Efficient Articulated Trajectory Reconstruction Using Dynamic Programming and Filters

Jack Valmadre[1,3] Yingying Zhu[2,3] Sridha Sridharan[1] Simon Lucey[3,1]

[1]Queensland University of Technology, Australia
[2]University of Queensland, Australia
[3]Commonwealth Scientific and Industrial Research Organisation, Australia
{jack.valmadre, yingying.zhu, simon.lucey}@csiro.au

Abstract. This paper considers the problem of reconstructing the motion of a 3D articulated tree from 2D point correspondences subject to some temporal prior. Hitherto, smooth motion has been encouraged using a trajectory basis, yielding a hard combinatorial problem with time complexity growing exponentially in the number of frames. Branch and bound strategies have previously attempted to curb this complexity whilst maintaining global optimality. However, they provide no guarantee of being more efficient than exhaustive search. Inspired by recent work which reconstructs general trajectories using compact high-pass filters, we develop a dynamic programming approach which scales linearly in the number of frames, leveraging the intrinsically local nature of filter interactions. Extension to affine projection enables reconstruction without estimating cameras.

1 Introduction

Trajectory basis approaches to Non-Rigid Structure from Motion (NRSfM) are able to model modes of deformation which traditional shape basis methods cannot [2]. Unfortunately, their application is often limited by the issue of reconstructability [3]. Articulated trajectories are a special class for which this is not true. A trajectory is said to be articulated if it must remain a constant distance from a parent trajectory at all times. The ability to reconstruct articulated trajectories is valuable to any monocular vision task involving a skeleton, whether it be human, animal or robotic.

Reconstructing an articulated trajectory given its parent trajectory, its projection in a known camera and the articulation length is a binary combinatorial problem [1], since a forwards/backwards ambiguity exists at each time instant (Figure 2). Although the projection constraints permit two solutions per frame, our intuition (and Newton's second law) suggests that any object which has mass should move smoothly. Motivated by trajectory basis NRSfM, Park and Sheikh [1] recently proposed to solve articulated trajectory reconstruction by searching the finite feasible set for the trajectory nearest to a low-dimensional

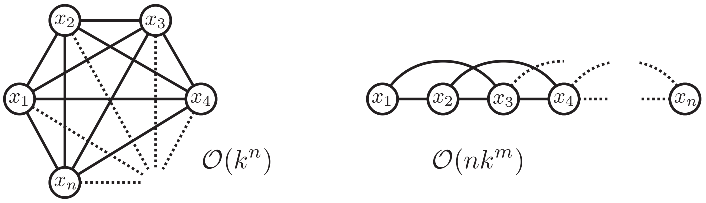


Fig. 1: Solving for the trajectory of an articulated joint is a binary ($k = 2$) combinatorial problem. Park and Sheikh [1] minimised the distance of the trajectory from a subspace, which has inseparable terms depending on the solution for all frames (*left*). The worst-case time complexity of their algorithm is exponential in the number of frames n. We present an approach which scales linearly in n by considering only local interactions of order m (*right*).

subspace representative of this Newtonian behaviour. This is a discrete optimisation problem with time complexity exponential in the number of frames. They presented a globally optimal branch and bound strategy to reduce the running time, however, it does so without theoretical guarantee.

The key contributions of this paper are that we:-

- Incorporate temporal prior using compact filters instead of a basis to yield a globally optimal solution which scales linearly in the number of frames and exponentially only in the support of the filter, which is generally a small constant. Asymptotic time complexity is verified experimentally.
- Provide an elegant re-formulation of the projection constraints which encompasses affine as well as full perspective cameras. This extension enables one to solve for articulated motion in camera co-ordinates without the need to estimate camera calibration and root node position, provided that the camera moves smoothly and a weak perspective assumption is reasonable. Practical 3D reconstructions are demonstrated for videos of humans and animals, in which the dimensions of the skeleton are a priori unknown and the background makes camera estimation difficult.

The paper is structured as follows. Related work is briefly reviewed in §2. The problem is defined in §3. Principles of dynamic programming which enable application to filter responses are covered in §4 and the calibration-less extension is introduced in §5. Experimental results are presented in §6 before concluding in §7.

2 Related Work

There is a large body of literature on NRSfM, which aims to jointly estimate cameras and deformable structure from 2D point correspondences alone. Unlike rigid SfM, this is an inherently ill-posed problem since the structure can vary between frames, resulting in more variables than equations. To counter this, two dominant paradigms for solving NRSfM have emerged, requiring either a) the 3D structure in each frame be restricted to a low-dimensional shape basis [4] or b) the path of each point through time be restricted to a low-dimensional trajectory basis [2]. Trajectory basis approaches are generally able to represent a larger range of motion and can adopt a content-agnostic basis (typically that of the Discrete Cosine Transform [DCT]), eliminating the need to solve for the unknown basis vectors. However in situations of poor reconstructability, the accuracy with which these methods can recover 3D structure is limited by the insufficient motion of the camera [3, 5]. Zhu et al. [6] showed that poor reconstructability can be overcome if rigid keyframes are available, encouraging a sparse vector of trajectory basis coefficients. Park and Sheikh [1] recently identified articulated trajectory reconstruction as a problem which is not prone to reconstructability due to the additional shape constraints. They solved the arising binary combinatorial problem using branch and bound, minimising the component of the trajectory which is orthogonal to a truncated DCT basis. Other recent work has considered using a shape basis with coefficients that lie on a trajectory basis [7], using a spatiotemporal basis [8] and being completely basis-agnostic in favour of a low-rank prior alone [9]. These approaches, however, can not take advantage of known articulation constraints.

