Modeling 3D Human Poses from Uncalibrated Monocular Images

Xiaolin K. Wei
Texas A&M University
xwei@cse.tamu.edu

Jinxiang Chai
Texas A&M University
jchai@cse.tamu.edu

Abstract

This paper introduces an efficient algorithm that reconstructs 3D human poses as well as camera parameters from a small number of 2D point correspondences obtained from uncalibrated monocular images. This problem is challenging because 2D image constraints (e.g. 2D point correspondences) are often not sufficient to determine 3D poses of an articulated object. The key idea of this paper is to identify a set of new constraints and use them to eliminate the ambiguity of 3D pose reconstruction. We also develop an optimization process to simultaneously reconstruct both human poses and camera parameters from various forms of reconstruction constraints. We demonstrate the power and effectiveness of our system by evaluating the performance of the algorithm on both real and synthetic data. We show the algorithm can accurately reconstruct 3D poses and camera parameters from a wide variety of real images, including internet photos and key frames extracted from monocular video sequences.

1. Introduction

A long standing challenge in computer vision is to reconstruct 3D structures of non-static objects (*e.g.*, deforming faces or articulated human bodies) from monocular image sequences. This paper presents an efficient algorithm to simultaneously reconstruct 3D human poses and camera parameters using a small number of 2D point correspondences obtained from images. The problem is challenging because 2D image constraints are often not sufficient to determine 3D poses of an articulated object. To address this problem, we introduce a set of new constraints which can be used to eliminate the ambiguity of 3D pose reconstruction. Our analysis shows that, under weak perspective projection model, we need at least five images to accurately reconstruct 3D human poses and camera parameters.

Mathematically, we formulate the reconstruction problem in a continuous optimization framework by maximizing the consistency between the reconstructed 3D poses and 2D point correspondences. We develop an efficient optimization process to find 3D human poses and camera parameters that best match various forms of the reconstruction constraints. The whole optimization process consists of two sequential optimization steps. We first estimate human skeletal size and camera parameters in an efficient gradient-based optimization framework, and then use the estimated skeletal size and camera parameters to reconstruct 3D human poses in joint angle space. The two-step optimization strategy not only significantly speeds up the whole reconstruction process but also enables the optimization process to avoid falling in local minima.

We demonstrate the performance of our system by evaluating the algorithm on both real and synthetic data. We show the algorithm can accurately reconstruct 3D joint angle poses, skeletal bone lengths, and camera parameters from internet photos or multiple frames from monocular video streams. The quality of the reconstruction results produced by our system depends on the number of input images and accuracies of 2D joint locations specified by the user. We, therefore, evaluate how increasing or decreasing the number of input images influences the 3D reconstruction error. We also evaluate the robustness of the reconstruction algorithm under different levels of input noise.

2. Background

Our system simultaneously reconstructs 3D human poses and camera parameters from a small number of 2D point correspondences obtained from monocular images. This section briefly reviews related work in reconstructing 3D articulated objects from monocular images.

Our work builds on the success of previous work in reconstructing 3D articulated objects from a single image [8, 2, 6]. Modeling 3D articulated objects from a single 2D image, however, is an inherently "ill-posed" problem. Previous approaches often rely on known skeletal size [8, 6] or strong anthropometry prior [2] to reduce the reconstruction ambiguity. For example, Taylor [8] assumed a known skeletal size, and present an analytical solution to recover 3D orientation of the bone segments up to an undetermined weak perspective camera scale. Barron and Kakadiaris [2] extended the idea to estimate both anthropometric param-

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

eters and a human body pose from a single image. They formulate the problem as a nonlinear optimization problem and impose a multivariate Gaussian prior on bone lengths to constrain the solution space. Parameswaran and Chellappa [6] solved the same problem with known skeletal size, accounting for projective foreshortening effects of a simplified skeleton model. Our work differs from previous work because our algorithm simultaneously reconstructs skeletal size, articulated poses and camera parameters from multiple monocular images. The use of multiple images allows us to eliminate the reconstruction ambiguity caused by unknown skeletal size and camera parameters.

Another approach to reconstruct 3D human poses from monocular images is to use data-driven techniques to reduce the reconstruction ambiguity. Previous work in this direction either learns the mapping between 2D image features (*e.g.* silhouettes) and 3D poses [7, 1, 3], or constructs pose priors to constrain the solution space [4]. This approach, however, has not demonstrated they can accurately reconstruct 3D poses with unknown skeletal size and camera parameters. Another limitation of the data-driven approach is that it can only model poses that are similar to the training data. Our approach does not have this limitation, and can model arbitrary human poses from multiple monocular images.

