Reconstructing 3D Human Pose from 2D Image Landmarks

Varun Ramakrishna, Takeo Kanade, Yaser Sheikh

Robotics Institute, Carnegie Mellon University
{vramakri,tk,yaser}@cs.cmu.edu

Abstract. Reconstructing an arbitrary configuration of 3D points from their projection in an image is an ill-posed problem. When the points hold semantic meaning, such as anatomical landmarks on a body, human observers can often infer a plausible 3D configuration, drawing on extensive visual memory. We present an *activity-independent* method to recover the 3D configuration of a human figure from 2D locations of anatomical landmarks in a single image, leveraging a large motion capture corpus as a proxy for visual memory. Our method solves for anthropometrically regular body pose and explicitly estimates the camera via a matching pursuit algorithm operating on the image projections. Anthropometric regularity (i.e., that limbs obey known proportions) is a highly informative prior, but directly applying such constraints is intractable. Instead, we enforce a necessary condition on the sum of squared limb-lengths that can be solved for in closed form to discourage implausible configurations in 3D. We evaluate performance on a wide variety of human poses captured from different viewpoints and show generalization to novel 3D configurations and robustness to missing data.

1 Introduction

Figure 1(a) shows the 2D projection of a 3D body configuration. From this 2D projection alone, human observers are able to effortlessly organize the anatomical landmarks in three-dimensions and guess the relative position of the camera. Geometrically, the problem of estimating the 3D configuration of points from their 2D projections is ill-posed, even when fitting a known 3D skeleton[1]. With human observers, the ambiguity is likely resolved by leveraging vast memories of likely 3D configurations of humans [2]. A reasonable proxy for such experience is available in the form of motion capture libraries [3], which contain millions of 3D configurations. The computational challenge is to tractably generalize from the configurations spanned in the corpus, ensuring anthropometric plausibility while discouraging impossible configurations.

[1] As noted in [1], each 2D end-point of a limb subtends a ray in 3D space. A sphere of radius equal to the length of the limb centered at any location on one of these rays intersects the other ray at two points (in general) producing a tuple of possible 3D limb configurations for each location on the ray.

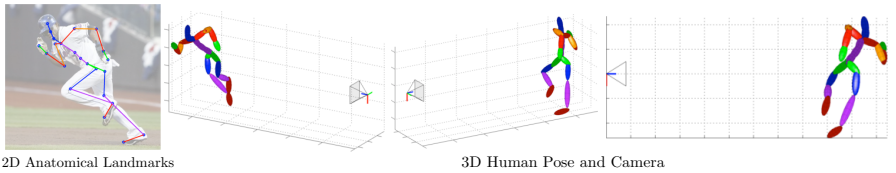


Fig. 1. Given the 2D location of anatomical landmarks on an image, we estimate the 3D configuration of the human as well as the relative pose of the camera.

Kinematic representations of human pose are high-dimensional and difficult to estimate directly. Allowing only statistically plausible configurations leads to compact representations that can be estimated from data. Linear dimensionality reduction (such as PCA) is attractive as it yields tractable and optimal estimation methods. It has been successfully applied to constrained deformable objects, such as faces [4] and action-specific body reconstruction, such as walking, [5]. However, as we add poses from varied actions, the complexity of the distribution of poses increases and, consequently, the dimensionality of the reduced model needs to be increased (see Figure 2). If we expand the dimensionality, linear models increasingly allow configurations that violate anthropometric constraints such as limb proportions, yet yield a projection in 2D that is plausible. The goal is therefore to develop an activity-independent model while ensuring anthropometric regularity.

In this paper, we present a method to reconstruct 3D human pose while maintaining compaction, anthropometric regularity, and tractability. To achieve compaction, we separate camera pose variability from the intrinsic deformability of the human body (because combining both leads to an approximately six-fold increase in the number of parameters [6]). To compactly model the intrinsic deformability across multiple actions, we use a sparse linear representation in an overcomplete dictionary. We estimate the parameters of this sparse linear representation with a matching pursuit algorithm. Enforcing anthropometric regularity through strict limb length constraints is intractable because satisfying multiple quadratic equality constraints on a least squares system is nonconvex [7]. Instead, we encourage anthropometric regularity by enforcing a necessary condition (i.e., an equality constraint on the sum of squared lengths) as a constraint that is applied in closed form [8]. We solve for the model coefficients and camera pose within the matching pursuit iterations, decreasing the reprojection error objective in each iteration.

Our core contributions are: (1) a new activity-independent representation of 3D human pose variability as a sparse embedding in an overcomplete dictionary, and (2) an algorithm, Projected Matching Pursuit, to estimate the sparse model from only 2D projections while encouraging anthropometric regularity. Within the matching pursuit iterations, we explicitly estimate both the 3D camera pose and the 3D body configuration. We evaluate our method to test generalization,

and robustness to noise and missing landmarks. We compare against a standard linear dimensionality reduction baseline and a nearest neighbor baseline.