There are a number of earlier works examining the specific problem of articulated SfM, where the observed structure is known to be a collection of linked rigid bodies. Tresadern and Reid [10] and Yan and Pollefeys [11] independently established that the measurement matrix of two rigid objects is of lower rank when they are connected at a joint or hinge, proposing factorisation-based algorithms to recover the structure when segmentation is known. Yan and Pollefeys additionally discovered the articulation topology and part segmentation using Generalised Principal Component Analysis. Paladini et al. [12] later gave an iterative algorithm which, unlike approaches reliant on factorisation, can recover articulated structure despite missing data. Fayad et al. [13] automatically assign point tracks to rigid parts using graph cuts and recover 3D structure in an alternation scheme. An issue with all of these methods is the requirement of reliable, dense point correspondences for each articulated segment, which frequently cannot be obtained due to the slender, often texture-less nature of articulated parts.

Our problem is also related to algorithms for articulated parts-based detection which employ a three-dimensional model. We should note, however, that unlike the tracking and detection literature, our proposed approach begins with the assumption that joint positions have been obtained. In the past, a slew of particle filter approaches have been developed for 3D human body tracking. Sidenbladh et al. [14] use a generative model of appearance with either generic temporal smoothness or action-specific motion prior, also incorporating limits on the range of joint angles. Sminchisescu and Triggs [15] propose a method for intelligent sampling of pose configurations, explicitly considering the two-fold ambiguity and pruning solutions based on kinematic limits and self-collision. Unlike particle filter approaches, our algorithm recovers a deterministic, optimal solution in a batch scenario. More recent approaches learn a prior over the space of human pose and motion in a Gaussian process framework [16,17]. These methods are capable of learning much higher-capacity models, but therefore require a large volume of training data covering the space of possible configurations. Physics-based approaches [18] have also been entertained, where a dynamical model is fit to a video sequence. A drawback to these approaches, however, is that they require a highly complex model to generalise beyond simple actions. Wei and Chai [19] alternatively propose an interactive process for monocular human motion capture which combines 3D keyframes, Newtonian physics and contact constraints. The central appeal of articulated trajectory reconstruction, as posed in [1], is that it employs only geometric constraints, allowing reconstruction of a much broader class of articulated structures.

The problem of estimating articulated structure from a single image has been considered in a number of notable works. Taylor [20] and Parameswaran and Chellappa [21] require the user to manually specify the direction of each bone, solving camera calibration using a weak perspective model with fixed relative bone-length proportions and using a full perspective model with a set of landmarks on the head, respectively. Barrón and Kakadiaris [22] and Wei and Chai [23] both find a single solution to the combinatorial problem by allowing a non-linear optimisation to converge to a local minimum. Weak-perspective scale is estimated using statistics of human anthropometry and rigid structure within the body, respectively. Agarwal and Triggs [24] adopt an appearance-based approach, learning a regression from silhouettes to 3D pose for a synthetic training set. This does not generalise well to a large range of pose configurations and silhouettes can be difficult to obtain.

A number of previous works in non-rigid reconstruction have entertained a temporal-differencing regularisation term (e.g. [25,26]). Valmadre and Lucey [5] recently argued that requiring that a trajectory lie on a low-frequency DCT basis is roughly equivalent to finding the trajectory with the minimum response to a similar such high-pass filter, since the DCT bases are the eigenvectors of symmetric convolution [27].

3 Problem Formulation

Let an articulated structure be an undirected graph with vertices $\mathcal{V} = \{1, \dots, p\}$ and edges $\mathcal{E}$. There exists a path between every pair of vertices and each edge $(i, j) \in \mathcal{E}$ is labeled with a length $\ell_{ij} > 0$. We will limit discussion to acyclic graphs to avoid degenerate cases which arise when the projection ray must intersect simultaneously with two spheres.