A number of researchers have also extended the factorization methods [9] to reconstruct articulated objects from monocular images [10, 11]. For example, Tresadern and Reid [10] developed a factorization method to recover segment lengths as well as joint angles using a large number of feature trajectories across the entire monocular sequence. Yan and Pollefeys [11] used the rank constraints to segment the feature trajectories and then built the kinematic chain as a minimum spanning tree of a graph constructed from the segmented motion subspaces. In order to utilize rank constraints for factorization and segmentation, this approach often requires a large number of 2D visible features across the entire sequences. Our approach is different because we introduce a set of new constraints for 3D articulated reconstruction and formulate the reconstruction problem in a continuous optimization framework. One benefit of the optimization framework is its ability to deal with missing features. More importantly, the number of corresponding features required for our algorithm is significantly smaller than the factorization methods.

3. Problem Statement

Given 2D joint locations of an articulated object at K frames, our goal is to reconstruct 3D joint locations relative to the camera. Without loss of generality, we focus our discussion on human skeletal models, though the basic reconstruction scheme that will be proposed in this section can easily be extended to recover other articulated objects.

Throughout the paper, we assume a weak perspective projection model, which is valid when the average variation of the depth of an articulated object along the line of sight is small compared to the distance between the camera and object. The unknown camera parameters are the scalars of the weak perspective projection across the K frames $\mathbf{s} = (s_1, ..., s_K)^T$, where s_k is the camera scale at frame $k, k = 1, ..., K$.

Our human skeletal model consists of $B = 17$ bones (see Figure 1.(a)): head, neck, back, and left and right clavicle, humerus, radius, hip, femur, tibia, and metatarsal. Let $\mathbf{l} = (l_1, ..., l_B)^T$ represent bone segment lengths, where $l_i, i = 1, ..., B$ is the length of the i-th bone. Since we are dealing with images obtained by single-view cameras, the absolute length of a bone segment cannot be inferred from the images. We, therefore, normalize the lengths for all the bones by assuming $l_1 = 1$.

We follow the representation of Taylor [8]. The 3D pose of an articulated object at frame k is thus represented by bone segment lengths $\mathbf{l} = (l_1, ..., l_B)^T$ and relative depth values of two end points for all bone segments $\mathbf{dz}_k = (dz_{k,1}, ..., dz_{k,B})^T$. Let $\mathbf{dz}$ be a long vector which stacks $\mathbf{dz}_k$ across all frames.

The goal of our paper is to reconstruct unknown camera parameters $\mathbf{s}$, bone lengths $\mathbf{l}$, and relative depths $\mathbf{dz}$, from 2D joint locations $\mathbf{x}_{k,i} = (u_{k,i}, v_{k,i})^T, k = 1, ..., K, i = 1, ..., B + 1$ at monocular frames.

4. Constraints for 3D Articulated Reconstruction

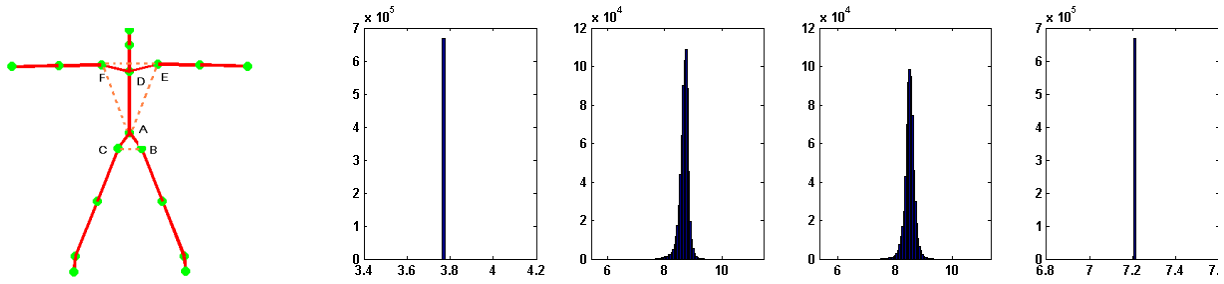
Simultaneous reconstruction of 3D human poses from monocular image sequences is challenging and often ambiguous. As discussed in Taylor [8], the solution to this reconstruction problem is up to an unknown weak perspective scale even with known skeletal size. The key idea of our paper is to identify a number number of new constraints which can be used to remove the reconstruction ambiguity.

