# 3D Shape Estimation from 2D Landmarks: A Convex Relaxation Approach

Xiaowei Zhou[†], Spyridon Leonardos[†], Xiaoyan Hu[‡†], Kostas Daniilidis[†]
[†] University of Pennsylvania [‡] Beijing Normal University
{xiaowz,spyridon,kostas}@cis.upenn.edu huxy@bnu.edu.cn

## Abstract

*We investigate the problem of estimating the 3D shape of an object, given a set of 2D landmarks in a single image. To alleviate the reconstruction ambiguity, a widely-used approach is to confine the unknown 3D shape within a shape space built upon existing shapes. While this approach has proven to be successful in various applications, a challenging issue remains, i.e., the joint estimation of shape parameters and camera-pose parameters requires to solve a non-convex optimization problem. The existing methods often adopt an alternating minimization scheme to locally update the parameters, and consequently the solution is sensitive to initialization. In this paper, we propose a convex formulation to address this problem and develop an efficient algorithm to solve the proposed convex program. We demonstrate the exact recovery property of the proposed method, its merits compared to alternative methods, and the applicability in human pose and car shape estimation.*

## 1. Introduction

Recognizing 3D objects from 2D images is a central problem in computer vision. In recent years, there has been an emerging trend towards analyzing 3D geometry of objects including shapes and poses instead of merely providing bounding boxes [37, 25, 4, 28, 36, 33]. The 3D geometric reasoning can not only provide richer information about the scene for subsequent high-level tasks, but also improve the performance of object detection. [20, 31, 3, 34].

Estimating the 3D geometry of an object from a single view is an ill-posed problem. But it is a possible task for a human observer, since human can leverage visual memory of object shapes. Inspired by this idea, more and more efforts have been made towards 3D model-based analysis leveraging the increasing availability of online 3D model databases. To address intra-class variability or nonrigid deformation, many recent works, e.g., [22],[32],[45],[35], have adopted a shape-space approach originated from the "active shape model" [14], where each shape is defined by a set of ordered landmarks and the shape to be estimated is assumed to be a linear combination of predefined basis shapes. For estimation, the 3D shape model is fitted to the landmarks annotated or detected in images. In this way, the problem turns into a 3D-to-2D shape fitting problem, where the shape parameters (weights of the linear combination) and the pose parameters (viewpoint) have to be estimated simultaneously.

While this approach has achieved promising results in various applications, the model inference is still a challenging problem, since the subproblems of shape and pose estimation are coupled: the pose needs to be known to deform the 3D model to match 2D landmarks, while the exact 3D model is required to estimate the pose. The joint estimation of shape and pose parameters usually results in a nonconvex optimization problem, and the orthogonality constraint on the pose parameters makes the problem more complicated. The previous methods often adopted an alternating scheme to alternately update the shape and pose parameters until convergence. Therefore, the algorithms were sensitive to initialization and may get stuck at locally-optimal solutions. As mentioned in many works, e.g., [32],[22], most of the failed cases were attributed to bad initialization. Some heuristics have been used to relieve this issue, such as trying multiple initializations [35] or using a viewpoint-aware detector for pose initialization [45]. But there is still no guarantee for global optimality.

In this paper, we propose a convex relaxation approach to addressing the aforementioned issue:

1. We use an augmented shape-space model, where a shape is represented as a linear combination of rotatable basis shapes. This model can give a linear representation of both intrinsic shape deformation and extrinsic viewpoint changes.

2. We use the convex relaxation of the orthogonality constraint to convert the entire problem into a spectral-norm regularized linear inverse problem, which is a convex program.

3. We develop an efficient algorithm to globally solve the proposed convex program.

The remainder of this paper is organized as follows. We first give a brief introduction to related work in Section 2. Then, we explain the formulation in Section 3 and provide the algorithm in Section 4. Next, we experimentally demonstrate the merits and applicability of the proposed method in Section 5. Finally, we conclude the paper with some discussions in Section 6.

## 2. Related Work

The most related work includes the papers that solve shape estimation by fitting a shape-space model to 2D landmarks. This approach has been successfully applied to reconstruction of a variety of objects including human poses [32, 35, 18, 42], cars [24, 22, 45, 26], faces [6, 21, 12], to name a few. Following are a few recent examples.

Ramakrishna et al. [32] proposed a sparse representation based approach to reconstructing 3D human pose from annotated landmarks in a still image. Wang et al. [35] adopted a 2D human pose detector [40] to automatically locate the joints and used a robust estimator to handle inaccurate joint locations. Fan et al. [18] proposed to improve the performance of [32] by enforcing locality when building the pose dictionary. Hejrati et al. [22] used the active shape model for 3D car reconstruction and produced 2D landmarks by a variant of deformable part models [19]. Lin et al. [26] proposed a method for joint 3D model fitting and fine-grained classification for cars. In some works, landmark locations were estimated jointly with shape fitting. For example, Zia [45] et al. developed a probabilistic framework to simultaneously localize 2D landmarks and recovery 3D object models. Zhou et al. [42] formulated human pose estimation as a matching problem, where the learned spatio-temporal pose model was matched to extracted trajectories in a video.

A common component or an intermediate step in these works is the 3D model fitting to 2D landmarks. As mentioned in the introduction, the previous work usually relied on nonconvex formulations, which may be sensitive to initialization. The convex formulation proposed in this paper can potentially serve as a building block to improve the performance of the existing methods.

Our work is also closely related to nonrigid structure from motion (NRSfM), where a deformable shape is recovered from multi-frame 2D-2D correspondences. The low-rank shape-space model has been frequently used in NRSfM, but the basis shapes are unknown. The joint estimation of shape/pose variables and basis shapes is typically solved via matrix factorization followed by metric rectification [9, 39]. In some recent works, iterative algorithms were employed for better precision [29, 15] or sequential processing [2], and the problem studied in this paper is analogous to the step of fixing basis shapes and updating the remaining variables in those iterative methods for NRSfM.