Let the configuration of an articulated structure at time t be denoted $\mathbf{x}_t = (\mathbf{x}_{t1}, \dots, \mathbf{x}_{tp}) \in \mathcal{R}^{3p}$, where $\mathbf{x}_{ti} \in \mathcal{R}^3$ is the 3D position of point i. Each edge in the graph provides the articulation constraint

$$\|\mathbf{x}_{tj} - \mathbf{x}_{ti}\|_2 = \ell_{ij}. \tag{1}$$

Each point i is observed at frame t by a pinhole camera $\mathbf{P}_t \in \mathcal{R}^{3\times 4}$ as projection $\mathbf{w}_{ti} \in \mathcal{R}^2$. Every observation further constrains the system by the projective equality

$$\begin{bmatrix}\mathbf{w}_{ti}\\1\end{bmatrix} \simeq \mathbf{P}_t \begin{bmatrix}\mathbf{x}_{ti}\\1\end{bmatrix} \qquad \Leftrightarrow \qquad \mathbf{Q}_{ti}\mathbf{x}_{ti} = \mathbf{u}_{ti}, \tag{2}$$

where (momentarily dropping subscripts t and i) $\mathbf{Q} = \mathbf{P}_{1:2,1:3} - \mathbf{w}\mathbf{P}_{3,1:3}$ and $\mathbf{u} = \mathbf{P}_{3,4}\mathbf{w} - \mathbf{P}_{1:2,4}$.[1] Each matrix $\mathbf{Q}_{ti}$ has a 1D right nullspace corresponding to the ray connecting the camera center and the projection on the image plane. The above linear constraint forms the algebraic error term for linear reconstruction algorithms in epipolar geometry.

The problem is thus to find the smoothest trajectory for all points $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{R}^{3np}$ which satisfies (1) and (2), given cameras $\{\mathbf{P}_t\}$, projections $\{\mathbf{w}_{ti}\}$ and the trajectory of a root node $\{\mathbf{x}_{t1}\}$, as well as the graph topology $\mathcal{E}$ and the articulation lengths $\{\ell_{ij}\}$.

3.1 Finite Set of Feasible Solutions

Park and Sheikh [1] recognised that combining equations (1) and (2) yields two feasible solutions per frame, representing the intersection of the projection ray with the articulation sphere (Figure 2). Unlike in [1], however, we give a general derivation of the binary ambiguity under the assumption of perspective *or* affine cameras using the nullspace vector of the linear projection equations instead of the camera center, which is only pertinent to perspective cameras.

Since each $\mathbf{Q}_{ti}$ is 2×3 and full rank, the position of a point can be decomposed

$$\mathbf{x}_{ti} = \mathbf{x}'_{ti} + \mathbf{x}^{\perp}_{ti}, \tag{3}$$

where $\mathbf{x}'_{ti} = \mathbf{Q}^{\dagger}_{ti}\mathbf{u}_{ti}$ lies in the row-space of $\mathbf{Q}_{ti}$ and $\mathbf{x}^{\perp}_{ti} = \alpha_{ti}\mathbf{q}^{\perp}_{ti}$ lies in the 1D nullspace of $\mathbf{Q}_{ti}$, which is spanned by $\mathbf{q}^{\perp}_{ti}$.[2] Substituting this result into the articulation constraint (1) yields a quadratic equation

$$\left\|\mathbf{x}'_{tj} + \alpha_{tj}\mathbf{q}^{\perp}_{tj} - \mathbf{x}_{ti}\right\|_2^2 = \ell^2_{ij} \tag{4}$$

which, assuming the articulation constraint can be satisfied, has two real roots

$$\alpha_{tj} = \bar{\alpha}_{tj}(\mathbf{x}_{ti}) \pm \varDelta\alpha_{tj}(\mathbf{x}_{ti}). \tag{5}$$

This is illustrated geometrically in Figure 2. We emphasise above that these roots depend on the position of parent joint i. Thus the two solutions are enumerated by a binary variable $s_{tj} \in \{-1, 1\}$,

$$\mathbf{x}_{tj}(s_{tj}, \mathbf{x}_{ti}) = \bar{\mathbf{x}}_{tj}(\mathbf{x}_{ti}) + s_{tj}\,\Delta\mathbf{x}_{tj}(\mathbf{x}_{ti}). \tag{6}$$

[1] The sub-matrix of rows $\{a, \ldots, b\}$ and columns $\{c, \ldots, d\}$ is denoted $\mathbf{A}_{a:b,c:d}$.

[2] The right matrix inverse is denoted $\mathbf{A}^{\dagger} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$.

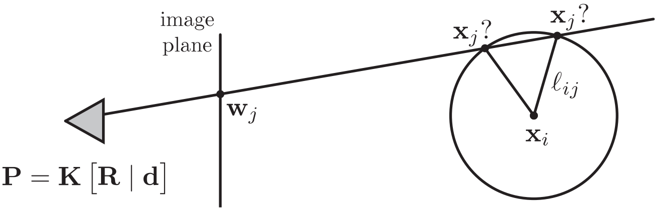


Fig. 2: With known camera $\mathbf{P}$, parent point $\mathbf{x}_i$, edge length ℓ_{ij} and 2D projection $\mathbf{w}_j$, there is a two-fold ambiguity over the position of the endpoint $\mathbf{x}_j$ in each frame. This is true of full perspective and affine cameras.

