Automated Articulated Structure and 3D Shape Recovery from Point Correspondences

Anonymous ICCV submission

Paper ID 272

Abstract

In this paper we propose a new method for the simultaneous segmentation and 3D reconstruction of interest point based articulated motion. We decompose a set of point tracks into rigid-bodied overlapping regions which are associated with skeletal links, while joint centres can be derived from the regions of overlap. This allows us to formulate the problem of 3D reconstruction as one of model assignment, where each model corresponds to the motion and shape parameters of an articulated body part. We show how this labelling can be optimised using a combination of pre-existing graph-cut based inference, and robust structure from motion factorization techniques. The strength of our approach comes from viewing both the decomposition into parts, and the 3D reconstruction as the optimisation of a single cost function, namely the image re-projection error. We show results of full 3D shape recovery on challenging real-world sequences with one or more articulated bodies, in the presence of outliers and missing data.

1. Introduction

Recovering the 3D structure of non-rigid objects (either deformable or articulated) from monocular sequences is a longstanding challenge in computer vision. One problem of particular interest is human motion analysis which involves the estimation of the 3D motion of articulated bodies from video streams. Estimating the 3D pose of the human body purely from image data taken by a single camera is an important problem with applications ranging from bio-mechanics to cinema post-production, computer gaming, animation and human behaviour analysis. In this paper we focus on the specific case of full 3D reconstruction using only the 2D positions of interest points tracked over time.

Most algorithms which perform 3D pose estimation of articulated bodies require prior knowledge of a model of its underlying structure, given by a kinematic chain [17, 2, 18], which often requires manual intervention to create. This high level of intervention is undesirable in many circumstances, for example in animation or gaming, an actor should be able to pick up and interact with a rigid object, effectively augmenting their skeletal structure, without the need for a graphical artist to generate a new model. Given this predefined 3D skeleton model, these approaches track the articulated motion, estimating the joint angles, but do not usually recover the full 3D shape of the object. In this paper we take a different approach, and demonstrate how the estimation of the full 3D shape, motion, and underlying skeletal structure of one or more articulated bodies can be derived directly from 2D correspondences in a single camera video sequence — without models.

Previous approaches to recovering both articulated structure and motion purely from 2D tracking data include factorization methods [30, 22] which model articulated motion as a set of intersecting motion subspaces. These methods require two steps: first a motion segmentation algorithm separates the 2D trajectories into different articulated parts; and second a factorization approach is used to estimate joint positions and articulation axes. Yan and Pollefeys [30] follow this with a third step that builds the kinematic chain automatically from the segmented subspaces. Each articulated part can be recovered as a rigid shape using the factorization method [20].

Such pipeline approaches are inherently unstable, a failure in any of the early stages of reconstruction can not be
recovered from, and such difficulties are often unapparent until the final reconstruction fails. As an alternative to this multi-stage formulation, we propose an algorithm which performs a simultaneous decomposition of the articulated body into its constituent parts and reconstructs the full 3D shape of the object, revealing its skeletal structure.

We tackle articulated reconstruction from 2D tracks using a piecewise approach. We assume that an articulated object can be approximated by a set of rigid segments that overlap on the joints. The key idea is to segment the object into its constituent rigid segments, while enforcing overlap between neighbouring segments. The body parts are reconstructed independently and points in the overlapping regions are then used to stitch them together to create a full 3D articulated body. Our assumption of rigidity for the individual links could be relaxed later – after model assignment, a non-rigid reconstruction algorithm could be applied to each of the body parts as a post-processing.

Piecewise solutions have been applied before with success to the 3D reconstruction of piecewise planar rigid scenes [1] and more recently to deformable surfaces [6, 16, 19, 24]. Our work demonstrates that they are equally applicable to the problem of articulated structure from motion. Our approach distinguishes itself from these as, on articulated data, it estimates semantically meaningful rigid parts and gives the location of joints, rather than returning surface regions. Compared to the works [6, 16] we do not require an initial estimate of a rest shape.