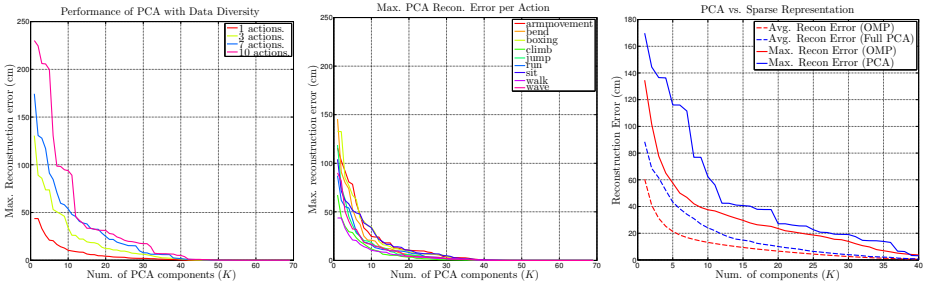
2 Related Work

For the single image pose recovery task, some of the earliest work is by Lee and Chen [1] who assumed known limb lengths and recovered pose by pruning a binary interpretation tree that enumerates the entire set of configurations for an articulated body using physical and structural pruning rules and user input. Taylor's approach [9] used known skeletal sizes to recover 3D pose up to a weak perspective scale; this method required human input to resolve the depth ambiguities at each joint. Jiang [10] used Taylor's method [9] to generate hypotheses followed by a nearest neighbor approach to prune the hypotheses. Parameswaran and Chellappa [11] used a strong prior on skeletal size and employed 3D model based invariants to recover the joint angle configuration but made restrictive assumptions on the 3D configurations possible. Other approaches, such as Barron and Kakadiaris [12], estimated anthropometry and pose using strong anthropometric priors on limb lengths by generating a set of plausible poses based on geometric constraints followed by a nonlinear minimization.

Discriminative approaches [13–16] have attempted to directly learn a mapping from 2D image measurements to 3D pose. Several approaches have recovered 3D pose from silhouettes. Elgammal and Lee [17] learned view-based activity manifolds from 2D silhouette data. Rosales and Sclaroff [18] described a method to learn the inverse mapping from silhouette to pose. Salzmann and Urtasun [13] proposed a method to impose physical constraints on the output of a discriminative predictor. Discriminative methods, in general, require large amounts of training data from varied viewpoints and deformations to be able to recover pose reliably and do not generalize well to data that is not represented by the training set.

Enforcing structural constraints optimally is usually intractable. In the context of deformable mesh reconstruction, Salzmann and Fua [19, 20] derived a convex formulation for constraining the solution space of possible 3D configurations by imposing convex inequality constraints on the relative distance between reconstructed points. Wei and Chai [21] and Valmadre and Lucey [22] describe deterministic algorithms to simultaneously estimate limb lengths and reconstruct human pose. These methods require multiple images and manual resolution of depth ambiguities at several joints.

In this paper, we present an automatic algorithm for recovering 3D body pose from 2D landmarks in a single image. To achieve this, we develop a statistical model of human pose variability that can describe a wide variety of actions, and an algorithm that simultaneously estimates 3D camera and body pose while enforcing anthropometric regularity.



(a) (b) (c)

Fig. 2. Data Complexity. (a) As more actions and, consequently, diverse poses are added to the training corpus, the maximum reconstruction error incurred by a linear dimensionality reduction model increases. (b) Maximum reconstruction error for each action separately using PCA. Each action can be compactly modeled with a linear basis. (c) Using a sparse representation in an overcomplete dictionary estimated using Orthogonal Matching Pursuit (OMP) achieves lower reconstruction error for 3D pose.

3 Sparse Representation of 3D Human Pose

A 3D configuration of P points can be represented by $\mathbf{X} = \left(\mathbf{X}_1^T, \ldots, \mathbf{X}_P^T\right)^T \in \mathbb{R}^{3P\times 1}$ of stacked 3D coordinates. Under weak perspective projection, the 2D coordinates of the points in the image are given by

$$\mathbf{x} = \left(\mathbf{I}_{P\times P} \otimes \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{R}\right) \mathbf{X} + \mathbf{t} \otimes \mathbf{1}_{P\times 1}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{2P\times 1}$, $\otimes$ denotes the Kronecker product, $\mathbf{s} \in \mathbb{R}^{2\times 2}$ is a diagonal scale matrix with s_x and s_y being the scales in the x and y directions , $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^{2\times 1}$ denote the rotation and translation parameters of the weak perspective camera that we collectively denote as $\mathcal{C}$. We assume the camera intrinsic parameters are known. Estimating $\mathbf{X}$ and $\mathcal{C}$ from only the image evidence $\mathbf{x}$ is, fundamentally, an ill-posed problem. We see from Equation 1 we have $3P + 7$ parameters that we need to estimate from only $2P$ equations.