with $\bar{\mathbf{x}}_{tj}(\mathbf{x}_{ti}) = \mathbf{x}'_{tj} + \bar{\alpha}_{tj}(\mathbf{x}_{ti})\,\mathbf{q}^{\perp}_{tj}$ and $\Delta\mathbf{x}_{tj}(\mathbf{x}_{ti}) = \mathit{\Delta}\alpha_{tj}(\mathbf{x}_{ti})\,\mathbf{q}^{\perp}_{tj}$. A complete trajectory is described linearly in terms of a vector of signs $\mathbf{s}_j \in \{-1, 1\}^n$

$$\mathbf{x}_j(\mathbf{s}_j, \mathbf{x}_i) = \begin{bmatrix} \Delta\mathbf{x}_{1j}(\mathbf{x}_{1i}) & & \\ & \ddots & \\ & & \Delta\mathbf{x}_{nj}(\mathbf{x}_{ni}) \end{bmatrix} \mathbf{s}_j + \begin{bmatrix} \bar{\mathbf{x}}_{1j}(\mathbf{x}_{1i}) \\ \vdots \\ \bar{\mathbf{x}}_{nj}(\mathbf{x}_{ni}) \end{bmatrix}. \tag{7}$$

3.2 Temporal Prior

Trajectory basis NRSfM suggests that a matrix $\mathbf{X} \in \mathcal{R}^{n\times 3}$ describing a 3D trajectory of length n can be efficiently represented $\mathbf{X} = \mathbf{\Phi}\mathbf{B}$, where $\mathbf{\Phi} \in \mathcal{R}^{n\times b}$ is a b-dimensional orthonormal basis and $\mathbf{B} \in \mathcal{R}^{b\times 3}$ is the matrix of basis coefficients. Vectorising this expression, $\mathbf{x} = \mathbf{\Phi}_3\boldsymbol{\beta}$ with $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$ and $\mathbf{\Phi}_3 = \mathbf{\Phi} \otimes \mathbf{I}_3$.[3,4]

Following their previous work on trajectory basis reconstruction with known cameras [3], Park and Sheikh [1] sought an articulated trajectory which minimised the distance from trajectory space

$$E(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{M}_3} \tag{8}$$

where $\mathbf{M}_3 = \mathbf{M} \otimes \mathbf{I}_3$ with $\mathbf{M} = \mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^T$ being the orthogonal projector to the column-space of $\mathbf{\Phi}$.[5] The problem for a single trajectory is therefore to minimise

$$f(\mathbf{s}_j) = E\left(\mathbf{x}_j\left(\mathbf{s}_j, \mathbf{x}_i\right)\right). \tag{9}$$

Recent work [5] has shown that smoothness can alternatively be encouraged in a trajectory by penalising its response to a compact high-pass filter, instead constructing $\mathbf{M} = \mathbf{G}^T\mathbf{G}$, where

$$\mathbf{G} = \begin{bmatrix} g_m \cdots g_1 & & \\ \ddots & \ddots & \ddots \\ & & g_m \cdots g_1 \end{bmatrix} \tag{10}$$

[3] The vec($\cdot$) operator stacks the columns of a matrix.

[4] The $\otimes$ operator denotes the Kronecker (tiled) product.

[5] A Mahalanobis distance is denoted $\|\mathbf{x}\|^2_{\mathbf{A}} = \mathbf{x}^T\mathbf{A}\mathbf{x}$.

represents convolution by a filter $\mathbf{g} \in \mathcal{R}^m$ having support $m \ll n$. Filters are an elegant formulation in that temporal constraints are enforced locally rather than globally. This property is at the core of the efficient solution proposed in Section 4.3. When temporal prior is enforced using trajectory filters, the objective in (8) can be expressed

$$E(\mathbf{x}) = \|\mathbf{g} * \mathbf{x}\|_F^2 = \sum_{t=1}^{n-m+1} \|g_m \mathbf{x}_t + \cdots + g_1 \mathbf{x}_{t+m-1}\|_2^2 . \tag{11}$$

Assuming the trajectory of one joint in an articulated structure is known, for example by general trajectory reconstruction [1,5] or by fixing it to the rigid background, the trajectories of all joints may be found by recursively solving this for all of the joint's children. In the following section we will demonstrate that the above problem can be solved in polynomial time for compact filters. Our method also inherits the advantage of not having to specify the basis size b.

4 Dynamic Programming

4.1 Acyclic Graphs

The combinatorial optimisation problem of minimising a general function $f : \{1, \dots, k\}^n \to \mathcal{R}$ has exponential time complexity $\mathcal{O}(k^n)$. However, a well-known result is that if the objective can be expressed

$$f(\mathbf{x}) = \sum_{i=1}^{n} g_i(x_i) + \sum_{(i,j) \in \mathcal{E}} h_{ij}(x_i, x_j) \tag{12}$$

and the undirected graph defined by the edge set $\mathcal{E}$ is acyclic, then the global minimum can be found using dynamic programming in $\mathcal{O}(nk^2)$ time. In the machine learning community, this is known as the max-sum algorithm for doing inference in graphical models [28]. Note that $\mathcal{E}$ is unrelated to the definition of articulated structure in the previous section. Since this paper is primarily concerned with solving for time sequences, we restrict ourselves to the case where the graph is a chain,

$$f(\mathbf{x}) = \sum_{i=1}^{n} g_i(x_i) + \sum_{i=1}^{n-1} h_i(x_i, x_{i+1}). \tag{13}$$