## 3. Formulation

### 3.1. Problem Statement

The problem studied in this paper can be described by the following linear system:

$$\boldsymbol{W} = \Pi \boldsymbol{S}, \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{3\times p}$ denotes the unknown 3D shape, which is represented by 3D locations of $p$ points. $\boldsymbol{W} \in \mathbb{R}^{2\times p}$ denotes their projections in a 2D image. $\Pi$ is the camera calibration matrix. To simplify the problem, the weak-perspective camera model is usually used, which is a good approximation when the object depth is much smaller than the distance from the camera. With this assumption, the calibration matrix has the following simple form:

$$\Pi = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \end{pmatrix}, \tag{2}$$

where $\alpha$ is a scalar depending on the focal length and the distance to the object.

There are always more variables than equations in (1). To make the problem well-posed, a widely-used assumption is that the unknown shape can be represented as a linear combination of predefined basis shapes, which is originated from the active shape model [14]:

$$\boldsymbol{S} = \sum_{i=1}^{k} c_i \boldsymbol{B}_i, \tag{3}$$

where $\boldsymbol{B}_i \in \mathbb{R}^{3\times p}$ for $i \in [1, k]$ represents a basis shape learned from training data, while $c_i$ denotes the weight of each basis shape. In this way, the reconstruction problem is turned into a problem of estimating several coefficients by fitting the model (3) to the landmarks in an image, which greatly reduces the number of unknowns.

Since the basis shapes are predefined, the relative rotation and translation between the camera frame and the frame defining the basis shapes need to be taken into account, and the 3D-2D projection is depicted by:

$$\boldsymbol{W} = \Pi \left( \boldsymbol{R} \sum_{i=1}^{k} c_i \boldsymbol{B}_i + \boldsymbol{T}\boldsymbol{1}^T \right), \tag{4}$$

where $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{T} \in \mathbb{R}^3$ correspond to the rotation matrix and the translation vector, respectively. $\boldsymbol{R}$ should be in the special orthogonal group

$$SO(3) = \{\boldsymbol{R} \in \mathbb{R}^{3\times 3} | \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_3, \det \boldsymbol{R} = 1\}. \tag{5}$$

Equation (4) can be further simplified as

$$\boldsymbol{W} = \bar{\boldsymbol{R}} \sum_{i=1}^{k} c_i \boldsymbol{B}_i, \tag{6}$$

where $\bar{\boldsymbol{R}} \in \mathbb{R}^{2\times 3}$ denotes the first two rows of the rotation matrix, and the translation $\boldsymbol{T}$ has been eliminated by centralizing the data, i.e. subtracting each row of $\boldsymbol{W}$ and $\boldsymbol{B}$ by its mean. Note that the scalar $\alpha$ in the calibration matrix has been absorbed into $c_1, \cdots, c_k$.

In the active shape model, the number of basis shapes is set to be small, which assumes that the unknown shape lies in a low-dimensional linear space. In many recent works [32, 41, 43, 44], it has been shown that the low-dimensional linear space is insufficient to model complex shape variation, e.g., human poses, and a promising approach is using an over-complete dictionary and representing an unknown shape as a sparse combination of atoms in the dictionary. Such a sparse representation implicitly encodes the assumption that the unknown shape should lie in a union of subspaces that approximates a nonlinear shape manifold.

Based on the sparse representation of shapes, the following optimization problem is often considered to estimate an unknown shape:

$$\begin{aligned} \min_{\boldsymbol{c},\bar{\boldsymbol{R}}} \ & \frac{1}{2}\left\|\boldsymbol{W} - \bar{\boldsymbol{R}}\sum_{i=1}^{k} c_i \boldsymbol{B}_i\right\|_F^2 + \lambda\|\boldsymbol{c}\|_1, \\ \text{s.t.} \ & \bar{\boldsymbol{R}}\bar{\boldsymbol{R}}^T = \boldsymbol{I}_2, \end{aligned} \tag{7}$$

where $\boldsymbol{c} = [c_1, \cdots, c_k]^T$ and $\|\boldsymbol{c}\|_1$ represents the $\ell_1$ norm of $\boldsymbol{c}$, which is the convex surrogate of the cardinality. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The cost function terms in (7) correspond to the reprojection error and the sparsity of representation, respectively.

The optimization in (7) is nonconvex and there is an orthogonality constraint. A commonly-used strategy is the alternating minimization scheme, in which two steps are alternated: fixing $\bar{\boldsymbol{R}}$ and updating $\boldsymbol{c}$ by solving the $\ell_1$ minimization problem; and fixing $\boldsymbol{c}$ and updating $\bar{\boldsymbol{R}}$ using certain rotation representations such as the quaternions, the exponential map or a manifold representation. Note that the Procrustes method cannot be directly applied here since $\bar{\boldsymbol{R}} \in \mathbb{R}^{2\times 3}$ is not a full rotation matrix and generally no closed-form solution exists [17]. Consequently, the whole algorithm may get stuck at local minima far away from the globally-optimal solution.

### 3.2. Proposed Model

We propose to use the following shape-space model:

$$\boldsymbol{S} = \sum_{i=1}^{k} c_i \boldsymbol{R}_i \boldsymbol{B}_i, \tag{8}$$

in which there is a rotation for each basis shape. The model in (8) implicitly accounts for the viewpoint variability and the projected 2D model is

$$\boldsymbol{W} = \Pi\sum_{i=1}^{k} c_i \boldsymbol{R}_i \boldsymbol{B}_i = \sum_{i=1}^{k} \boldsymbol{M}_i \boldsymbol{B}_i, \tag{9}$$

where $\boldsymbol{M}_i \in \mathbb{R}^{2\times 3}$ is the product of $c_i$ and the first two rows of $\boldsymbol{R}_i$, which satisfies

$$\boldsymbol{M}_i \boldsymbol{M}_i^T = c_i^2 \boldsymbol{I}_2. \tag{10}$$

The motivation of using the models in (8) and (9) is to achieve a linear representation of shape variability in 2D, such that we can get rid of the bilinear form in (6), which is a necessary step towards a convex formulation.