If the points form a semantic group that deform in a structured way, such as anatomical landmarks on a human body, we can reduce the number of parameters that need to be estimated using dimensionality reduction methods that learn the correlations between the points [23]. Linear dimensionality reduction methods (e.g., Principal Component Analysis (PCA)) can be used to represent the points as a linear combination of a small number of basis poses,

$$\mathbf{X} = \boldsymbol{\mu} + \sum_{i=1}^{K} \mathbf{b}_i \omega_i, \tag{2}$$

where K is the number of basis poses, $\mathbf{b}_i$ are the basis poses, ω_i are the coefficients, and $\boldsymbol{\mu} \in \mathbb{R}^{3P\times 1}$ is the mean pose computed from training data. Under

this model, we now have to estimate only $K+7$ parameters instead of the original $3P+7$ parameters.

A direct application of PCA to all the poses contained in the corpus[2] raises difficulties as shown in Figure 2(a). For a single action, PCA performs well. As the diversity in actions in the data increases, the number of PCA components required for accurate reconstruction increases, and the assumption of a low dimensional linear subspace becomes strained. In particular, the *maximum* reconstruction error increases as the diversity in the data is increased because PCA inherits the occurrence statistics of poses in the corpus and not just the extent of variability.

3.1 Sparse Representation in an Overcomplete Dictionary

In Figure 2(b) we see that each individual action is compactly representable by a linear basis. Therefore, an arbitrary pose can be compactly represented by some subset of the set of all bases,

$$\begin{aligned} \mathbf{X} = \quad & \boldsymbol{\mu} + \textstyle\sum_{i=1}^{K} \mathbf{b}_i \omega_i, \\ & \{\mathbf{b}_i\}_{i \in I_{\mathbf{B}^*}} \in \mathbf{B}^* \subset \mathcal{B}, \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}$ is the mean pose, $\mathcal{B} \in \mathbb{R}^{3P \times (\sum_{i=1}^{N_a} N_b^i)}$ is an overcomplete dictionary of basis components created by concatenating N_b^i bases computed from N_a different actions, $\mathbf{B}^*$ is an optimal subset of $\mathcal{B}$, and $I_{\mathbf{B}^*}$ are the indices of the optimal basis $\mathbf{B}^*$ in $\mathcal{B}$. We validate this observation in Figure 2(c) by using Orthogonal Matching Pursuit (OMP) [24, 25] to select a sparse set of basis vectors to reconstruct each 3D pose in a test corpus. The sparse representation is able to achieve lower reconstruction error with higher compaction on the test set than using a full PCA model. It is instructive to note the behavior in Figure 2(c) of the maximum reconstruction error, which usually correspond to atypical poses. For human poses, we conclude that the sparse representation demonstrates greater generalization ability than full PCA.

3.2 Anthropometric Regularity

Linear models allow cases where the 2D projection appears to be valid (i.e., the reprojection error is minimized), but the configuration in 3D violates anthropometric quantities such as the proportions of limbs. Enforcing anthropometric regularity (i.e., that limb lengths follow known proportions) would discourage such implausible configurations. For a limb[3] between the ith and j^{th} landmark locations, we denote the normalized limb length as l_{ij}. The normalized limb

[2] We use the Carnegie Mellon Motion Capture Database [3] to obtain a large corpus of 3D human poses.

[3] We loosely define a limb to be a rigid length between two consecutive anatomical landmarks in the tree.

lengths are set by normalizing with respect to the longest limb of the mean pose ($\boldsymbol{\mu}$). For a 3D pose $\mathbf{X}$, we can ensure anthropometric regularity by enforcing

$$\begin{aligned} \|\mathbf{X}_i - \mathbf{X}_j\|_2 = l_{ij}, \\ \forall (i,j) \in \mathcal{L} \end{aligned} \tag{4}$$

where $\mathcal{L} = \{(i,j)\}_{i=1}^{N_l}$ is the set of pairs of joints between which a limb exists and N_l is the total number of limbs in the model. Unfortunately, applying quadratic equality constraints on a linear least squares system is nonconvex. A *necessary* condition for anthropometric regularity is

$$\sum_{\forall (i,j) \in \mathcal{L}} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 = \sum_{\forall (i,j) \in \mathcal{L}} l_{ij}^2. \tag{5}$$

This constraint limits the sum of the squared distances between valid landmarks to be equal to the sum of squares of the limb lengths[4]. The feasible set of the constraint in Equation 5 contains the feasible set of the constraints in Equation 4. The necessary condition on the sum of squared limb lengths is therefore a relaxation of the constraints in Equation 4. As shown in [8], this necessary condition can be applied in closed form.

4 Projected Matching Pursuit

We solve for the pose and camera by minimizing the reprojection error in the image. The resulting optimization problem can be stated as follows

$$\begin{aligned} \min_{\boldsymbol{\Omega}, \mathcal{C}, I_{\mathbf{B}^*}} \quad & \|\mathbf{x} - (\mathbf{I} \otimes \mathbf{sR})(\mathbf{B}^* \boldsymbol{\Omega} + \boldsymbol{\mu}) - \mathbf{t} \otimes \mathbf{1}\|_2 \\ \text{s.t.} \quad & \sum_{\forall (i,j) \in \mathcal{L}} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 = \sum_{\forall (i,j) \in \mathcal{L}} l_{ij}^2, \\ & \mathbf{B}^* \subset \mathcal{B}. \end{aligned} \tag{6}$$

Although the problem is non-linear, non-convex, and combinatorial, it has the following useful property in the set of arguments $(\mathcal{C}, \boldsymbol{\Omega}, I_{\mathbf{B}^*})$: we can solve optimally, or near-optimally, for each subset of the arguments given the rest. This property suggests a coordinate descent-style algorithm. Algorithm 1 describes a matching pursuit algorithm we refer to as *Projected Matching Pursuit* for coordinate descent on the reprojection error objective.