Dynamic programming gives a recursive definition for "partial solutions"

$$f_i^*(x_{i+1}) = \min_{x_i} \left[f_{i-1}^*(x_i) + h_i(x_i, x_{i+1}) \right] + g_{i+1}(x_{i+1}) \tag{14}$$

such that the solution to the original problem is easily computed

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{x_n} f_{n-1}^*(x_n). \tag{15}$$

Given a table of the previous partial solution f_{i-1}^* evaluated for all x_i, the following partial solution f_i^* can be computed for all x_{i+1} in $\mathcal{O}(k^2)$ time. Starting with a trivial base case, the solution is found by induction in $\mathcal{O}(nk^2)$ time.

4.2 Sequences with Terms of m Consecutive Elements

The solution presented above is the special case of a lesser known result in combinatorial optimisation, which is that if the objective can be expressed as a sum of terms which depend on m consecutive elements

$$f(\mathbf{x}) = \sum_{i=1}^{n-m+1} h_i(x_i, \ldots, x_{i+m-1}), \tag{16}$$

then the global minimum can be found in $\mathcal{O}(nk^m)$ time using dynamic programming [29]. The recursive definition for a partial solution, which depends on $m-1$ elements and can be evaluated exhaustively in $\mathcal{O}(k^m)$ time, becomes

$$f_i^*(x_{i+1}, \ldots, x_{i+m-1}) = \min_{x_i} \left[f_{i-1}^*(x_i, \ldots, x_{i+m-2}) + h_i(x_i, \ldots, x_{i+m-1}) \right], \tag{17}$$

and the final solution is given

$$\min_{x_1, \ldots, x_n} f(x_1, \ldots, x_n) = \min_{x_{n-m+2}, \ldots, x_n} f_{n-m+1}^*(x_{n-m+2}, \ldots, x_n), \tag{18}$$

providing a solution by induction in $\mathcal{O}(nk^m)$ time.

4.3 Application to Filter Responses

The general distance-from-subspace objective in (8) can be expressed

$$f(\mathbf{s}) = \sum_{t=1}^{n} \sum_{u=1}^{n} a_{tu} \, s_t \, s_u + \sum_{t=1}^{n} b_t \, s_t + c, \tag{19}$$

highlighting the maximally-connected interaction between variables (Figure 1). However, the filter objective from (11) fits the form of (16), defining

$$h_{\mathbf{x},t}(s_t, \ldots, s_{t+m-1}) = \left\| g_m \, \mathbf{x}_t(s_t) + \cdots + g_1 \, \mathbf{x}_{t+m-1}(s_{t+m-1}) \right\|_2^2, \tag{20}$$

where each $\mathbf{x}_t$ maps $\{-1, 1\} \to \mathcal{R}^3$. This makes it possible to minimise the response of a sequence of length n to a filter with support m in $\mathcal{O}(n2^m)$ time by taking into account the known banded-diagonal structure of $\mathbf{M}$ in (8). Previous work [5] has found an effective filter choice to be a linear combination of the responses to $[1, -1]$ and $[-1, 2, -1]$ filters, arguing that longer-support filters simply introduce unwanted degrees of freedom. The framework generalises trivially to multiple filters by modifying $h_{\mathbf{x},t}$.

5 Calibration-less Reconstruction

Articulated trajectory reconstruction typically assumes that the camera and root node positions can be recovered from the background using rigid SfM. For many “real-world” video sequences of interest, however, the background may lack sufficient structure or visual texture to reliably estimate cameras in this manner. We recognise that, since articulated trajectories are immune to reconstructability issues which arise from camera motion, it may be practical to assume a constant camera and reconstruct the relative motion of the structure within the camera reference frame. Small non-smooth camera motion (jitter) can be removed using 2D stabilisation. For a full perspective camera, we would still require an estimate of the root trajectory in the camera reference frame. Adopting an affine camera model allows us to estimate a scale parameter instead.

If the object maintains approximately constant distance from the camera, reconstruction can be achieved by assuming constant scale $\sigma_t = 1$. If the object possesses approximately-rigid sub-structure, rigid structure from motion can be used to estimate scale [23, 30]. Finally, if the camera is zooming in and out, scale may be computed from background objects which maintain their relative out-of-plane orientation to the camera, such as trees and road-signs.

Once camera scale is known, the length of each edge in the articulation graph can be estimated by its maximum observed projection

$$\ell_{ij} = \max_t \frac{\|\mathbf{w}_{tj} - \mathbf{w}_{ti}\|_2}{\sigma_t} \tag{21}$$

as in [30]. This is a reasonable estimate due to the slow decay of the cosine function at the origin. This also ensures that the articulation and projection constraints will be feasible. If some edges are known to be of equal length, the maximum over several edges can be used to improve the estimate.

6 Experiments

The reconstruction accuracy and running time of our proposed dynamic programming algorithm have been evaluated on synthetic projections of the freely available CMU human body Motion Capture (MoCap) dataset.[6] The calibrationless extension has been demonstrated on real world examples.