4.1. Bone projection constraints

Taylor [8] introduced bone projection constraints for recovering 3D articulated poses with known skeletal size. The bone projection constraints consider the relationship between 3D end points of a bone segment and their 2D projections. Under the weak perspective projection model, the 3D coordinates of a point $\mathbf{p} = (x, y, z)^T$ in the scene and the 2D coordinates $\mathbf{x} = (u, v)^T$ in the image space should satisfy the following equation:

$$\begin{aligned} u &= sx \\ v &= sy \end{aligned} \tag{1}$$

where the scalar s denotes the unknown camera scale for a weak perspective camera projection. Note that we describe



(a) (b)

Figure 1. (a) Our human skeleton model consists of 17 bones and a full-body human pose is represented by 6 degrees of freedom (DoFs) for root position and orientation, and 31 DoFs for joint angles. (b) The histograms of the distances of $\overline{BC}$, $\overline{AE}$, $\overline{AF}$ and $\overline{EF}$ (from left to right), which are computed from millions of human poses from the online CMU mocap database, show the effectiveness of four rigid body constraints.

the 3D coordinates of the point with respect to the camera.

We now consider the projection of the i bone segment onto the image space. The length of the i-th bone can be computed as follows:

$$l_i^2 = \|\mathbf{p}_{i_1} - \mathbf{p}_{i_2}\|^2, \ i = 1, ..., B \tag{2}$$

where the vectors $\mathbf{p}_{i_1}$ and $\mathbf{p}_{i_2}$ represent the 3D coordinates of two end points for the i-th bone segment respectively. After combining Equation (1) with Equation (2), we obtain the following bone projection constraints:

$$dz_i^2 = l_i^2 - \frac{\|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|^2}{s^2}, \ i = 1, ..., B \tag{3}$$

where dz_i represents the relative depth of two end points: $z_{i_1} - z_{i_2}$.

A simple extension of Taylor [8] to multiple monocular images will not work because the bone projection constraints are not sufficient to reconstruct the skeletal size, 3D poses and camera parameters at the same time. The total number of bone projection constraints (KB) is much smaller than the total number of unknowns ($KB + K + B - 1$), which include KB unknown relative depths of bone segments, K unknown camera scales, and $B - 1$ unknown bone lengths.

4.2. Bone symmetry constraints

Studies in anthropometry show human skeletons satisfy symmetry properties. We therefore can use bone symmetry constraints to reduce the solution space. We impose bone symmetry constraints on seven bones, including clavicle, humerus, radius, hip, femur, tibia, and metatarsal. In a mathematical term, we have

$$l_{i_1}^2 = l_{i_2}^2, \tag{4}$$

where i_1 and i_2 are indices to symmetric bones. Totally, we have seven bone symmetry constraints.

The solution to the reconstruction problem is still not unique because the number of constraints ($KB + 7$) is still fewer than the number of unknowns ($KB + K + B - 1$).

4.3. Rigid body constraints

We introduce a new set of constraints–rigid body constraints–to eliminate the reconstruction ambiguity for 3D articulated poses. The rigid body constraints consider the relationship of three points located on the same rigid body. The constraints preserve the distances between *any* two points regardless of the movement of a human body. For example, for a standard human body skeletal model, the root, left and right hip joints, which are often assumed to be located on the same rigid body bone, satisfies the rigid body constraints.

We define four rigid body constraints based on the joints located on the torso: $\triangle ABC$, $\triangle ADE$, $\triangle ADF$ and $\triangle DEF$ (see Figure 1.(a)). To evaluate how well these four constraints are satisfied for human skeletal models, we compute the distances of $\overline{BC}$, $\overline{AE}$, $\overline{AF}$ and $\overline{EF}$ using millions of prerecorded human poses in the online CMU mocap database[1]. Figure 1.(b) shows the distance distributions for the four rigid body constraints. The distance distributions show the four rigid-body constraints are well satisfied.

Let us consider the rigid body constraints for one triangle (*e.g.*, $\triangle ABC$). Due to the rigid body property, the relative depths of the point A, B and C should satisfy the following condition:

$$dz_{B,C} = dz_{A,B} - dz_{A,C}, \tag{5}$$

where $dz_{B,C}$, $dz_{A,B}$, and $dz_{A,C}$ represent the relative depths of the points A, B, and C. After combining Equation (5) with Equation (3), we obtain the following rigid body constraints:

$$l_{B,C}^2 - \frac{\|\mathbf{x}_B - \mathbf{x}_C\|^2}{s^2} = (dz_{A,B} - dz_{A,C})^2 \tag{6}$$

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

where $l_{B,C}$ is the length of the line segment BC. Equation (6) can also be rewritten as an equation of dz^2:

$$(l_{B,C}^2 - \frac{\|\mathbf{x}_B - \mathbf{x}_C\|^2}{s^2} - dz_{A,B}^2 - dz_{A,C}^2)^2 = 4dz_{A,B}^2 dz_{A,C}^2 \quad (7)$$

In total, we enforce $4K$ rigid body constraints across all the frames. We also introduce four new unknowns $\mathbf{e} = (l_{B,C}, l_{A,F}, l_{A,E}, l_{E,F})^T$ to the reconstruction problem.