The strength of our approach comes from viewing both the decomposition into parts, and the 3D reconstruction as the optimisation of a single cost function, namely the image re-projection error, subject to a spatial constraint that neighbouring points should also belong to the same model. This gives us the ability to switch back and forth from the assignment of points to parts, and fitting a rigid model to the parts, in a hill-climbing approach, allowing us to recover from previous mistakes and refine our current model estimates as we go. In practice, we make use of the rigid estimation techniques of [10] and the model assignment techniques of [16].

A significant advantage of our formulation over previous motion segmentation algorithms [26, 5] is that it does not require the number of motions to be known in advance, and we exploit the spatial prior that points which are physically close are likely to belong to the same model. Our only necessary assumptions are that we find a minimum of three tracked interest points on each rigid part, which is needed to perform 3D reconstruction, and that at least one point is located in the intersection of body parts — this last constraint is due to the fact that we rely on points belonging to multiple models to guarantee the spatial consistency of the global 3D shape. Both of these constraints are guaranteed by our inference model, providing that each point has at least two neighbours, and that the graph of points in the human skeleton is path connected. See section 2.1 for more details.

**Contribution** This paper is the first to perform simultaneous segmentation and 3D reconstruction of articulated motion from 2D point tracks, making us substantially more robust, and providing an improvement in the quality of reconstruction. While previous methods have generated 3D information in order to extract skeletal structures [30], we are the first to show the resulting 3D reconstruction on complex articulated structures and not just articulated pairs [22] and to evaluate the quality of our 3D models.

### 1.1. Related work

**3D Pose Estimation** The problem of 3D pose estimation from a monocular video sequence is an important one and evidence of this is the large number of works that have addressed it in recent years. An exhaustive review is out of the scope of this paper but we refer the reader to [7] for a more complete overview. Two broad classes of strategies have been used for 3D pose inference: *Generative (top-down)* algorithms optimise a cost function to align an appearance based 3D model with image features [17, 2]; *Discriminative (bottom-up or recognition-based)* use training sets of (pose; image) pairs to recognise the pose in a specific image. While generative approaches require prior knowledge of a 3D kinematic model and often require manual initialisation, discriminative methods are dependent on the amount and quality of the training data.

Our approach falls into a different category of methods, those termed *articulated structure-from-motion algorithms (A-SfM)* which use only 2D correspondences as input and
focus on the inference of the 3D shape and motion of the object without the use of a model \([30, 22, 11]\) – a model-free and data driven approach. These algorithms require an initial motion segmentation step to divide the object into its constituent parts.

**Motion Segmentation.** Motion segmentation is a particularly challenging problem in the case of articulated motion due to the dependencies between the linked parts. The original solution to the multi-body segmentation problem \([3]\), based on rigid factorization \([20]\), was influential but unable to solve problems containing dependent motions. This was remedied by \([31]\), who built an affinity matrix from the data and used its dominant eigenvectors to separate dependent motions. However, it performed poorly in the presence of articulated motion. The GPCA algebraic framework \([27, 25]\) can also deal with dependent subspaces and missing data. However, in practice it cannot be applied to more than a few subspaces as the number of required samples grows exponentially with the number of subspaces.

Concerning articulated SfM methods, while Tresadern and Reid \([22]\) used a RANSAC \([21]\) approach to segmentation, Yan and Pollefeys proposed a segmentation algorithm specifically designed to tackle the articulated motion case \([29]\). A set of linear subspaces is estimated through local sampling and an affinity matrix is built computing the principal angles between them, followed by spectral clustering to give the segmentation result. Despite outperforming all other motion segmentation algorithms in the cases of articulated motion, this algorithm is highly dependent on the correct detection of the rank of the trajectories, and consequently is sensitive to noise.

Our approach is closer to the energy based multiple model fitting approach to motion segmentation proposed by PEARL \([8]\) that reformulates it as a labeling problem where both the labels (model parameters) and their assignment to data points are computed simultaneously. However, the technique described in PEARL can only be applied to disjoint motions and not to sequences of articulated motion.