6.1 Accuracy and Efficiency

The performance of our proposed algorithm was compared to an implementation of the branch and bound method using the same tools as [1] on synthetic projections of sequences from CMU MoCap. In these experiments the ground truth perspective camera (constant throughout the sequence) and root node trajectory were supplied to the algorithm. Note that [1] recommends following the discrete optimisation with a non-linear refinement. Since we propose an alternative to the discrete optimisation alone, there is no non-linear refinement in any of the experiments.

Figure 3 shows conclusively that dynamic programming is orders of magnitude more efficient for long sequences. In fact, despite employing a branch and

[6] `http://mocap.cs.cmu.edu/`

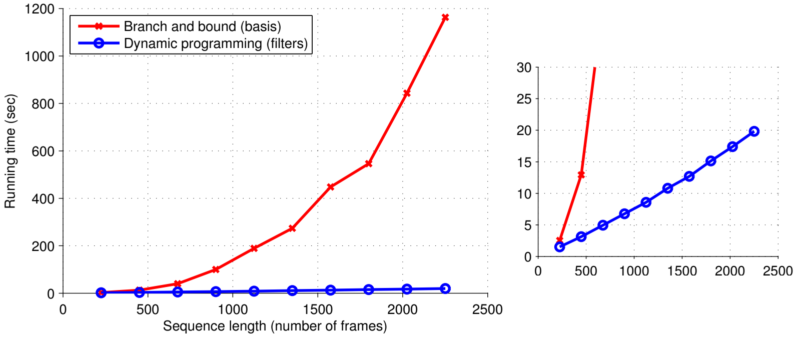


Fig. 3: Running time versus sequence length for the two reconstruction algorithms. The trajectory basis objective may still take exponential time to solve despite the branch and bound algorithm (*left*). The dynamic programming method guarantees a solution in linear time (*right*). Both are globally optimal. The experiment is for an 18-joint human body sequence from CMU MoCap.

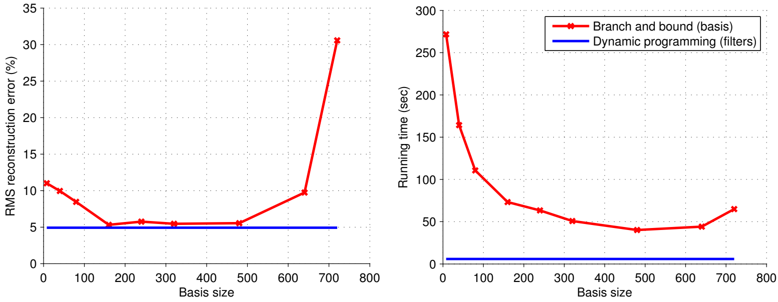
bound strategy, the time complexity of the competing method still appears to grow exponentially. Note that both implementations were written in Matlab and neither is highly optimised. The fact that the running times are very similar for short sequences suggests that this is a fair comparison.

Another advantage of the filter-based approach is that there is no need to specify a basis size. While Park and Sheikh [1] identified that articulated trajectory reconstruction is relatively insensitive to number of DCT bases used, Figure 4 shows that the basis size needs to be chosen relatively large, and failure to do this correctly will still result in a poor reconstruction. We also show empirically that choosing the wrong basis size affects the running time.

6.2 Calibration-less Reconstruction

Figure 5 compares the performance of general trajectory reconstruction with known cameras [3] to that of articulated trajectory reconstruction with i) known perspective cameras, known lengths and known root [1], ii) known weak perspective scale and known lengths and iii) known articulation topology only. Note that results are normalised to the ground truth by an optimal rigid transformation before comparison. This graph highlights that calibration-less articulated reconstruction complements the existing domain of trajectory reconstruction problems, providing a method which does not need camera estimation at the cost of decreased accuracy for faster cameras. In contrast, the original trajectory reconstruction algorithm [3] fails when camera motion is slow.

Reconstructions are presented in Figures 6, 7, 8, 9 and 10 for several real videos using the calibration-less approach. For the purposes of demonstration, 2D point correspondences were manually labeled. All of these sequences would be challenging or impossible cases for automatic full perspective camera estimation.



(a) Error versus basis size

(b) Running time versus basis size

Fig. 4: While the basis size in [1] is not critical, an incorrect choice still adversely affects the reconstruction (a). Filter prior obtains a slightly better reconstruction than the best basis reconstruction. The filter objective could be minimised using branch and bound but the reconstruction will be identical. The running time of the branch and bound method varies with the basis size (b). Results were averaged over 8 × 800-frame sequences.

7 Conclusion

This paper has presented a globally optimal solution to articulated trajectory reconstruction with worst-case time complexity linear in the number of frames instead of exponential. This demonstrates that there are major practical advantages to the theoretical insight that temporal prior can be imposed using compact trajectory filters rather than a trajectory basis. Experimental results show an order of magnitude speed increase for sequences that are thousands of frames long.