The model in (9) is equivalent to the affine-shape model in existing literature [5, 38], which uses an augmented linear space to represent the shape variation in 2D caused by both intrinsic shape deformation and extrinsic viewpoint changes. This representation also appears in most NRSfM literature [9, 29]. As mentioned in [38], the augmented linear space can represent any 2D shape produced by the 3D shape model projected into the image plane, but the increase of degree of freedom may result in invalid shapes. In this work, we try to reduce the possibility of invalid cases by enforcing the orthogonality constraint on $\boldsymbol{M}_i$s and the sparsity constraint on the number of activated basis shapes. We will show that these constraints can be conveniently imposed by minimizing a convex regularizer.

Next, we will consider to replace the orthogonality constraint in (10) by its convex counterpart. The following lemma has been proven in literature [23, Section 3.4]:

**Lemma 1.** *The convex hull of the Stiefel manifold* $\mathcal{Q} = \left\{\boldsymbol{X} \in \mathbb{R}^{m\times n} | \ \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_n\right\}$ *equals the unit spectral-norm ball* $conv(\mathcal{Q}) = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} | \ \|\boldsymbol{X}\|_2 \leq 1\}$. $\|\boldsymbol{X}\|_2$ *denotes the spectral norm (a.k.a. the induced 2-norm) of a matrix* $\boldsymbol{X}$*, which is defined as the largest singular value of* $\boldsymbol{X}$.

Based on Lemma 1, we have the following proposition:

**Proposition 1.** *Given a scalar* $s$*, the convex hull of* $\mathcal{S} = \left\{\boldsymbol{Y} \in \mathbb{R}^{m\times n} | \ \boldsymbol{Y}^T\boldsymbol{Y} = s^2\boldsymbol{I}_n\right\}$ *equals the spectral-norm ball with a radius of* $|s|$*:* $conv(\mathcal{S}) = \{\boldsymbol{Y} \in \mathbb{R}^{m\times n} | \ \|\boldsymbol{Y}\|_2 \leq |s|\}$.

The proof is straightforward since there is a linear mapping between $\mathcal{S}$ and $\mathcal{Q}$ by $\boldsymbol{Y} = s\boldsymbol{X}$.

Consequently, the tightest convex relaxation to the constraint in (10) is given by $\|\boldsymbol{M}_i\|_2 \leq |c_i|$.

Finally, with the modified shape model, the relaxed orthogonality constraint and the assumption of sparse repre-

sentation, we propose to minimize the $\ell_1$-norm of the coefficient vector for shape recovery under noiseless cases:

$$\begin{aligned} \min_{c_1,\cdots,c_k,\boldsymbol{M}_1,\cdots,\boldsymbol{M}_k} \quad & \sum_{i=1}^{k} |c_i|, \\ \text{s.t.} \quad & \boldsymbol{W} = \sum_{i=1}^{k} \boldsymbol{M}_i \boldsymbol{B}_i, \\ & \|\boldsymbol{M}_i\|_2 \leq |c_i|,\ \forall i \in [1,k] \end{aligned} \tag{11}$$

which is obviously equivalent to the following problem:

$$\begin{aligned} \min_{\boldsymbol{M}_1,\cdots,\boldsymbol{M}_k} \quad & \sum_{i=1}^{k} \|\boldsymbol{M}_i\|_2, \\ \text{s.t.} \quad & \boldsymbol{W} = \sum_{i=1}^{k} \boldsymbol{M}_i \boldsymbol{B}_i. \end{aligned} \tag{12}$$

The formulation in (12) is a linear inverse problem, where we estimate a set of orthogonal matrices by minimizing their spectral norms. Interestingly, the conditions for exact recovery using such a convex program has been theoretically analyzed in [13]. We will provide numerical results to demonstrate the exact recovery property in Section 5.1.

Considering noise in real applications, we can solve:

$$\min_{\boldsymbol{M}_1,\cdots,\boldsymbol{M}_k} \frac{1}{2} \left\| \boldsymbol{W} - \sum_{i=1}^{k} \boldsymbol{M}_i \boldsymbol{B}_i \right\|_F^2 + \lambda \sum_{i=1}^{k} \|\boldsymbol{M}_i\|_2. \tag{13}$$

The problem (13) is our final formulation. It is a penalized least-squares problem. We have following remarks:

1. The problem in (13) is convex programming, which can be solved globally. We will provide an efficient algorithm to solve it in Section 4.

2. Notice that $\|\cdot\|_2$ in the above formulations denotes the spectral norm of a matrix instead of the $\ell_2$-norm of a vector. As we will show in Section 4, minimizing the spectral norm of a matrix is equivalent to minimizing the $\ell_\infty$-norm of the vector of singular values, which will simultaneously shrink the norm of the matrix towards zero and enforce its singular values to be equal. Therefore, by spectral-norm minimization, we can not only minimize the number of activated basis shapes but also enforce each transformation matrix $\boldsymbol{M}_i$ to be orthogonal (an orthogonal matrix has equal singular values).

3. In practice, we may estimate $\boldsymbol{M}_i$s by only considering reprojection errors at visible landmarks, i.e., including a binary weight matrix in the first term of (13). The missing landmarks can be hallucinated from the reconstructed shape model as their locations are known on the basis shapes.

### 3.3. Reconstruction

After solving (13), we recover $c_i$ and $\boldsymbol{R}_i$ from the estimated $\boldsymbol{M}_i$, and reconstruct the 3D shape by (8). Specifically, $c_i = \|\boldsymbol{M}_i\|_2$ and $\bar{\boldsymbol{R}}_i = \boldsymbol{M}_i / c_i$. Note that $c_i = -\|\boldsymbol{M}_i\|_2$ is also a feasible solution. To eliminate the ambiguity, we assume that $c_i \geq 0$ and impose this constraint when training the shape dictionary. Finally, the third row of $\boldsymbol{R}_i$ is recovered by the cross product of the rows in $\bar{\boldsymbol{R}}_i$.