5. Simultaneous Reconstruction of Skeletal Size and Camera Parameters

We now discuss how to use constraints described in Section 4 to reconstruct human skeletal size and camera parameters as well as relative depth values based on 2D joint locations defined in K monocular frames. We formulate the reconstruction problem in a continuous unconstrained optimization framework. We define our constraints as "soft" constraints because 2D joint locations extracted from images often contain noise.

However, even with known skeletal size and camera parameters, there are still two possible solutions for the relative depths dz_i of the bone segments, representing the pose ambiguity that has been previously discussed in the work of Taylor [8]. According to Equation 3, the two possible solutions for relative depths are

$$dz_i = \pm\sqrt{l_i^2 - s^2\|\mathbf{x}_{i,1} - \mathbf{x}_{i,2}\|^2}. \quad (8)$$

The equation shows that the relative depths of the bone segments are up to an unknown sign.

To address this ambiguity, we choose to optimize the objective function with respect to squares of all unknowns. We stack squares of all unknowns into a long vector $\mathbf{X} = \{\mathbf{l}^2, \bar{\mathbf{s}}^2, \mathbf{dz}^2, \mathbf{e}^2\}$, where $\bar{\mathbf{s}} = (\frac{1}{s_1}, ..., \frac{1}{s_K})^T$. The overall objective function includes three terms described in Equation (3), (4) and (7):

$$\arg\min_{\mathbf{X}} E_p(\mathbf{l}^2, \bar{\mathbf{s}}^2, \mathbf{dz}^2) + \lambda_1 E_s(\mathbf{l}^2) + \lambda_2 E_r(\mathbf{e}^2, \bar{\mathbf{s}}^2, \mathbf{dz}^2) \quad (9)$$

where E_p, E_s, and E_r represent bone projection, bone symmetry and rigid body constraints respectively. The three terms represents square differences between the left and right side of the Equation 3, 4 and 7 respectively. The weights λ_1 and λ_2 control the importance of each constraint term.

We analytically derive the Jacobian of the object function and then run the optimization with the Levenberg-Marquardt algorithm in the Levmar library [5]. The optimization converges very quickly. We set initial values for l_i^2, $\overline{s_k}^2$, $dz_{k,i}^2$, e_i^2 to 10, 0.1, 0, 10 respectively.

5.1. How many images are needed?

The reconstruction problem consists of squares of four groups of unknowns: the relative depths of the bone segments across the K frames $\mathbf{dz}^2 \in R^{KB}$, the camera parameters $\mathbf{s}^2 \in R^K$, the lengths of the bone segments $\mathbf{l}^2 \in R^{B-1}$, and the lengths of four extra bone segments $\mathbf{e}^2 \in R^4$ in rigid body constraints. Therefore, there are totally $KB + K + B + 3$ unknowns. One intriguing question here is how many constraints are needed for an unique 3D reconstruction of human skeletal lengths and camera parameters.

We assume constraints are independent of each other. If we want to uniquely reconstruct all unknowns without *any* ambiguity, we need at least the same number of constraints. In total, we have KB bone projection constraints, $4K$ rigid-body constraints, and 7 bone symmetry constraints.

In sum, we use $KB + 4K + 7$ independent constraints to reconstruct $KB + K + B + 3$ unknowns. To have a unique reconstruction, we need to ensure that the number of constraints is not lower than the number of unknowns.

$$\begin{aligned} KB + 4K + 7 &\geq KB + K + B + 3 \\ K &\geq \tfrac{B-4}{3} \\ K &\geq \tfrac{13}{3} \end{aligned} \quad (10)$$

According to Equation 10, we need at least five key frames to remove the reconstruction ambiguity. Otherwise, the system will be ill-posed and produce ambiguous results.