We also contribute a reformulation which incorporates affine as well as full perspective cameras, showing that this allows reconstruction without solving for perspective cameras if weak perspective scale can be heuristically estimated. Reconstructions are demonstrated for videos of humans and animals.

Acknowledgements

Dr. Simon Lucey is the recipient of an Australian Research Council Future Fellowship (project FT0991969).

References

1. Park, H.S., Sheikh, Y.A.: 3D reconstruction of a smooth articulated trajectory from a monocular image sequence. In: ICCV, IEEE (2011) 201–208
2. Akhter, I., Sheikh, Y.A., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space. In: NIPS, MIT Press (2008)

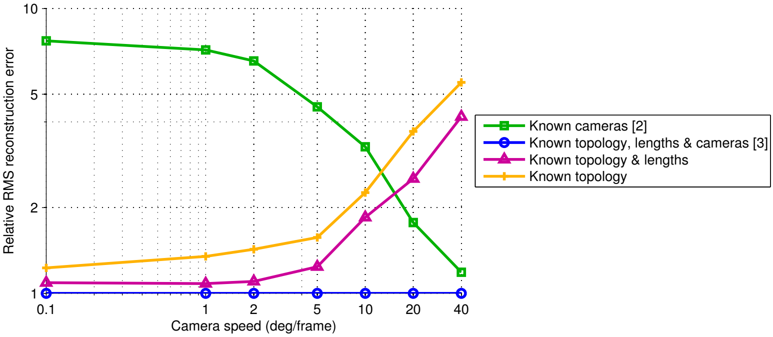


Fig. 5: RMS reconstruction error versus camera speed. Results are shown relative to Park and Sheikh's version of the articulated reconstruction problem [1], which assumes known cameras, articulation topology and edge lengths, and is unaffected by camera speed. As expected, Park et al.'s general trajectory reconstruction problem [3], which assumes only known cameras, becomes inaccurate without fast camera motion. Our calibration-less extension complements these approaches, assuming only known topology and lengths and relying on slow camera motion. When lengths are heuristically estimated, slightly worse results are obtained. Results are averaged over a subset of CMU MoCap sequences.

3. Park, H.S., Shiratori, T., Matthews, I., Sheikh, Y.A.: 3D reconstruction of a moving point from a series of 2D projections. In: ECCV. Volume 6313., Springer (2010) 158–171
4. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: CVPR. Volume 2., IEEE (2000) 690–696
5. Valmadre, J., Lucey, S.: General trajectory prior for non-rigid reconstruction. In: CVPR, IEEE (2012)
6. Zhu, Y., Cox, M., Lucey, S.: 3D motion reconstruction for real-world camera motion. In: CVPR, IEEE (2011)
7. Gotardo, P.F.U., Martinez, A.M.: Non-rigid structure from motion with complementary rank-3 spaces. In: CVPR, IEEE (2011) 3065–3072
8. Akhter, I., Simon, T., Khan, S., Matthews, I., Sheikh, Y.A.: Bilinear spatiotemporal basis models. In: SIGGRAPH, ACM Press (2012)
9. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. In: CVPR. (2012)
10. Tresadern, P.A., Reid, I.: Articulated structure from motion by factorization. In: CVPR. Volume 2., IEEE (2005) 1110–1115
11. Yan, J., Pollefeys, M.: A factorization-based approach to articulated motion recovery. In: CVPR. Volume 2., IEEE (2005) 815–821
12. Paladini, M., Del Bue, A., Stošić, M., Dodig, M., Xavier, J., Agapito, L.: Factorization for non-rigid and articulated structure using metric projections. In: CVPR, IEEE (2009) 2898–2905
13. Fayad, J., Russell, C., Agapito, L.: Automated articulated structure and 3D shape recovery from point correspondences. In: ICCV, IEEE (2011) 431–438
14. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3D human figures using 2D image motion. In: ECCV, Springer (2000) 702–718

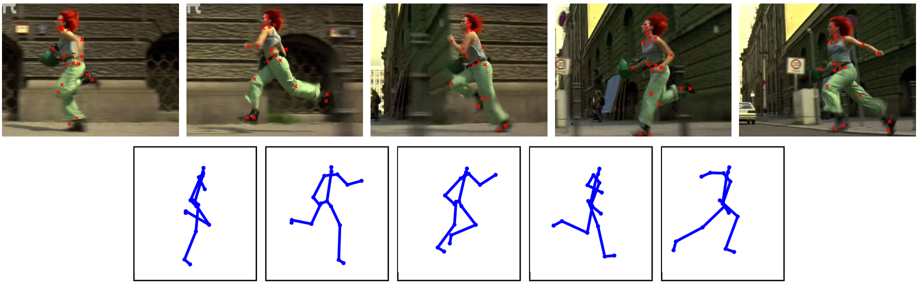


Fig. 6: The reconstruction of several frames of a sequence from the movie "Run Lola Run." Camera estimation would be difficult as significant perspective effects are only observed for a handful of frames. Reconstruction is shown from two novel views. Our human skeleton comprises 18 joints.