## 4. Optimization

### 4.1. Proximal operator of the spectral norm

Before deriving the specific algorithm to solve (13), we first prove the following proposition, which will serve as an important building block in our algorithm.

**Proposition 2.** *The solution to the following problem*

$$\min_{\boldsymbol{X}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_2 \tag{14}$$

*is given by* $\boldsymbol{X}^* = \mathcal{D}_\lambda(\boldsymbol{Y})$*, where*

$$\mathcal{D}_\lambda(\boldsymbol{Y}) = \boldsymbol{U}_Y \, diag\left[\boldsymbol{\sigma}_Y - \lambda \mathcal{P}_{\ell_1}(\boldsymbol{\sigma}_Y / \lambda)\right] \, \boldsymbol{V}_Y^T, \tag{15}$$

$\boldsymbol{U}_Y$*,* $\boldsymbol{V}_Y$ *and* $\boldsymbol{\sigma}_Y$ *denote the left singular vectors, right singular vectors and the singular values of* $\boldsymbol{Y}$*, respectively.* $\mathcal{P}_{\ell_1}$ *is the projection of a vector to the unit* $\ell_1$*-norm ball.*

*Proof.* The problem in (14) is a proximal problem [30]. The proximal problem associated with a function $F$ is defined as

$$\text{prox}_{\lambda F}(\boldsymbol{Y}) = \arg\min_{\boldsymbol{X}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \lambda F(\boldsymbol{X}), \tag{16}$$

with the solution denoted by $\text{prox}_{\lambda F}(\boldsymbol{Y})$ and named the proximal operator of $F$.

For the problem (14), $F(\boldsymbol{X}) = \|\boldsymbol{X}\|_2 = \|\boldsymbol{\sigma}_X\|_\infty$, where $\|\cdot\|_\infty$ means the $\ell_\infty$ norm. It says that $F$ is a spectral function operated on singular values of a matrix. Based on the property of spectral functions [30, Section 6.7.2], we have

$$\text{prox}_{\lambda F}(\boldsymbol{Y}) = \boldsymbol{U}_Y \, \text{diag}\left[\text{prox}_{\lambda f}(\boldsymbol{\sigma}_Y)\right] \, \boldsymbol{V}_Y^T, \tag{17}$$

where $f$ is the $\ell_\infty$-norm. The proximal operator of the $\ell_\infty$-norm can be computed by Moreau decomposition [30, Section 6.5]:

$$\text{prox}_{\lambda f}(\boldsymbol{\sigma}_Y) = \boldsymbol{\sigma}_Y - \lambda \mathcal{P}_{\ell_1}(\boldsymbol{\sigma}_Y / \lambda), \tag{18}$$

given that the $\ell_1$-norm is the dual norm of the $\ell_\infty$-norm. □

### 4.2. Algorithms

We present the algorithm to solve (13). The noiseless case (12) can be solved similarly. Our algorithm is based on the Alternating Direction Method of Multipliers (ADMM) [8] and the proximal operator of the spectral norm.

We first introduce an auxiliary variable $\boldsymbol{Z}$ and rewrite (13) as follows

$$\begin{aligned} \min_{\widetilde{\boldsymbol{M}},\widetilde{\boldsymbol{Z}}} \ & \frac{1}{2}\left\|\boldsymbol{W}-\boldsymbol{Z}\widetilde{\boldsymbol{B}}\right\|_F^2+\lambda\sum_{i=1}^{k}\|\boldsymbol{M}_i\|_2, \\ \text{s.t.} \ & \widetilde{\boldsymbol{M}}=\boldsymbol{Z}, \end{aligned} \tag{19}$$

where we concatenate $\boldsymbol{M}_1,\cdots,\boldsymbol{M}_k$ as column-triplets of $\widetilde{\boldsymbol{M}}$ and $\boldsymbol{B}_1,\cdots,\boldsymbol{B}_k$ as row-triplets of $\widetilde{\boldsymbol{B}}$.

The augmented Lagrangian of (19) is

$$\begin{aligned} \mathcal{L}_\mu\left(\widetilde{\boldsymbol{M}},\boldsymbol{Z},\boldsymbol{Y}\right)=&\frac{1}{2}\left\|\boldsymbol{W}-\boldsymbol{Z}\widetilde{\boldsymbol{B}}\right\|_F^2+\lambda\sum_{i=1}^{k}\|\boldsymbol{M}_i\|_2 \\ &+\left\langle\boldsymbol{Y},\widetilde{\boldsymbol{M}}-\boldsymbol{Z}\right\rangle+\frac{\mu}{2}\left\|\widetilde{\boldsymbol{M}}-\boldsymbol{Z}\right\|_F^2, \end{aligned} \tag{20}$$

where $\boldsymbol{Y}$ is the dual variable and $\mu$ is a parameter controlling the step size in optimization. Then, the ADMM alternates the following steps until convergence:

$$\widetilde{\boldsymbol{M}}^{t+1}=\arg\min_{\widetilde{\boldsymbol{M}}}\mathcal{L}_\mu\left(\widetilde{\boldsymbol{M}},\boldsymbol{Z}^t,\boldsymbol{Y}^t\right); \tag{21}$$

$$\boldsymbol{Z}^{t+1}=\arg\min_{\boldsymbol{Z}}\mathcal{L}_\mu\left(\widetilde{\boldsymbol{M}}^{t+1},\boldsymbol{Z},\boldsymbol{Y}^t\right); \tag{22}$$

$$\boldsymbol{Y}^{t+1}=\boldsymbol{Y}^k+\mu\left(\widetilde{\boldsymbol{M}}^{t+1}-\boldsymbol{Z}^{t+1}\right). \tag{23}$$