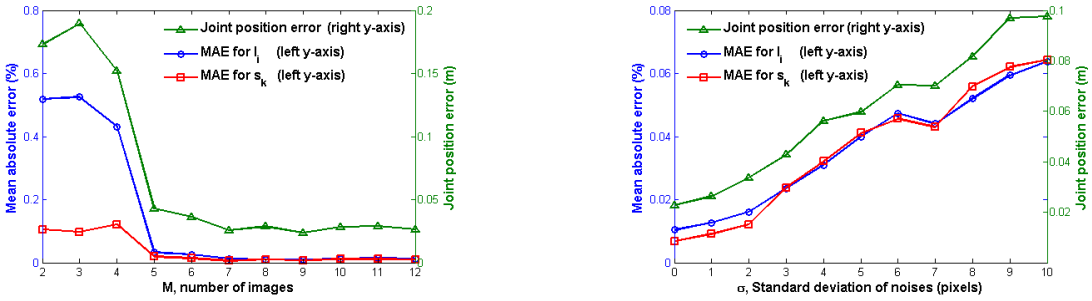
6. Reconstruction of 3D Joint-angle Poses

For most applications (e.g., human motion tracking or video-based motion capture), we need to estimate both human skeletal size and 3D human poses. This section discusses how to reconstruct 3D human poses using the reconstructed bone lengths $\mathbf{l}$ and $\mathbf{e}$, camera parameters $\mathbf{s}$, and squares of relative depths $\mathbf{dz}^2$.

We cannot directly compute the 3D joint positions from $\mathbf{dz}^2$ due to undetermined signs for $\mathbf{dz}$ (see Equation 8). There are a finite number of possible poses given the estimated $\mathbf{dz}^2$. To reduce this ambiguity, we enforce joint angle limit constraints from biomechanics community and solve the 3D human poses in the joint angle space in order to efficiently incorporate them.

We represent a full-body human pose with joint-angle values of 17 joints. These joints are head (2 Dofs), neck (2 Dofs), back (3 Dofs), and left and right clavicle (2 Dofs), humerus (3 Dofs), radius (1 Dofs), femur(3 Dofs), tibia (1 Dofs), and metatarsal (2 Dofs). A full-body pose is, therefore, represented by a 37 dimensional vector. Let the vector $\mathbf{q}_k \in R^{37}$ denote the joint-angle pose at frame k. The 3D coordinates of joint i at frame k can be computed as a function of the joint-angle pose $\mathbf{q}_k$ and bone lengths $\mathbf{l}$:

$$\mathbf{p}_{k,i} = \mathbf{f}_i(\mathbf{q}_k; \mathbf{l}) \quad (11)$$



(a) (b)

Figure 2. Quantitative evaluation on synthetic data: (a) Reconstruction errors of 3D poses, bone lengths, and camera scales vs. the different numbers (K) of images used for reconstruction. (b) Reconstruction errors vs. different levels of input noise (σ).

where the vector function $\mathbf{f}_i$ is a forward kinematics function which maps a joint-angle pose $\mathbf{q}_k$ and skeletal size $\mathbf{l}$ to 3D coordinates of joint i.

Similarly, we formulate the 3D pose recovery problem as a constrained nonlinear optimization problem. The constrained optimization formulation finds the unknowns that minimize the objective function while satisfying the joint angle limit constraints:

$$\begin{aligned} &\arg\min_{\{\mathbf{q}_k\}} \; E'_p + \lambda_1 E'_r \\ &s.t. \quad \mathbf{q}^l \leq \mathbf{q}_k \leq \mathbf{q}^u \end{aligned} \tag{12}$$

where E'_p and E'_r represent bone projection and rigid body constraints in the joint angle space respectively. The vector $\mathbf{q}^l$ and $\mathbf{q}^u$ are the lower and upper bounds for 3D human poses respectively, which are adopted from biomechanics community. The weight λ_1 tradeoffs the importance of the following two constraint terms.

The bone projection term, E'_p, measures consistency between the reconstructed poses and 2D joint locations in images. More specifically, it computes the distance between the projections of 3D joints and specified 2D joint locations:

$$E'_p = \sum_{k=1}^{K} \sum_{i=1}^{B} (s_k f_i^x(\mathbf{q}_k;\mathbf{l}) - u_{k,i})^2 + (s_k f_i^y(\mathbf{q}_k;\mathbf{l}) - v_{k,i})^2 \tag{13}$$

where the function f_i^x and f_i^y represent the x and y coordinates of the i-th joint. The scalars $u_{k,i}$ and $v_{k,i}$ 2D coordinates of the i-th joint at frame k.

The rigid body term, E'_r, ensures the lengths of extra bone segments $\mathbf{e}$ remain constant across all frames:

$$E'_r = \sum_{k=1}^{K} \sum_{i=1}^{4} (\|\mathbf{f}_{i_1}(\mathbf{q}_k;\mathbf{l}) - \mathbf{f}_{i_2}(\mathbf{q}_k;\mathbf{l})\| - e_i)^2 \tag{14}$$

where $\mathbf{f}_{i_1}$ and $\mathbf{f}_{i_2}$ represent 3D joint locations of two end points of the i-th extra bone segment.