4.1 Algorithm

The combinatorial challenge of picking the optimal set of basis vectors from an overcomplete dictionary to represent a given signal is NP-hard. However, techniques exist to solve the sparse representation problem approximately with guarantees [25, 26]. Greedy approaches such as orthogonal matching pursuit (OMP)

[4] Note that since we are using normalized limb-lengths, these constraints become constraints on limb proportions rather than on limb lengths.

Algorithm 1 Projected Matching Pursuit

1. Initialize $\mathbf{r}_0 = \mathbf{x} - (\mathbf{I} \otimes \mathbf{sR})\,\boldsymbol{\mu} - \mathbf{t} \otimes \mathbf{1}$
2. While $(\|\mathbf{r}_t\| \geq tol)$
   3. $i_{\max} = \arg\max_{\mathrm{i}} \langle \mathbf{r}_t,\ (\mathbf{I} \otimes \mathbf{s}_t\mathbf{R}_t)\,\mathbf{B}_i \rangle$
   4. $\mathbf{B}^* = [\mathbf{B}^*\ \ \mathbf{B}_{i_{\max}}]$
   5. Solve: $\{\mathcal{C}^*, \boldsymbol{\Omega}^*\} = \arg\min \|\hat{\mathbf{x}} - (\mathbf{I} \otimes \mathbf{sR})\,\mathbf{B}^*\boldsymbol{\Omega}\|_2$
      subject to constraints in Equation (8) using Section 4.2 & Section 4.3
   6. Recompute residual $\mathbf{r}_{t+1} = \mathbf{x} - (\mathbf{I} \otimes \mathbf{s}^*\mathbf{R}^*)\,(\mathbf{B}^*\boldsymbol{\Omega}^* + \boldsymbol{\mu}) - \mathbf{t}^* \otimes \mathbf{1}$
   7. Set $\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Omega}^*$
8. Return $\{\mathcal{C}^*, \boldsymbol{\Omega}^*, \mathbf{B}^*\}$

[27, 25] reconstruct a signal $\mathbf{v}$ with a sparse linear combination of basis vectors from an overcomplete dictionary $\mathcal{B}$. It proceeds in a greedy fashion by choosing, at each iteration, the basis vector from $\mathcal{B}$ that is most aligned with the residual $\mathbf{r}$ (the residual is set equal to $\mathbf{v}$ in the first iteration). The new estimate of the signal $\hat{\mathbf{v}}$ is computed by reconstructing using the basis vectors selected at the current iteration and the new residual ($\mathbf{r} = \mathbf{v} - \hat{\mathbf{v}}$) is computed. The iterations proceed on the residual until K basis vectors are chosen or a tolerance on the residual error is reached.

In our scenario, we do not have access to the signal of interest, namely the 3D pose $\mathbf{X}$. Instead, we are only given the projection of the original 3D pose in the image $\mathbf{x}$. We present a matching pursuit algorithm for reconstructing a signal from its projection and an overcomplete dictionary. At each iteration of our algorithm, the optimal basis set $\mathbf{B}^*$ is augmented by matching the image residual with basis vectors projected under the current camera estimate and adding the basis vector which maximizes the inner product to the optimal set. Given the current optimal basis set $\mathbf{B}^*$, the pose and camera parameters are re-estimated as outlined in Section 4.2 and Section 4.3. The algorithm terminates when the optimal basis set has reached a predefined size or the image residual is smaller than a tolerance value. The procedure is summarized in Algorithm 1. We have an intuitive and feasible initialization in the mean 3D pose computed from the training corpus.

4.2 Estimating Basis Coefficients with Anthropometric Regularization

To encourage anthropometric regularity we enforce the necessary constraint from Equation 5 which limits the sum of squared limb lengths. We can write each 3D landmark $\mathbf{X}_i = \mathbf{E}_i\mathbf{X}$, where $\mathbf{E}_i = [\cdots\ \mathbf{0}\ \mathbf{I}_{3\times 3}\ \mathbf{0}\ \cdots]$ is a $3 \times 3P$ matrix that selects out the ith landmark.

We can write $\mathbf{E}_{ij} = \mathbf{E}_i - \mathbf{E}_j$, and express each limb length as $\|\mathbf{E}_{ij}\mathbf{X}\| = l_{ij}$. Equation 5 can now be rewritten in matrix form as:

$$\|\mathbf{C}\mathbf{X}\|_2^2 = \sum_{\forall(i,j)\in\mathcal{L}} l_{ij}^2, \tag{7}$$

where $\mathbf{C}$ is a $3N_l \times 3P$ matrix of the N_l stacked $\mathbf{E}_{ij}$ matrices. Where N_l is the number of limbs.