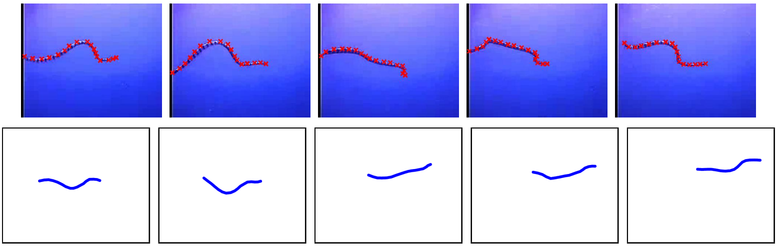


Fig. 7: Reconstruction of a sea snake filmed underwater by a diver. Note the absence of any rigid background. The skeleton consists of 17 joints.

15. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3D human tracking. In: CVPR. Volume 1., IEEE (2003) 69–76
16. Urtasun, R., Fleet, D.J., Fua, P.: 3D people tracking with Gaussian Process Dynamical Models. In: CVPR. Volume 1., IEEE (2006) 238–245
17. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR, IEEE (2010) 623–630
18. Brubaker, M.A., Fleet, D.J.: The Kneed Walker for human pose tracking. In: CVPR, IEEE (2008)
19. Wei, X., Chai, J.: VideoMocap: Modeling physically realistic human motion from monocular video sequences. In: SIGGRAPH, ACM Press (2010)
20. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: CVPR, IEEE (2000) 677–684
21. Parameswaran, V., Chellappa, R.: View independent human body pose estimation from a single perspective image. In: CVPR, IEEE (2004) 16–22
22. Barrón, C., Kakadiaris, I.A.: Estimating anthropometry and pose from a single uncalibrated image. Computer Vision and Image Understanding **81**(3) (2001) 269–284
23. Wei, X., Chai, J.: Modeling 3D human poses from uncalibrated monocular images. In: ICCV, IEEE (2009)
24. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. PAMI **28** (2006) 44–58

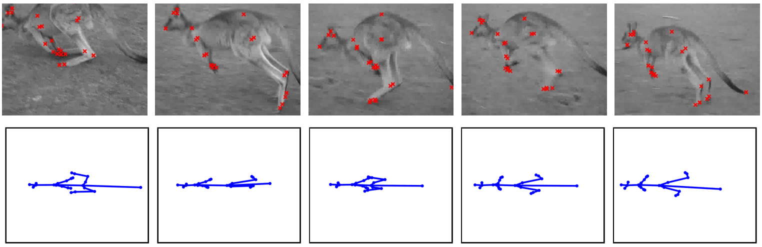


Fig. 8: Reconstruction of a jumping kangaroo. Insufficient background texture is available to reliably estimate cameras. The skeleton consists of 23 joints.

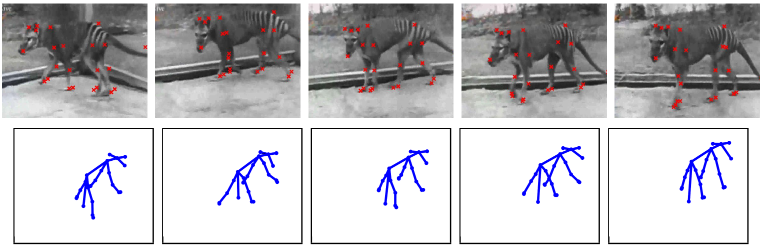


Fig. 9: Reconstruction from the last known footage of the extinct thylacine (Tasmanian tiger). The skeleton consists of 23 joints.

25. Salzmann, M., Urtasun, R.: Physically-based motion models for 3D tracking: A convex formulation. In: ICCV, IEEE (2011) 2064–2071
26. Micusik, B.: Trajectory reconstruction from non-overlapping surveillance cameras with relative depth ordering constraints. In: ICCV, IEEE (2011) 922–928
27. Martucci, S.A.: Symmetric convolution and the discrete sine and cosine transforms. IEEE Transactions on Signal Processing **42**(5) (1994) 1038–1051
28. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2007)
29. Felzenszwalb, P.F., Zabih, R.: Dynamic programming and graph algorithms in computer vision. PAMI **33**(4) (2011) 721–40
30. Valmadre, J., Lucey, S.: Deterministic 3D human pose estimation using rigid structure. In: ECCV, Springer (2010) 467–480

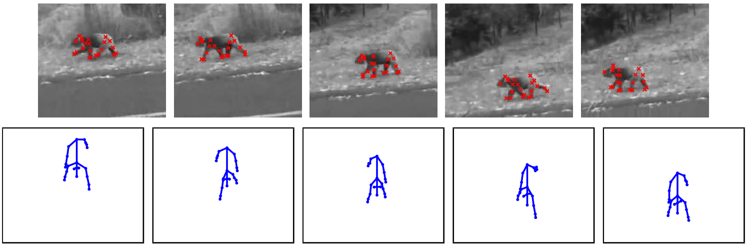


Fig. 10: Reconstruction of a walking koala from shaky video footage, obtained following video stabilisation. The skeleton consists of 22 joints.