**Articulated Structure from Motion (A-SfM)** Articulated SfM algorithms model articulated motion as a set of intersecting motion subspaces — the intersection of two motion subspaces implies the existence of a connection between the two corresponding parts. After segmentation articulation constraints are imposed during factorization to recover the location of joints and axes \([22, 28]\). While Tresadern and Reid \([22]\) only deal with articulated pairs, Yan and Pollefeys go further \([30]\) estimating the kinematic chain of articulated objects with a more complex structure, by building the minimum spanning tree from the segmented subspaces. Factorization is first used to recover the shape and motion of the individual parts, then joints and axes are calculated and used to combine parts into a single coordinate system, and recovering the articulated shape and motion as a whole. Ross et al. \([15]\) instead propose a probabilistic approach to learn the structure of an articulated object while inferring its pose given a time series of 2D feature tracks. The drawbacks of this method are that it generally places joints in the middle of segments, instead of the endpoints and has difficulty to recover from a poor initialisation.

These A-SfM methods methods offer an attractive solution to 3D pose estimation since they are model-free and do not require any training data. However they suffer from a number of weaknesses: the quality of the motion segmentation step is critical for a successful 3D reconstruction — misclassified points can lead to large errors in the motion and shape estimates; motion segmentation algorithms on the other hand either require the number of constituent object parts to be known in advance or are sensitive to its mis-estimation via the detection of the rank of the trajectories; missing data cannot be dealt with; and finally, their demonstrations are on simple articulated motions, crucially never on a full body. These algorithms focus less on the full 3D reconstruction of the body and more on estimating the location of joints and articulation axes to estimate the skeleton structure.

In this paper we advance the state of the art in articulated structure from motion by addressing these weaknesses. We jointly solve the problems of segmentation and 3D reconstruction via optimisation of a single cost function which allows us to recover from incorrect solutions and does not require the number of articulated body parts to be known in advance. We describe a strategy to deal with missing data in the measurements in a principled manner, by directly incorporating partial sequences into our optimisation. In segmentation, imposing the spatial constraint that neighbouring points should belong to the same model results in increased robustness and fewer mismatches in our approach. Finally, we demonstrate our complete system on challenging full body human articulated sequences.

## 2. Problem Formulation

Consider an articulated object described by a set of \(P\) point tracks, observed by an orthographic camera in a sequence of \(F\) image frames. We assume this articulated object can be accurately approximated by a set of rigid segments that form an articulated forest. We make no as-

---

1. Our assumption of an orthographic camera, is a simplifying one and need not hold in the general case. Reconstruction from a projective camera can be performed in the same manner as we describe, simply by using bundle adjustment \([23]\) instead of \([10]\).

2. Again, this is a simplifying assumption, the fact that the graph formed is a forest (i.e. contains no cycles) is not used in either the fitting of points to models or the assignment of models to points. However, the absence of cycles does guarantee that there are no impossible to resolve constraints.
We consider a set of point tracks $P$, and assume that tracks spatially adjacent to one another are connected in a graph structure. We express this by writing that each point track $p$ is connected to a set of neighbours $N_p$ (see section 2.4 for details on how the neighbourhood is built). Given this graph, and a set of models $M$, where each model is parameterized with the rotation and translation associated with a rigid motion. We choose an overlapping assignment of models to points $m = \{m_1, m_2, \ldots, m_P\}$ by optimising the following cost function:

$$\text{arg min}_{m \in (2^P)^P} C(m) = \sum_{p \in P} \left( \sum_{\alpha \in m_p} U_p(\alpha) \right) + \text{MDL}(m), \quad (1)$$

subject to the following constraints:

$$\forall p \in P \exists \alpha : p \in I_\alpha; \quad (2)$$

$$\forall q \in N_p \land q \in I_\alpha \implies \alpha \in m_p, \quad (3)$$

where $m_p$ is the subset of models assigned to point $p$, $U_p(\alpha)$ is the cost associated with assigning point $p$ to model $\alpha$, (computed as the image reprojection error given in (6)), and $\text{MDL}(m)$ is a minimum description length prior [9, 4], that penalises the total number of active models used to explain the data.