Given the optimal basis set $\mathbf{B}^*$ and the camera $\mathcal{C}$, solving for the coefficients of the linear model $\mathbf{\Omega}$ can now be formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{\Omega}} \quad & \|\hat{\mathbf{x}} - \mathbf{s}\mathbf{R} \otimes \mathbf{I}_{P\times P}\mathbf{B}^*\mathbf{\Omega}\|_2 \\ \text{s.t.} \quad & \|\mathbf{C}\mathbf{B}^*\mathbf{\Omega} - \mathbf{C}\boldsymbol{\mu}\|_2^2 = \sum_{\forall(i,j)\in\mathcal{L}} l_{ij}^2, \end{aligned} \tag{8}$$

where $\hat{\mathbf{x}} = \mathbf{x} - \mathbf{s}\mathbf{R} \otimes \mathbf{I}_{P\times P}\boldsymbol{\mu} - \mathbf{t} \otimes \mathbf{1}_{P\times 1}$. The above problem is a linear least squares problem with a single quadratic equality constraint that can be solved optimally in closed form as shown in [8].

There also exists a natural lower bound on the length of the limb between the estimated joint locations, $\mathbf{X}_i^*$ and $\mathbf{X}_j^*$, in terms of the image projections $\mathbf{x}_i$ and $\mathbf{x}_j$. Using the triangle inequality we can show that

$$\|\mathbf{X}_i^* - \mathbf{X}_j^*\| \geq \|\mathbf{s}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\|. \tag{9}$$

The above inequality shows that the estimated limb lengths are bounded by the length of the limbs in the image. Thus we can guarantee that the estimated limb length will not collapse to zeros as long as the limb has finite length in the image.

4.3 Estimating Camera Parameters

Given the pose $\mathbf{X} = \mathbf{B}^*\mathbf{\Omega} + \boldsymbol{\mu}$, and the image projections $\mathbf{x}$, we need to recover the weak perspective camera parameters $\mathcal{C}$. We solve this as an instance of the Orthogonal Procrustes problem [28]. We first write $\mathbf{x}$ and $\mathbf{X}$ in matrix form as $x \in \mathbb{R}^{2\times P}$ and $\mathcal{X} \in \mathbb{R}^{3\times P}$ respectively. We denote the mean-centered image projections as $\hat{x} = \mathbf{s}\mathbf{R}\mathcal{X}$. Using the singular value decomposition, we can write

$$\mathbf{M} = \hat{x}\mathcal{X}^T(\mathcal{X}\mathcal{X}^T)^{-1} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{10}$$

We obtain the scale $\mathbf{s}$ by taking the first 2×2 section of the matrix $\mathbf{D}$ and the rotation by setting $\mathbf{R} = \mathbf{U}\mathbf{V}^T$.

5 Evaluation

We perform quantitative and qualitative evaluation of our method. We use the Carnegie Mellon motion capture database for quantitative tests and compare our

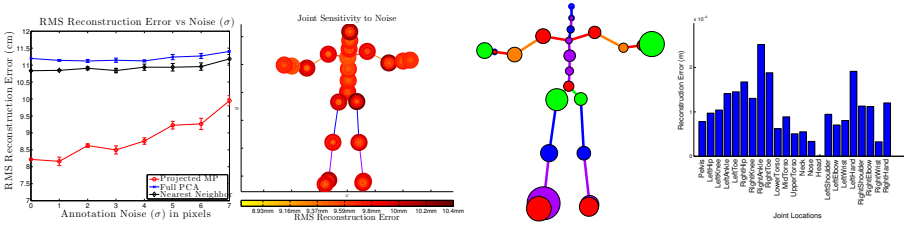


Fig. 3. Quantitative evaluation on optical motion capture. (a) We compare our method against two model baselines - a nearest neighbor approach and a linear model that uses PCA on the entire corpus. Reconstruction error is reported against annotation noise σ on a test corpus. (b) We evaluate the sensitivity of the reconstruction to each anatomical landmark annotation. (c) We show the sensitivity in reconstruction to missing landmarks. The radius of each circle indicates the relative magnitude of error in 3D incurred when the landmark is missing (d) The additional reconstruction incurred when the landmark is missing.

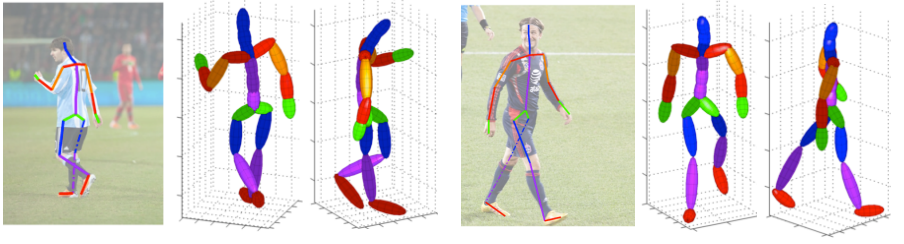


Fig. 4. Our method is able to handle missing data. We show examples of reconstruction with missing annotations. The missing limbs are marked with dotted lines. We are able to reconstruct the pose and impute the missing landmarks in 3D.

results against using a representation baseline (direct PCA on the entire corpus) and a non-parametric nearest neighbor method.