We initialize the optimization in the joint angle space with the reconstruction results in the position space. More specifically, the values of $\mathbf{dz}$ are initialized by the root square of the estimated $\mathbf{dz}^2$ with random signs. The initial joint angle values $\mathbf{q}_k$ are then computed by applying inverse kinematics to all of the reconstructed joint locations. We optimize the constrained objective function using the Levenberg-Marquardt algorithm with boundary constraints in the Levmar library [5]. The optimization process converges fast because of known skeletal lengths and camera parameters.

The joint angle limit constraints significantly reduce the ambiguity in $\mathbf{dz}^2$. But for some poses, they are not sufficient to remove all the ambiguity. When this happens, we allow the user to specify the sign of dz_i for bones that still have the ambiguity and run the optimization again.

7. Experimental Results

The performance of our reconstruction algorithm have been evaluated in a number of experiments.

7.1. Quantitative evaluation on synthetic and real data

The quality of the reconstruction results produced by our system depends on the number of input images and accuracies of input constraints. We, therefore, have evaluated how increasing or decreasing the number of input images influences the 3D reconstruction error. We have also evaluated the robustness of the reconstruction algorithm under different levels of noise.

Minimum number of images. We evaluated the reconstruction errors for different number of images (K). We randomly selected K 3D poses from the online CMU mocap database and render each pose with random camera parameters. We evaluated the 3D reconstruction error by comparing with ground-truth data. Figure 2.(a) shows the recon-

Figure 3. Internet photos: ten images are used for reconstruction and the reconstructed poses in four images are shown from the original viewpoint and a new viewpoint.

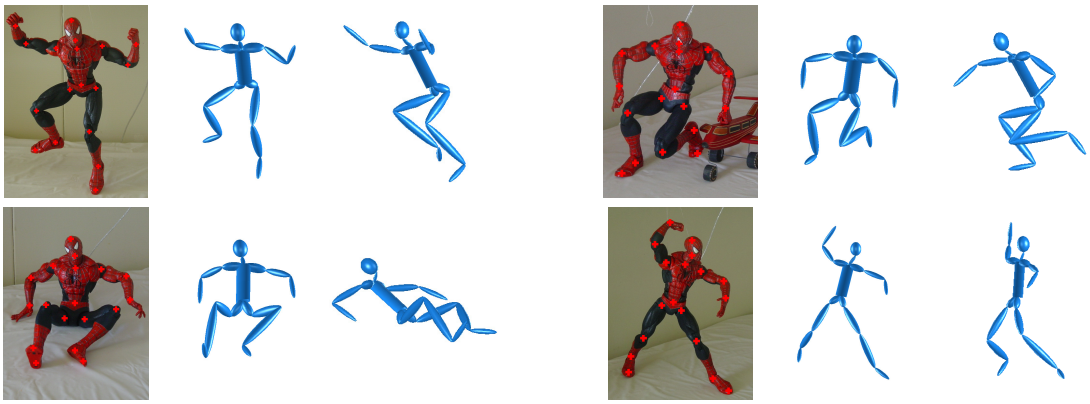


Figure 4. An articulated spiderman: twelve images are used for reconstruction and the reconstructed poses in four images are shown from the original viewpoint and a new viewpoint.