The notation $p \in I_\alpha$ is short-hand for “$p$ is an interior point of model $\alpha$”, where, as in topology, an interior point of a model or set $\alpha$, is defined as one whose neighbours must also belong to $\alpha$. As such, constraint (3) is the definition of an interior point; while constraint (2) states that every point must be an interior point of at least one model.

Note that this differs from a conventional MRF formulation in that: firstly, while a particular point must belong to at least one model, it may belong to multiple models; and secondly, two neighbouring points in the graph must always share at least one model in common – this condition is enforced by constraints (2,3). This condition that neighbouring points must share models functions as a smoothing constraint, eliminating outliers and encouraging the use of a single model to explain spatially coherent regions. To optimise this problem we use the recent work [16] which proposed an extension to $\alpha$-expansion that allows the application of efficient graph-cut based inference to these multiple-model assignment problems.

2.2. 3D Reconstruction of Rigid Segments

Given a set of points that belong to the same rigid body segment, the model associated with them will be described by the 3D coordinates of each point $s_j = [X_j Y_j Z_j]$ and the set of rotation matrices and translation vectors that align the shape with the coordinate system at each frame. In this work we consider an orthographic camera model, a good mathematical approximation of the imaging process when

3We use the term active model, to refer to a model which has at least one point belonging to it.
the relief of the object is small considered to its distance to the camera — a valid assumption in the case of the human body. The image coordinates of point \( s_j \) at image \( i \) will be given by:

\[
\mathbf{w}_{ij} = \begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \begin{bmatrix} X_j \\ Y_j \\ Z_j \end{bmatrix} + \mathbf{t}_i = \mathbf{R}_i \mathbf{s}_j + \mathbf{t}_i
\]

(4)

where \( \mathbf{R}_i \) is a \( 2 \times 3 \) matrix containing the first two rows of the rotation matrix (i.e. \( \mathbf{R}_i \mathbf{R}_i^T = \mathbf{I}_{2 \times 2} \)) and \( \mathbf{t}_i \) is a 2-vector describing the translation. In the case of full data, if the image coordinates \( w_{ij} \) are registered to the image centroid, the translation can be eliminated. Stacking the image coordinates of all \( P \) points in all \( F \) frames gives the registered measurement matrix

\[
\begin{bmatrix}
\mathbf{w}_1 \\
\mathbf{w}_2 \\
\vdots \\
\mathbf{w}_F
\end{bmatrix} =
\begin{bmatrix}
\mathbf{R}_1 \\
\mathbf{R}_2 \\
\vdots \\
\mathbf{R}_F
\end{bmatrix}
\begin{bmatrix}
\mathbf{s}_1 \\
\mathbf{s}_2 \\
\vdots \\
\mathbf{s}_P
\end{bmatrix} = \mathbf{RS}.
\]

(5)

Estimating the model parameters \( \mathbf{R} \) and \( \mathbf{S} \) for each rigid segment can be formulated as the factorization problem [20] which minimizes image reprojection error:

\[
\arg \min_{\mathbf{R}, \mathbf{S}} \sum_{i=1}^{F} ||\mathbf{w}_i - \mathbf{RS}||^2 \quad \text{s.t.} \quad \mathbf{R}_i \mathbf{R}_i^T = \mathbf{I}_{2 \times 2}.
\]

(6)

In this paper instead of using the classical solution to factorization [20] we adopt the solution of Marques and Costeira [10] which has the advantage of providing rotation matrices that are guaranteed to lie on the constraint manifold, and allows us to deal with missing data.

2.3. Aligning body segments

Since each segment is reconstructed independently in its own reference frame, to obtain a final reconstruction the unconnected body parts must be stitched together into a single articulated body. This can be achieved by aligning the shared points between segments imposing the constraint that they must have the same 3D coordinates. A disambiguation step is needed to solve for their relative depths and reflection states. This step is common to all piecewise reconstruction methods [6, 16, 19, 24]. First the relative depth ambiguity is solved by registering the \( z \) coordinate of the centroids of both overlapping regions.