For all experiments, an overcomplete shape basis dictionary was constructed by concatenating the shape bases learnt for a set of human actions. We use a model with 23 anatomical landmarks. Each pose in the motion capture corpus was aligned by procrustes analysis to a reference pose. Shape bases were then learnt for the following motion categories- *'running', 'waving', 'arm movement', 'walking', 'jumping', 'jumping jacks', 'run', 'sit', 'boxing','bend'* by collecting sequences from the CMU Motion Capture Dataset and concatenating PCA components which retained 99% of the energy from each motion category.

5.1 Quantitative Evaluation

Optical Motion Capture. To evaluate our methods we test our algorithm on a sequence of mixed activities from the CMU motion capture database. We take care to ensure that the motion capture frames come from sequences that were not used in the training of the shape bases. We project 30 frames of motion capture of diverse poses into 4 synthetically generated camera views. We then run our algorithm on the 2D projections of the joint locations to obtain the camera location and the pose of the human. We report 3D joint position error with increasing annotation noise σ in Figure 3(a).

We compare our method against two baselines. The first baseline uses as a linear model, a basis computed by performing PCA on the entire training corpus. Anthropometric constraints are enforced as in Section 4.2. The second baseline uses a non-parametric, nearest neighbor approach that retains all the training data. The 2D projections in each test example are matched to every 3D pose in the corpus by estimating the best-fit camera using the method in Section 4.3. The 3D pose that has the least reprojection error under the best-fit camera estimate is returned. The results are reported in Figure 3. We find that our method that used Projected Matching Pursuit achieves the lowest RMS reconstruction error. We also tested the effect of imposing an equality constraint on the sum-of-squared limb length ratios and find that we deviate from the ground truth on our test set by 13.1% on average.

We evaluate the comparative importance of the anatomical landmarks by performing two experiments:

Joint Sensitivity. We test the sensitivity of the reconstruction to each landmark individually. Each pose in the testing corpus is projected into 2D with synthetically generated cameras and each landmark is perturbed with Gaussian noise independently. Figure 3(b) shows the sensitivity of the reconstruction to each landmark. The maximum length of a limb in the image is 200 pixels, the minimum limb length is 20 pixels, and the average length of a limb in the image is 94.5. pixels The noise is varied to about 10% of the average limb length in the image.

Missing Data. An advantage of our formulation is the ability to handle missing data. In Figure 5 we show examples of reconstructions obtained with incomplete annotations. We perform an ablative analysis of the joint annotations by removing each annotation in turn and measure the increase in the reconstruction error. We plot our results in Figure 3(d). The radius of each circle is indicative of the error incurred when the annotation corresponding to that joint is missing. We find that the extremal joints are most informative and help in constraining the reconstruction.

5.2 Qualitative Evaluation

Comparison with recent work. We compare reconstructions obtained by our method to recent work by Valmadre et al. [22]. Their method requires multiple images of the same person and requires a human annotator to resolve depth

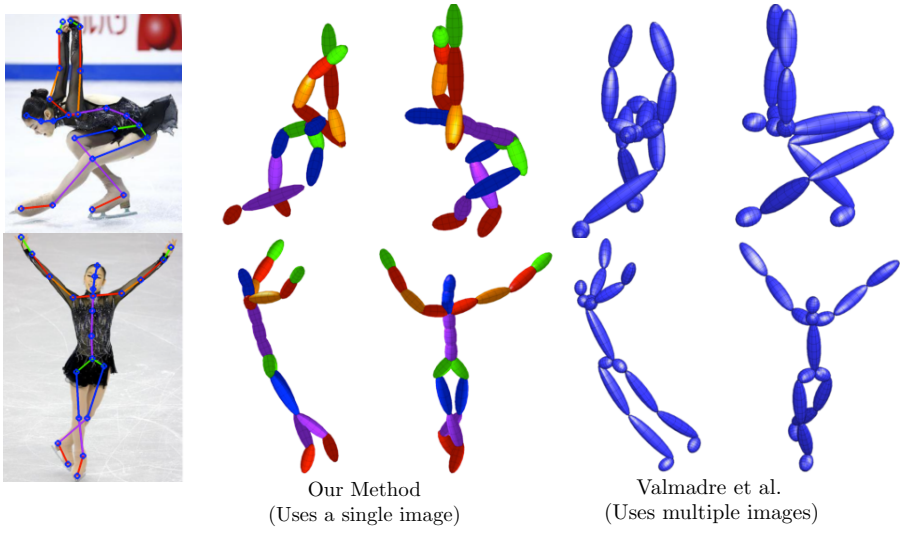


Fig. 5. Comparsion with recent work. Valmadre et al., estimate human pose using multiple images and requires additional annotation to resolve ambiguities. Our method achieves realistic results operating on a single image and does not require additional annotation

ambiguities. We present our comparative results in Figure 5. Our method is applied per frame to images of the ice skater Yu-Na Kim and compared to the reconstructions obtained by Valmadre et al. We can see in Figure 5 that we are able to obtain good reconstructions per image, without the requirement of a human annotator resolving the depth ambiguities.