For the step in (21), we have

$$\begin{aligned} &\min_{\widetilde{\boldsymbol{M}}}\mathcal{L}_\mu\left(\widetilde{\boldsymbol{M}},\boldsymbol{Z}^t,\boldsymbol{Y}^t\right) \\ =&\min_{\widetilde{\boldsymbol{M}}}\frac{1}{2}\left\|\widetilde{\boldsymbol{M}}-\boldsymbol{Z}^t+\frac{1}{\mu}\boldsymbol{Y}^t\right\|_F^2+\frac{\lambda}{\mu}\sum_{i=1}^{k}\|\boldsymbol{M}_i\|_2 \\ =&\min_{\boldsymbol{M}_1,\cdots,\boldsymbol{M}_k}\sum_{i=1}^{k}\left\{\frac{1}{2}\left\|\boldsymbol{M}_i-\boldsymbol{Q}_i^t\right\|_F^2+\frac{\lambda}{\mu}\|\boldsymbol{M}_i\|_2\right\}, \end{aligned} \tag{24}$$

where $\boldsymbol{Q}_i^t$ is the $i$-th column-triplet of $\boldsymbol{Z}^t-\frac{1}{\mu}\boldsymbol{Y}^t$. Therefore, we can update each $\boldsymbol{M}_i$ by solving a proximal problem based on Proposition 2:

$$\boldsymbol{M}_i^{t+1}=\mathcal{D}_{\frac{\lambda}{\mu}}(\boldsymbol{Q}_i^t),\ \ \forall i\in[1,k]. \tag{25}$$

For the step in (22), $\mathcal{L}_\mu\left(\widetilde{\boldsymbol{M}}^{t+1},\boldsymbol{Z},\boldsymbol{Y}^t\right)$ is a quadratic form of $\boldsymbol{Z}$ and admits the following closed-form solution:

$$\boldsymbol{Z}^{t+1}=\left(\boldsymbol{W}\widetilde{\boldsymbol{B}}^T+\mu\widetilde{\boldsymbol{M}}^{t+1}+\boldsymbol{Y}^t\right)\left(\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^T+\mu\boldsymbol{I}\right)^{-1}. \tag{26}$$

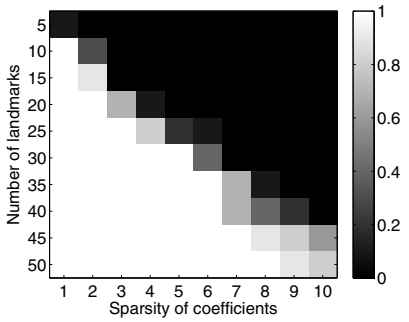


Figure 1. The frequency of exact recovery on synthesized data.

It can be proven that the sequences of values produced by the ADMM iterations in (21) to (23) converge to the optimal solutions of the primal problem in (19) [8], which are also the optimal solutions to the original problem in (13).

## 5. Experiments

### 5.1. Simulation

We aim to investigate whether the spectral-norm minimization in (12) can exactly solve the ill-posed inverse problem based on the prior knowledge of sparsity and orthogonality under noiseless cases.

More specifically, we randomly simulate $k$ basis shapes $\boldsymbol{B}_1,\cdots,\boldsymbol{B}_k\in\mathbb{R}^{3\times p}$ ($p$ is varying, $k=50$) with entries sampled independently from the normal distribution $\mathcal{N}(0,1)$, and simulate $k$ rotation matrices $\boldsymbol{R}_1,\cdots,\boldsymbol{R}_k$ as well as coefficients $c_1,\cdots,c_k$. Only $z$ randomly-selected coefficients are nonzero with values sampled from the uniform distribution $\mathcal{U}(0,1)$. Then, $\boldsymbol{M}_i=c_i\bar{\boldsymbol{R}}_i\in\mathbb{R}^{2\times 3}$ and $\boldsymbol{W}=\sum_{i=1}^{k}\boldsymbol{M}_i\boldsymbol{B}_i$. We use $\boldsymbol{W}$ as the input and solve (12) to estimate $\boldsymbol{M}_i$s. The solution is regarded as exact if $\|\hat{\boldsymbol{M}}-\widetilde{\boldsymbol{M}}\|_F/\|\widetilde{\boldsymbol{M}}\|_F<10^{-3}$, where we concatenate $\boldsymbol{M}_i$s in $\widetilde{\boldsymbol{M}}$, and $\hat{\boldsymbol{M}}$ is the algorithm estimate.

Figure 1 reports the frequency of exact recovery with varying $p$ (number of landmarks) and $z$ (sparsity of the underlying coefficients), which is evaluated over 10 randomly-generated instances for each setting. Note that the number of unknowns ($6k$) is much larger than the number of equations ($2p$). The proposed convex program can exactly solve the problem with a frequency equal to 1 in the lower-triangular area, where the number of landmarks is sufficiently large and the coefficients are truly sparse. This demonstrates the power of convex relaxation, which has proven to be successful in various inverse problems, e.g., compressed sensing [11] and matrix completion [10]. The performance drops in more difficult cases in the upper-triangular area. This observation is analogous to the phase transition in compressive sensing, where the recovery probability also depends on the number of observations and the underlying signal sparsity [16].

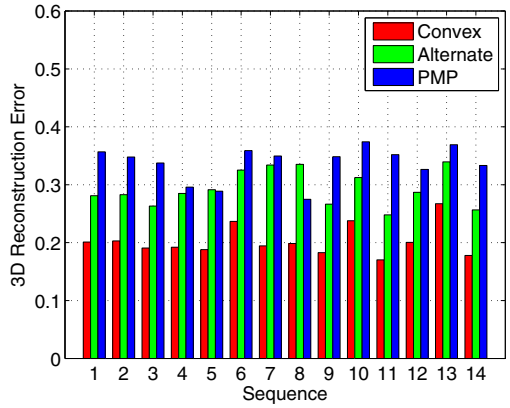


Figure 2. The mean reconstruction error for each sequence of Subject 15 from the MoCap dataset. Three methods are compared: "Convex" denotes the proposed convex method; "Alternate" means the alternating minimization method; "PMP" represents the method proposed in [32].

| | Convex | Alternate | PMP |
|---|---|---|---|
| Subject 13 | 0.259 | 0.293 | 0.390 |
| Subject 14 | 0.258 | 0.308 | 0.393 |
| Subject 15 | 0.204 | 0.286 | 0.340 |

Table 1. The mean errors over all sequences of three subjects from the MoCap dataset.