Second, the reflection, or sign, ambiguity is solved by choosing, from the two possible configurations, the one that results in a more consistent motion between the two overlapping segments. We measure this as the sum of the 3D Euclidean distance between all pairs of corresponding points in the overlap. This will ensure that we choose the configuration that results in overlapping points representing the same 3D points in space.

2.4. Guaranteeing a Valid Reconstruction

As previously mentioned, our approach has two requirements, (i) to perform a reconstruction at least 3 points must belong to each active model, and (ii) to reduce the sign, or depth ambiguity, to a single binary decision per skeletal structure, models must be path connected by overlapping regions i.e. if model \( A \) intersects with \( B \) and \( B \) intersects with \( C \), there is only one sign ambiguity to resolve for the entirety of \( A, B \), and \( C \).

Both of these properties are guaranteed by our inference approach. Property (i) holds for any neighbourhood structure in which every point has at least two neighbours. For a model \( \alpha \) to be active, it must be an interior model of at least one point, \( \alpha \in I_p \). If this point \( p \) is neighbours with at least two points \( q \) and \( r \) then \( \alpha \in m_q \) and \( \alpha \in m_r \) by constraint (3).

Property (ii) holds providing the underlying neighbourhood structure is path connected. This is a consequence of the fact that, if two points \( p_1 \) and \( p_2 \) are path connected by the sequence \( \{ p_1, p_2, \ldots, p_n \} \), the models \( \alpha_1 \) and \( \alpha_n \) are also path connected by the sequence of models \( \{ \alpha_1, \alpha_2, \ldots, \alpha_n \} \) where \( \alpha_k \in I_{p_k} \). This last statement holds as (3) guarantees that the interior models of neighbouring points must overlap. Neither of these properties need hold in a conventional MRF, such as those used by [8] where each point only belongs to exactly one model, and an active model may only have one point assigned to it.

Choice of neighbourhood structure The neighbourhood structure used by the algorithm depends on, both the distance measure chosen to tell how far apart points are, and a graph-building technique such as k-nearest neighbours, or minimum spanning tree. Ideally, the distance function chosen should capture the distances on the body between points i.e. it should separate the lower torso from the elbow. [16] suggested that the distance between points should be taken as the average distance between points over all frames. However [14] observed that the maximum distance between points over all frames is a better measure of the distance within a deformable object. We take a hybrid approach, and look at the average of the five greatest distances between a pair of points. This provides the benefits of [14], while remaining robust to noise.

To guarantee that properties (i) and (ii) hold, some care must be taken when choosing the neighbourhood structure of the graph. For example, use of k-nearest neighbours where \( k \geq 2 \) would guarantee property (i), while use of a minimum spanning tree would guarantee property (ii). There seems to be no standard method that guarantees the required properties, and does not lead to an over-connect graph, so in practice we use 6-nearest neighbour as an initialisation and add it additional minimum cost edges, un-
til we force the existence of at least two paths that do not share edges between every pair of points.

2.4.1 Initialisation

To initialise our approach we must propose a set of possible labels and corresponding model parameters to each of the \( P \) points. To avoid becoming stuck in a bad local optimum, we initialise with an excess of models, choosing the initial set \( M \) by fitting one model to each point \( p \in P \), and all of its neighbours. Given these initial labels the initial model parameters can be naturally recovered using the method described in Section 2.2.

2.5. Missing Data and Multiple Articulated Objects

Neither the graph-cut based inference of section 2.1, nor the reconstruction algorithm of 2.2 requires complete point tracks, and can be applied to partial tracks. The only difficulty with the use of partial tracks is the generation of their neighbourhood structure. As some points are only visible for a short period of time, they may well be linked to the wrong section of the body, for example, points on the arm may be mistakenly linked to those on the torso, and while this may give a good reconstruction for the frames in which the points are visible, in other frames it can leave artefacts. To avoid these difficulties, we include points with partial data directly in our framework, but give them an empty neighbourhood. Because of the MDL prior, these points without neighbours will belong to a common model used elsewhere in the reconstruction. The procedure is equivalent to assigning partial tracks to the active model which minimises the reconstruction error.