Internet Images. We downloaded images of people in a variety of poses from the internet. The 2D joint locations were manually annotated. We present the results in Figures 7(a) and 6. In Figure 6 we first obtained individual camera and pose estimates for each of the annotated human figures. We then fixed the camera upright at an arbitrary location and placed the human figures using the estimated relative rigid pose. It can be seen that the camera estimates are consistent as the actors are placed in their correct locations.

Non-standard viewpoints. We also test our method on images taken from non-standard viewpoints. We reconstruct the pose and relative camera from photographs downloaded from the internet taken from viewpoints that have generally been considered difficult for pose estimation algorithms. We are able to recover the pose and the viewpoint of the algorithm for such examples as shown in Figure 7(b).

Monocular video. We demonstrate our algorithm on a set of key frames extracted from monocular video in Figure 7(c). The relative camera estimates are aligned to a single view-point to obtain a sequence of the person performing an action. Note that we are able to estimate the relative pose between the camera and the human correctly resulting in a realistic reconstruction of the sequence.

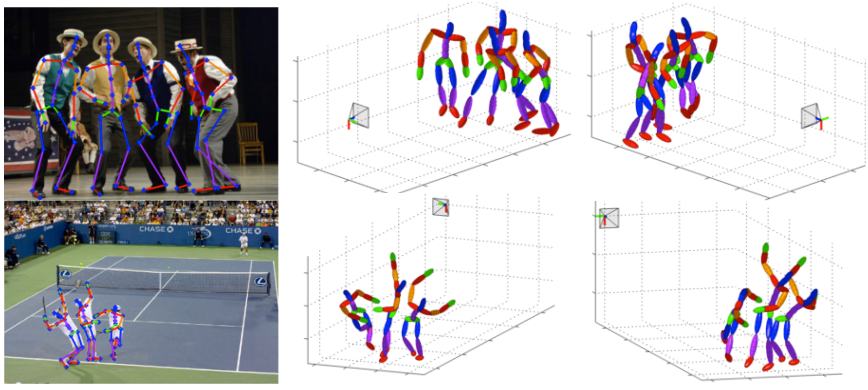
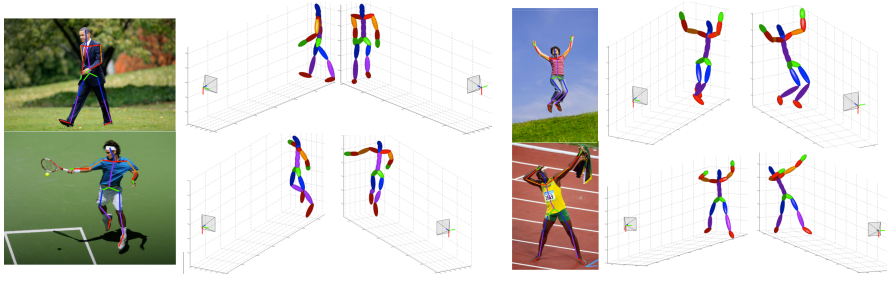
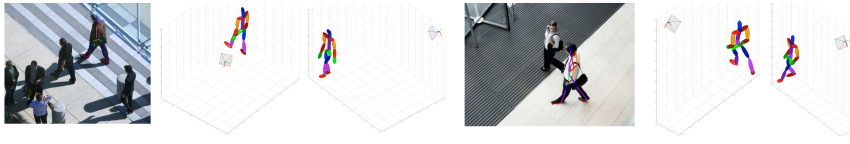


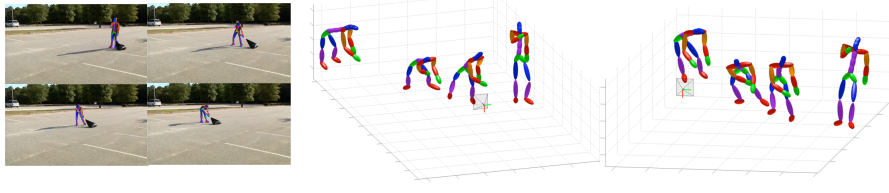
Fig. 6. Reconstruction with multiple people in the same view. The camera estimation is accurate as the people are placed consistently.



(a) Reconstruction of people in arbitrary poses from internet images.



(b) Reconstruction of people viewed from varied viewpoints.



(c) Our algorithm applied to four frames of an annotated video.

Fig. 7. We acheive realistic reconstructions for people in (a) arbitrary poses, (b) captured from varied viewpoints and (c) monocular video streams.

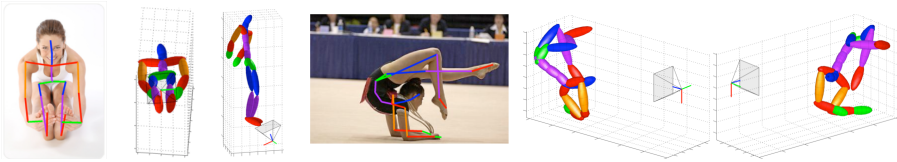


Fig. 8. Failure Cases. The method does not recover the correct pose when there are strong perspective effects and if the mean pose is not a good initialization.