## 5.2. Applications

### 5.2.1 Human Pose Estimation

The applicability of sparse shape representation for 3D human pose recovery has been thoroughly studied in previous work [32, 35, 18]. In this paper, we aim to illustrate the advantage of the proposed convex program compared to the alternating minimization widely used in previous work. We carry out evaluation on the MoCap dataset [1] and use the sequences from Subject 86 as training data and the sequences from Subject 13, 14 and 15 as testing data. All of the selected subjects are conducting a large variety of activities such as running, jumping, boxing, basketball, etc.

Since there are thousands of training shapes, using all of them as basis shapes is impractical. For our method, we solve the following problem to learn a shape dictionary:

$$\begin{aligned} \min_{\boldsymbol{B}_1,\cdots,\boldsymbol{B}_k,\boldsymbol{C}} \quad & \sum_{j=1}^{n} \frac{1}{2} \|\boldsymbol{S}_j - \sum_{i=1}^{k} C_{ij}\boldsymbol{B}_i\|_F^2 + \beta \sum_{i,j} C_{ij} \\ \text{s.t.} \quad & C_{ij} \geq 0,\ \|\boldsymbol{B}_i\|_F \leq 1, \\ & \forall i \in [1,k],\ j \in [1,n], \end{aligned} \tag{27}$$

where $\boldsymbol{B}_i$s are the basis shapes to be learned, $\boldsymbol{S}_i$s denote the training shapes (aligned by the Procrustes method), and $C_{ij}$ represents the $i$-th coefficient for the $j$-th training shape.

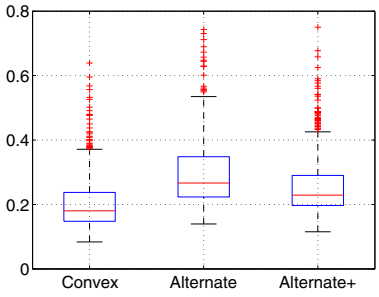


Figure 3. The barplots of estimation errors on the MoCap dataset (Subject 15) for the proposed method ("Convex"), the alternating minimization ("Alternate") and the alternating minimization initialized by the convex method ("Alternate+").

We initialize the dictionary by uniformly selecting $k$ shapes from the training data and locally solving (27) by alternately updating $\boldsymbol{C}$ and $\boldsymbol{B}_i$s, a strategy widely used in dictionary learning literature [27]. We use the 15 joints model as shown in Figure 4 and set $k = 64$

We compare the proposed method to Projected Matching Pursuit (PMP) from Ramakrishna et al. [32][1]. We also implement an alternating minimization method that solves the model in (7) by alternately updating the shape parameter $\boldsymbol{c}$ via $\ell_1$ minimization and updating the pose parameter $\bar{\boldsymbol{R}}$ via manifold optimization. The manifold optimization is implemented with the Manopt toolbox [7] to update $\bar{\boldsymbol{R}}$ by the trust-region algorithm over the Stiefel manifold. The alternating minimization is initialized by the mean shape of the training shapes. For both of the proposed method and the alternating minimization method, we set the regularization parameter as $\lambda = 0.1$ for all sequences.

The reconstruction error is evaluated by the Euclidean distance between the reconstructed shape and the true shape up to a similarity transformation. The mean errors for the 14 testing sequences from Subject 15 are shown in Figure 2. The subject is conducting various activities in different sequences [1]. The proposed convex algorithm clearly outperforms the alternative methods and achieves a stable performance for all sequences. The mean error over all of the sequences for each subject is given in Table 1.

To verify that the alternating minimization depends on initialization, we initialize the alternating minimization with the solution of our method. The results for Subject 15 are shown in Figure 3. The error of the alternating minimization is apparently decreased with a smaller variance by using the better initialization. The mean objective values of alternating minimization with and without the convex initialization are 0.17 and 0.24, respectively[2]. The accuracy

[1]The code is downloaded from the authors' website http://www.cs.cmu.edu/~vramakri/research.html

[2]The objective of the convex formulation is different and therefore not compared.