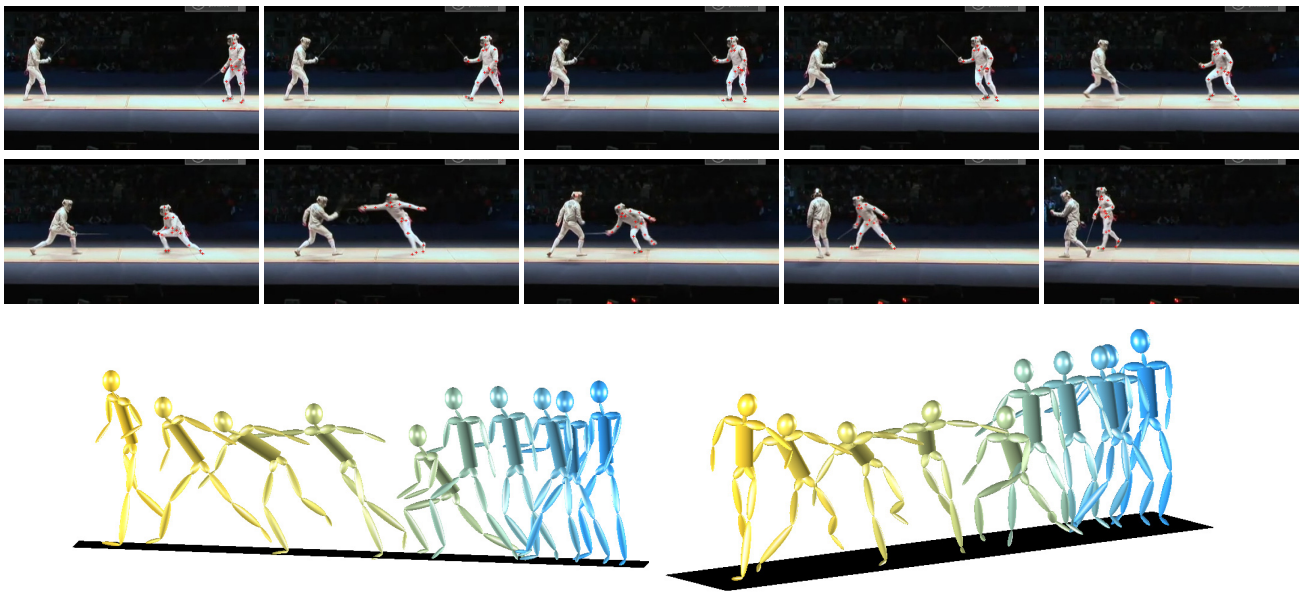


Figure 5. Key frames from monocular video stream. (top) Ten key frames selected from a 135-frame video sequence are used for reconstruction. (bottom) The reconstructed key poses are rendered from different viewpoints.

struction errors. The errors of 3D poses are computed as average Euclidean distances of 3D joints (in meters). The errors of bone lengths **l** and camera scales **s** are measured as the mean absolute reconstruction errors (MAEs). Figure 2.(a) confirms our analysis on the minimum number (five) of images needed for accurate reconstruction. The reconstruction errors are very large when the number of images is less than five. For example, the relative errors for bone length reconstruction are higher than 40% when only four images are used for reconstruction. But when we use the minimum number of images (five) to reconstruct bone lengths, the relative errors drop significantly to 2%.

Robustness to noisy constraints. The joint locations specified by user (*e.g.*, 2D joint locations) are often noisy. This experiment evaluated the robustness of our algorithm under various levels of input noise using the same CMU mocap database. We randomly selected 15 human poses from the database, synthesized 2D joint locations for each pose, and added Gaussian white noise to these 2D joint locations subsequently . We controlled the noise level with the standard deviation (σ). Figure 2.(b) reports the reconstruction errors under different levels of noise.

7.2. Qualitative evaluation on real images

We evaluated the performance of our system by testing the algorithm on a number of real monocular images. Our results are best seen in the accompanying video.

Internet photos. We downloaded ten random photos from a popular basketball player. Figure 3 shows four of the reconstructed 3D human poses from the original camera viewpoint and a new viewpoint.

Articulated toy. We have tested the performance of our algorithm on an articulated toy – spiderman. We posed the toy in twelve different poses and take a snapshot for each pose. The camera was about one meter away from the toy. Figure 4 shows four of the reconstructed poses from the original camera viewpoint and a new viewpoint.

Monocular video sequence. This experiment demonstrates that our algorithm can be used for reconstructing 3D poses and unknown human skeleton size from a small set of key frames extracted from monocular video streams. We tested our reconstruction algorithm on ten key frames of an Olympic fencing video download from internet [2] The reconstruction results are shown in Figure 5.

Missing features. Our algorithm is also capable of handling missing feature points from input images. Figure 6 shows the reconstruction results with images containing missing features. Twelve images are used for 3D reconstruction and four of the reconstructed poses are rendered from a new viewpoint.

Among all testing examples, the poses from monocular images are different and the joint locations specified by the user are highlighted with red crosses.

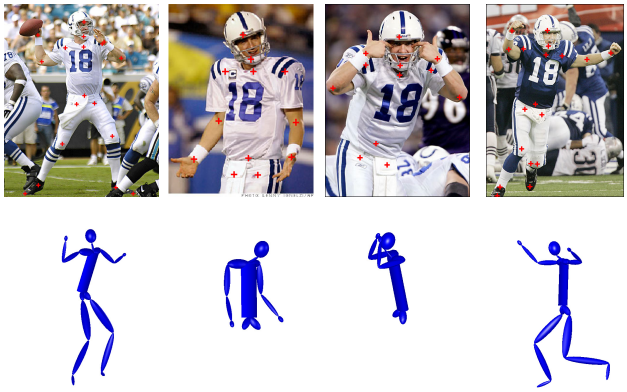


Figure 6. Reconstruction with missing data. (top) The input images containing missing features. (bottom) The 3D reconstructed poses rendered from a new viewpoint. Twelve images are used for 3D reconstruction.

