory triangulation. *Journal of Mathematical Imaging and Vision*, 2004. 6

[23] H.-J. Lee and Z. Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 1985. 1

[24] P. Merz and B. Freisleben. Greedy and local search heuristics for unconstrained binary quadratic programming. *Journal of Heuristics*, 2002. 1

[25] C. Olsson, A. P. Eriksson, and F. Kahl. Solving large scale binary quadratic problems: Spectral methods vs. semidefinite programming. In *CVPR*, 2007. 1

[26] K. E. Ozden, K. Cornelis, L. V. Eycken, and L. V. Gool. Reconstructing 3D trajectories of independently moving objects using generic constraints. *CVIU*, 2004. 1

[27] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *CVPR*, 2004. 6

[28] H. S. Park, T. Siratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010. 1, 4, 6

[29] P. Peursum, S. Venkatesh, and G. West. A study on smoothing for particle-filtered 3D human body tracking. *IJCV*, 2010. 5

[30] J. C. Picard and P. M. Ratliff. Minimal cost cut equivalent networks. *Management Science*, 1973. 3

[31] S. Poljak, F. Rendl, and H. Wolkowicz. A recipe for semidefinite relaxation for (0,1)-quadratic prgramming. *Journal of Global Optimization*, 1995. 1

[32] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *CVIU*, 2000. 6

[33] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 2008. 6

[34] R. Urtasun, D. J. Fleet, and P. Fua. Temporal motion models for monocular and multiview 3-D human body tracking. *CVIU*, 2006. 1, 6

[35] J. Valmadre and S. Lucey. Deterministic 3D human pose estimation using rigid structure. In *ECCV*, 2010. 1, 6

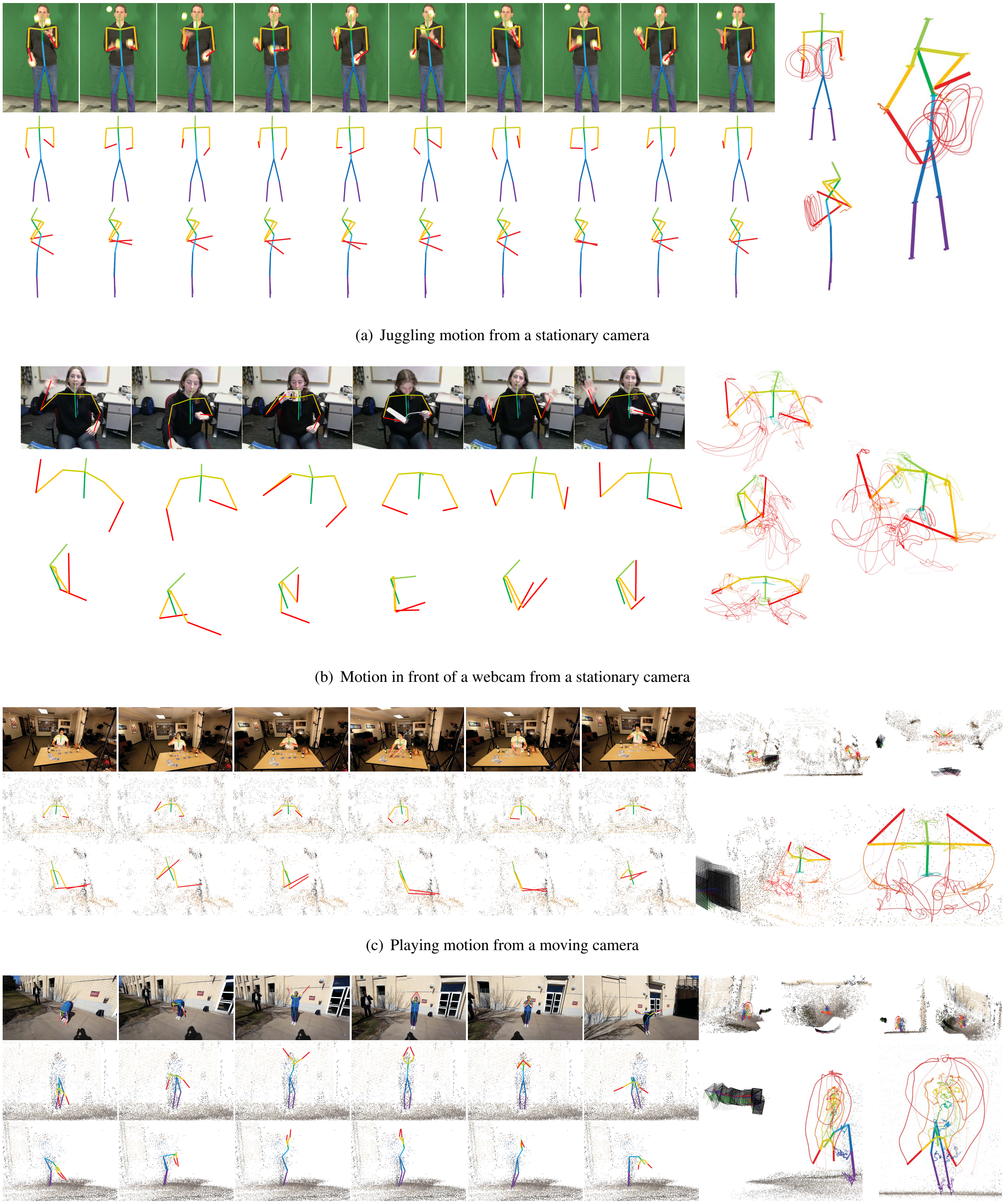
[36] M. Vondrak, L. Sigal, and O. C. Jenkins. Physic simulation for probabilistic motion tracking. In *ECCV*, 2008. 1, 6

[37] X. Wei and J. Chai. Modeling 3D human poses from uncalibrated monocular images. In *ICCV*, 2009. 1, 6

[38] X. Wei and J. Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. *SIGGRAPH*, 2010. 1, 6

[39] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *IJCV*, 2006. 6

[40] J. Yan and M. Pollefeys. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *PAMI*, 2008. 6



(a) Juggling motion from a stationary camera

(b) Motion in front of a webcam from a stationary camera

(c) Playing motion from a moving camera

(d) Yoga motion from a moving camera

Figure 4. (a) Juggling motion, (b) motion in front of the webcam from a stationary camera, (c) playing card motion, and (d) yoga motion from a moving camera. Image measurements are superimposed on images in the top row and 3D reconstruction of the motion corresponding to the images are shown from different views in the second and the third rows. The right-most figures summarize motion by showing whole trajectories.