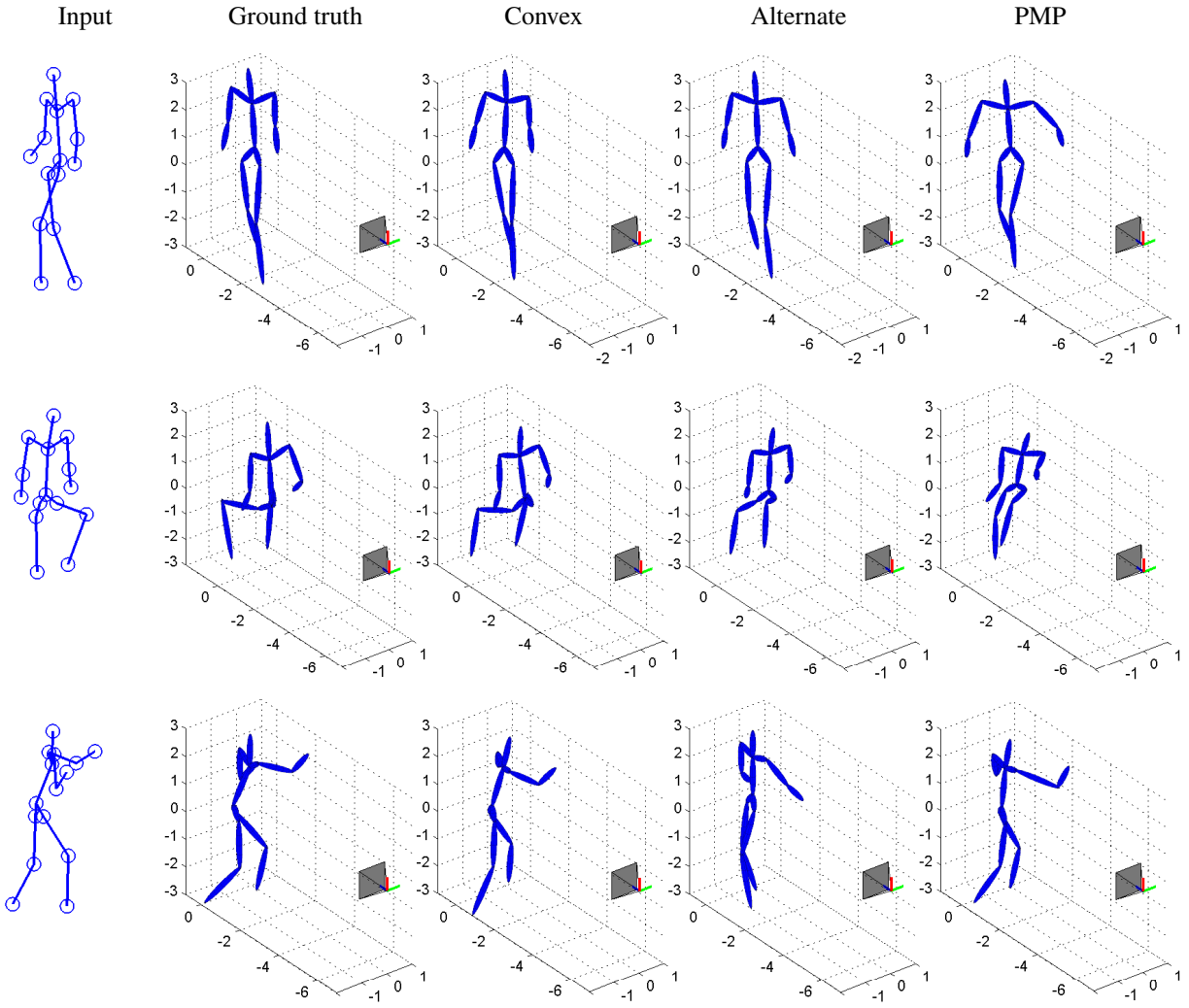


Figure 4. Examples of human pose estimation. The columns from left to right correspond to the input 2D landmarks, the ground-truth 3D pose, and the reconstructions from the proposed method, the alternating minimization, and the PMP method [32], respectively.

of "Alternate+" is worse than the convex formulation. This might be attributed to the fact that the shape model in (8) offers more degree of freedom than the original model in (3) to represent complex deformation of a human skeleton.

The reconstructed poses for three selected frames are visualized in Figure 4. We can see that all methods perform well in the first example, where the shape (walking) is close to the mean shape (standing straight). But the accuracies of the alternative methods degrade in the other two examples, where the shape is far away from the mean shape, while our method still obtains appealing reconstructions.

#### 5.2.2 Car Reconstruction

We demonstrate the applicability of the proposed method for 3D car shape estimation using the recently-published Fine-Grained 3D Car dataset [26], which provides images of cars, 2D landmark annotations and corresponding 3D models. We concatenate the 3D models of 15 cars as the shape dictionary and try to reconstruct the 3D models of other cars from the visible landmarks annotated in the images (∼40 points per image). The 3D models provided in the dataset were reconstructed by the authors instead of true CAD models. Therefore, we only show some qualitative results. As illustrated in Figure 5, our method can successfully reconstruct the models of various classes such as sedan, SUV and pick-up truck. For comparison, we also show the results of an alternative method proposed in the original paper [26], which uses the perspective camera model and nonlinear optimization. The alternative method initialized by the mean shape performs well in the sedan example but relatively poor in the SUV and truck examples, where the models deviate far away from the mean shape. Similar results were reported in the original paper [26] and