7.3. Comparisons

Comparison with optical mocap systems. We compared our reconstruction results with ground truth data obtained from a twelve-camera optical motion capture system *Vicon*. The CMU online motion capture database[3] includes a wide variety of 3D human mocap data as well as synchronized monocular video sequences. We selected ten key frames from a "basketball" video sequence and reconstructed 3D human poses with our algorithm. Figure 7 shows a side-by-side comparison between ground-truth poses and reconstructed poses. The numerical errors of the 3D reconstructed poses is about 5 centimeter every joint. Note that the Vicon mocap system uses a different type of skeletal model. Before numerical evaluation, we need to map the 3D joint positions estimated by the Vicon system to our skeletal model.

Comparison with anthropometric prior. Anthropometric prior is often used to constrain the solution space for human skeletal estimation. We have done an experiment to evaluate the reconstruction accuracy with and without anthropometric prior. We applied Principle component analysis (PCA) to all skeletal models in the online CMU mocap database (5 bases to keep 99% of the energy), and modeled a Gaussian anthropometric prior in the 5D space. We used the priors to constrain the solution space of the bone lengths. Our preliminary results (via cross validation) showed the prior did not improve the reconstruction results, and sometimes it even produced worse results. This might be due to the effectiveness of the current system and the

[2] http://www.youtube.com

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

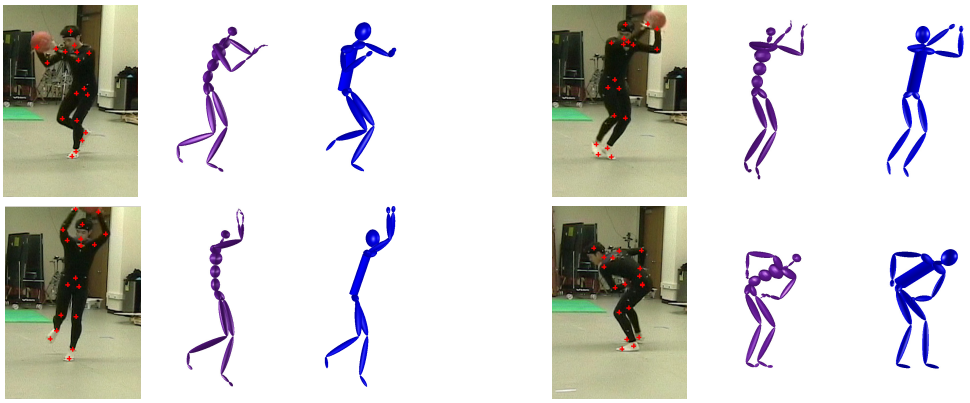


Figure 7. Side-by-side comparisons between the reconstructed poses and ground truth data from optical mocap systems: (left) input images with annotated 2D joint location; (middle) 3D poses recorded by a twelve-camera vicon system in a new viewpoint. (right) 3D poses reconstructed by our algorithm.

weak generalization ability of the priors.

8. Conclusion

We present a new technique for simultaneous reconstruction of 3D articulated poses and camera parameters from uncalibrated monocular images. One nice property of our algorithm is that our system does not require any prior knowledge on 3D poses or skeletal lengths. The key idea of the paper is to identify a number of new constraints to eliminate the reconstruction ambiguities. We formulate the reconstruction problem in a nonlinear optimization framework by maximizing the consistency between 3D poses and reconstruction constraints. Our analysis and experiments show that we need at least five single view images to accurately reconstruct the lengths of an unknown symmetric human skeletal model. Our experiments also show the algorithm can efficiently reconstruct 3D human poses from a variety of source images such as internet photos or monocular video streams.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2: 882–888.

[2] C. Barron and I. A. Kakadiaris. Estimating anthropometry and pose from a single image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000. 1:669-676.

[3] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2: 681–688.

[4] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2: 334-341.

[5] M. Lourakis. levmar: Levenberg-marquardt algorithms in C/C++. http://www.ics.forth.gr/∼lourakis/levmar, Feb. 2009.

[6] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2: 16-22.

[7] R. Rosales and S. Sclaroff. Infering body pose without tracking body parts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000. 2: 506–511.

[8] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Computer Vision and Image Understanding*, 2000. 80(3): 349-363.

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

[10] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2: 1110 - 1115.

[11] J. Yan and M. Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 1: 712-719.