6 Discussion

We presented a new representation for human pose as a sparse linear embedding in an overcomplete dictionary. We develop a matching pursuit algorithm for estimating the sparse representation of 3D pose and the relative camera from only 2D image evidence while simultaneously maintaining anthropometric regularity. Every step in the matching pursuit iterations is computed in closed form, therefore the algorithm is efficient and takes on average 5 seconds per image to converge. We are able to achieve good generalization to a large range of poses and viewpoints. A case where the algorithm does not result in good reconstructions are in images with strong perspective effects where the weak perspective assumptions on the camera model are violated and in poses where the mean pose is not a reasonable initialization (See Figure 8).

Acknowledgements:

This research was funded (in part) by the Intel Science and Technology Center on Embedded Computing, NSF CRI-0855163, and DARPA's Mind's Eye Program. We also thank Daniel Huber and Tomas Simon for providing valuable feedback on the manuscript.

References

1. Lee, H.J., Chen, Z.: Determination of 3D Human Body Postures from a Single View. Computer Vision, Graphics, and Image Processing **30** (1985) 148 – 168
2. Peelen, M.V., Downing, P.E.: The Neural Basis of Visual Body Perception. Nature Reviews Neuroscience (**8**) 636–648
3. MoCap: Carnegie Mellon University Graphics Lab Motion Capture Database. (`http://mocap.cs.cmu.edu`)
4. Matthews, I., Baker, S.: Active Appearance Models Revisited. International Journal of Computer Vision **60** (2003) 135–164
5. Safonova, A., Hodgins, J.K., Pollard, N.S.: Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. ACM Transactions on Graphics (SIGGRAPH 2004) **23** (2004)
6. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-Time Combined 2D+3D Active Appearance Models. In: CVPR, IEEE. (2004) 535 – 542

7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
8. Gander, W.: Least Squares with a Quadratic Constraint. Numerische Mathematik (1981)
9. Taylor, C.: Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. CVIU (2000) 349–363
10. Jiang, H.: 3D Human Pose Reconstruction Using Millions of Exemplars. In: ICPR, IEEE. (2010) 1674 –1677
11. Parameswaran, V., Chellappa, R.: View Independent Human Body Pose Estimation from a Single Perspective Image. CVPR, IEEE (2006) 16–22
12. Barron, C., Kakadiaris, I.A.: Estimating Anthropometry and Pose from a Single Uncalibrated Image. CVIU (2001) 269–284
13. Salzmann, M., Urtasun, R.: Implicitly Constrained Gaussian Process Regression for Monocular Non-Rigid Pose Estimation. In: Advances in Neural Information Processing Systems. (2010) 2065–2073
14. Agarwal, A., Triggs, B.: 3D Human Pose from Silhouettes by Relevance Vector Regression. In: CVPR, IEEE. (2004) 882–888
15. Mori, G., Malik, J.: Recovering 3D Human Body Configurations using Shape Contexts. PAMI, IEEE **28** (2006) 1052–62
16. Shakhnarovich, G., Viola, P., Darrell, T.: Fast Pose Estimation with Parameter-Sensitive Hashing. ICCV, IEEE (2003) 750
17. Elgammal, A., Lee, C.S.: Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In: CVPR, IEEE. (2004) 681–688
18. Rosales, R., Sclaroff, S.: Specialized Mappings and the Estimation of Human Body Pose from a Single Image. In: Proceedings of the Workshop on Human Motion, 2000. (2000) 19 –24
19. Salzmann, M., Fua, P.: Reconstructing Sharply Folding Surfaces: A Convex Formulation. CVPR, IEEE (2009) 1054–1061
20. Moreno-Noguer, F., Porta, J., Fua, P.: Exploring Ambiguities for Monocular Non-rigid Shape Estimation. In: ECCV. Volume 6313 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (2010) 370–383
21. Wei, X.K., Chai, J.: Modeling 3D Human Poses from Uncalibrated Monocular Images. In: ICCV, IEEE. (2009) 1873–1880
22. Valmadre, J., Lucey, S.: Deterministic 3D Human Pose Estimation using Rigid Structure. In: ECCV, Berlin, Heidelberg, Springer-Verlag (2010) 467–480
23. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 681 –685
24. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In: 1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers. (1993) 40 –44 vol.1
25. Tropp, J.A., Gilbert, A.C.: Signal Recovery from Random Measurements via Orthogonal Matching Pursuit. IEEE Transactions on Information Theory **53** (2007) 4655–4666
26. Tropp, J.: Greed is Good: Algorithmic Results for Sparse Approximation. IEEE Transactions on Information Theory **50** (2004) 2231 – 2242
27. Mallat, S., Zhang, Z.: Matching Pursuits with Time-Frequency Dictionaries. IEEE Transactions on Signal Processing **41** (1993) 3397 –3415
28. Schnemann, P.: A Generalized Solution of the Orthogonal Procrustes Problem. Psychometrika **31** (1966) 1–10 10.1007/BF02289451.