3. Experimental Results

We evaluate our approach on some of the more challenging articulated sequences in the literature. First, against the ‘dance’(figures 4, 5), ‘digger’ (figure 2), and ‘toy’ (figure 3) sequences from [30], and further on the ‘skin’ sequence of [12] (figure 6), which has 3 dimensional ground truth. Despite the relatively poor quality, and under connected neighbourhoods (in both human cases, the torso can be separated into two sections linked only by a single point); the neighbour structure is sufficient to guaranteed properties (i), and (ii) of section 2.4, and the decomposition into models and final reconstruction is convincing. Compared to [30], our assignment of points to models is much smoother, with no outliers. This can be attributed to our requirement that adjacent models must overlap, which functions as a smoothing term, suppressing outliers. Our failure to identify the head, as a model separate from the body in the ‘dance’ data-set (figure 5) can also attributed to the same smoothing term. In this sequence, the points on top of the head are incorrectly tracked, and [30] labels them as belonging to the torso (see figure 4). With relatively few points belonging to the head, and with points on either side of it belonging to the torso, the smoothing effects acts to suppress it.

We perform substantially better than [30] on the ‘digger’ data-set (figure 2), showing our approach to be both more robust to outliers (c.f. blue point bottom row, far left), and more discriminative, as we both detect the motion of the bucket on the right most digger, and successfully separate movement of the third digger in the background. Note that for this sequence, we followed [30] in thresholding the size of connecting edges – this allows for multiple disconnected objects.

The ‘skin’ data-set from [12] was acquired using a Motion Capture setup consisting of 12 infra-red cameras tracking the 3D positions of approximately 350 reflective markers placed in a regular grid pattern on the body of a male subject. This provides us with ground truth 3D locations to evaluate our reconstructions. We evaluate our performance by orthographically projecting the 3D sequence, and then using our method to recover the 3D coordinates. We obtain a mean reconstruction error of 8.10% on this sequence, without taking into account the hands (see figure 6 column 1), the error falls to 6.56%.

4. Conclusion

We have presented a novel data-driven approach for the problems of simultaneous segmentation and 3D reconstruction of articulated motion. Without making any assumptions of about the skeletal structure of the object we reconstruct, we are able to obtain both high quality 3D reconstructions, and semantically meaningful decomposition into articulated parts. Compared to existing motion segmentation approaches, we strongly benefit from spatial smoothing priors, which both increase our robustness to outliers, and make it easier for us to recover semantically informative segmentations. We improve substantially on previous articulated SFM methods which were only demonstrated on simple two bodied sequences with full data, by demonstrating our complete system on challenging full body human articulated sequences and providing a principled solution to dealing with missing data.

Figure 4. Point assignment of dance sequence due to [30] please compare with column 1 of figure 5, and see section 3.
Figure 5. Best viewed in colour. Reconstruction results from the ‘dance’ data-set [30]. From left to right 1. Original image and point location and decomposition. 2. Generated neighbourhood structure using the technique described in section 2.4. 3. Resulting decomposition into rigid overlapping models and estimated 3d reconstructions. 4. Estimated skeletal structure, and model assignment. Note that each node represents an intersection between two rigid models, and each edge the connecting model between two points. The location of the nodes is found by averaging all points which lie in the intersection.

References


Figure 6. Best viewed in colour. Reconstruction results from the data-set [13]. From left to right 1. Comparison between original ground truth (black) and location estimated (green) from a novel view point. 2. Generated neighbourhood structure using the technique described in section 2.4. 3 and 4. Resulting decomposition into rigid overlapping models and estimated 3d reconstructions. 5 and 6. Estimated skeletal structure from different view points. Note that each node represents an intersection between two rigid models, and each edge the connecting model between two points.