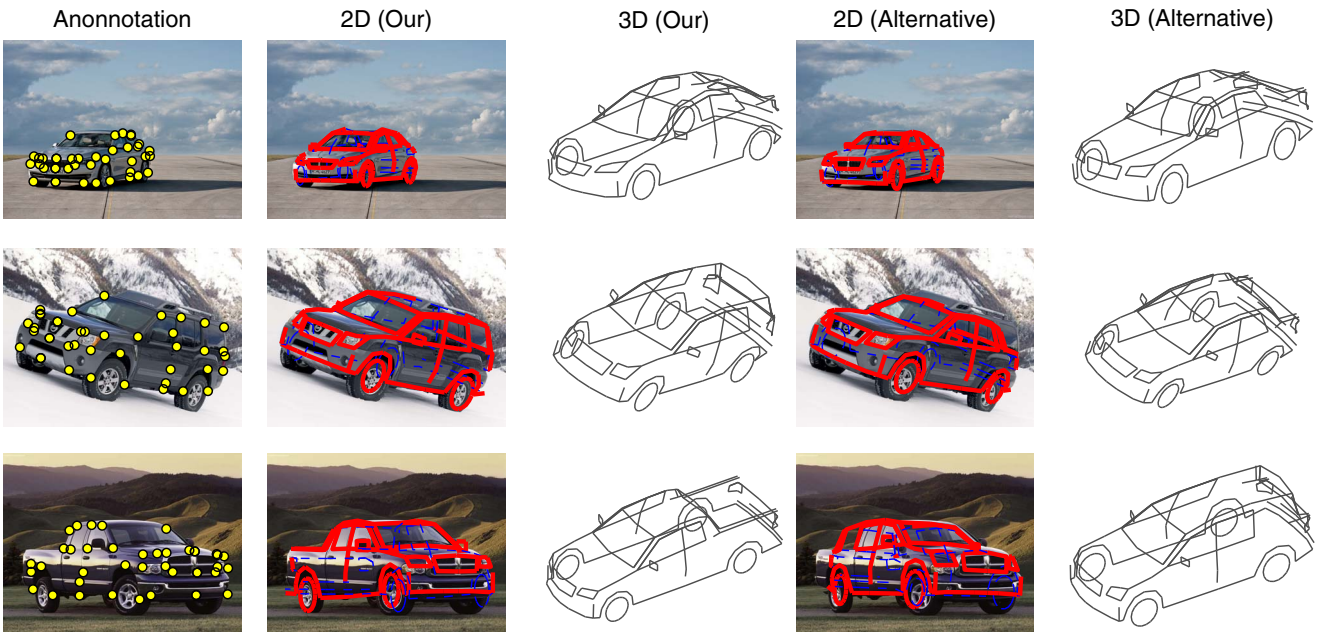


Figure 5. Examples of car reconstruction. The columns from left to right correspond to the input 2D landmarks, the 2D fitted models and 3D reconstructions of the proposed method, and the results of the alternative method [26], respectively. Only visible landmarks ($\sim 40$ per image) are used for shape fitting. The 3D models are visualized in novel views. The car models from top to bottom are the BMW 5 Series 2011 (sedan), the Nissan Xterra 2005 (SUV) and the Dodge Ram 2003 (pick-up truck), respectively.

the authors proposed to use the class-specific mean shape for better initialization. Instead, our method can achieve appealing results with arbitrary initialization.

### 5.3. Computational Time

Our algorithm is implemented in MATLAB and tested on a desktop with a Intel i7 3.4GHz CPU and 8G RAM. In our experiments, the ADMM algorithm generally converges within 500 iterations to reach a stopping criterion of $10^{-4}$. In the experiments of human pose estimation, for example, the computational time of our algorithm is 0.33s per frame, while those of the alternating minimization and the PMP algorithm [32] are 0.44s and 3.02s, respectively.

## 6. Discussion

In summary, we proposed a method for aligning a 3D shape-space model to 2D landmarks by solving a convex program, which guarantees global optimality. Intuitively, we adopted an augmented 3D shape model to achieve a linear representation of shape variability in 2D and proposed to use the spectral-norm regularization to penalize invalid cases caused by the augmentation.

The exactness of using convex relaxation for linear inverse problems with various assumptions, e.g., sparsity and orthogonality, has been theoretically analyzed in literature, e.g., [13]. In our experiments, we observed that the estimates satisfied the original constraints in most cases, and all reported results were the outputs of the proposed algorithm without refinement. In cases where the relaxation is not tight, postprocessing steps may be employed to enforce the exactness, e.g., projecting the estimated rotation matrix to $SO(3)$ or forcing the basis shapes to share the same rotation. This might be helpful in real applications of modeling rigid objects, although we did not use them in our experiments.

In this paper, we assume that the 2D landmarks and 3D-2D correspondences are given. Our method can be naturally extended to handle large errors in landmark localization in practice. For examples, the $\ell_1$-norm can be used to replace the squared loss in (13) to make the model more robust against outliers, and the optimization can be solved by ADMM as well. Another possible solution is to use RANSAC as proposed in [24], since the shape model can be estimated using only a portion of the landmarks. Also, there is a great potential to integrate the proposed shape model with existing landmark-localization methods to simultaneously localize 2D landmarks and recover shapes.

**Acknowledgments**: Grateful for support through the following grants: NSF-DGE-0966142, NSF-IIS-1317788, NSF-IIP-1439681, NSF-IIS-1426840, ARL MAST-CTA W911NF-08-2-0004, ARL RCTA W911NF-10-2-0016, and ONR N000141310778. Xiaoyan Hu was supported by NSFC (No.61103086 and 61170186).

## References


[1] Mocap: Carnegie mellon university motion capture database. http://mocap.cs.cmu.edu/. 6

[2] A. Agudo, L. Agapito, B. Calvo, and J. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *CVPR*, 2014. 2

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010. 1

[4] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 1

[5] A. Blake and M. Isard. *Active contours*. Springer, 2000. 3

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003. 2

[7] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. 6

[8] S. Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. 5

[9] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 2, 3

[10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. 5

[11] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008. 5

[12] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 2

[13] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012. 4, 8

[14] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 1, 2

[15] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear modeling via augmented lagrange multipliers (balm). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1496–1508, 2012. 2

[16] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009. 5

[17] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 3

[18] X. Fan, K. Zheng, Y. Zhou, and S. Wang. Pose locality constrained representation for 3d human pose reconstruction. In *ECCV*, 2014. 2, 6

[19] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2

[20] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *Advances in Neural Information Processing Systems*, 2012. 1

[21] L. Gu and T. Kanade. 3D alignment of face in a single image. In *CVPR*, 2006. 2

[22] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *Advances in Neural Information Processing Systems*, 2012. 1, 2

[23] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010. 3

[24] Y. Li, L. Gu, and T. Kanade. Robustly aligning a shape model and its application to car alignment of unknown pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1860–1876, 2011. 2, 8

[25] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 1

[26] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, 2014. 2, 7, 8

[27] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. 6

[28] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006. 1

[29] M. Paladini, A. Del Bue, J. Xavier, L. Agapito, M. Stošić, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision*, 96(2):252–276, 2012. 2, 3

[30] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013. 4

[31] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In *ECCV*, 2012. 1

[32] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, 2012. 1, 2, 3, 6, 7, 8

[33] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, 2012. 1

[34] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *CVPR*, 2013. 1

[35] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014. 1, 2, 6

[36] X. K. Wei and J. Chai. Modeling 3d human poses from uncalibrated monocular images. In *ICCV*, 2009. 1

[37] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012. 1

[38] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+ 3d active appearance models. In *CVPR*, 2004. 3

[39] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006. 2

[40] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2

[41] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. Metaxas, and X. Zhou. Sparse shape composition: A new framework for shape prior modeling. In *CVPR*, 2011. 3

[42] F. Zhou and F. De la Torre. Spatio-temporal matching for human detection in video. In *ECCV*, 2014. 2

[43] S. Zhu, L. Zhang, and B. M. Smith. Model evolution: an incremental approach to non-rigid structure from motion. In *CVPR*, 2010. 3

[44] Y. Zhu, D. Huang, F. De la Torre Frade, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *CVPR*, 2014. 3

[45] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2608–2623, 2013. 1, 2