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

Xiaowei Zhou[1], Spyridon Leonardos[1], Xiaoyan Hu[12], Kostas Daniilidis[1]
1 University of Pennsylvania. 2 Beijing Normal University

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 instead of merely providing bounding boxes. Estimating the 3D configuration 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 models.

A popular approach is the "active shape model" [3], where each shape is defined by a set of landmarks and the shape to be estimated is assumed to be a linear combination of predefined basis shapes. Given 3D-2D correspondences, the 3D deformable model is fitted to the landmarks annotated or detected in images. While this approach has proven to be successful in various applications [4, 6], a challenging issue remains, i.e., the joint estimation of shape parameters and camera-pose parameters requires to solve a nonconvex optimization problem. The existing methods often adopt an alternating minimization scheme to locally update the variables, and consequently the solution is sensitive to initialization as illustrated in Figure 1.

In this paper, we propose a convex formulation to address this problem. We use an augmented shape-space model, where a shape is represented as a linear combination of rotatable basis shapes giving a linear representation of both shape and viewpoint variability. Then, we use the convex relaxation of the orthogonality constraint to convert the entire problem into a convex program. Finally, we develop an efficient algorithm to solve the problem.

With the weak-perspective camera model and the sparse representation of shapes, the following problem is considered to estimate a shape:

$$\begin{aligned} \min_{c,\bar{R}} \quad & \frac{1}{2}\left\| W - \bar{R}\sum_{i=1}^{k} c_i B_i \right\|_F^2 + \lambda \|c\|_1, \\ \text{s.t.} \quad & \bar{R}\bar{R}^T = I_2, \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{2\times p}$ denotes the 2D landmarks, $\bar{R} \in \mathbb{R}^{2\times 3}$ represents the first two rows of a rotation matrix, $B_1, \cdots, B_k \in \mathbb{R}^{3\times p}$ are basis shapes learned from training data and $c = [c_1, \cdots, c_k]^T$ is the coefficient vector to represent the unknown shape S, such that $S = \sum_{i=1}^{k} c_i B_i$. The optimization in (1) is nonconvex and there is an orthogonality constraint. A commonly-used strategy is to alternate between the updates of $\bar{R}$ or c while fixing the other one, which only gives a locally-optimal solution.

We propose to use the shape model $S = \sum_{i=1}^{k} c_i R_i B_i$, in which there is a rotation for each basis shape, and the corresponding 2D model is

$$W = \Pi \sum_{i=1}^{k} c_i R_i B_i = \sum_{i=1}^{k} M_i B_i, \tag{2}$$

where $M_i \in \mathbb{R}^{2\times 3}$ is the product of c_i and the first two rows of R_i, which satisfies $M_i M_i^T = c_i^2 I_2$. The motivation of using this model is to achieve a linear representation of shape and viewpoint variability, such that we can get rid of the bilinear form in (1).

Next, we replace the orthogonality constraint on M_is by its convex counterpart. The following lemma has been proven in literature [5, Section 3.4] and Proposition 1 can be derived from it.

Lemma 1. *The convex hull of* $\mathcal{Q} = \{X \in \mathbb{R}^{m\times n} | \; X^T X = I_n\}$ *equals the unit spectral-norm ball* $conv(\mathcal{Q}) = \{X \in \mathbb{R}^{m\times n} | \; \|X\|_2 \le 1\}$. $\|X\|_2$ *denotes the spectral norm, which is defined as the largest singular value of* X.

Proposition 1. *The convex hull of* $\mathcal{S} = \{Y \in \mathbb{R}^{m\times n} | \; Y^T Y = s^2 I_n\}$ *equals* $conv(\mathcal{S}) = \{Y \in \mathbb{R}^{m\times n} | \; \|Y\|_2 \le |s|\}$.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

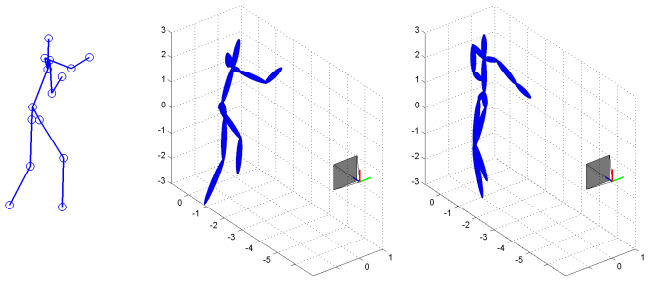

Figure 1: An example of 3D human pose recovery. The columns from left to right correspond to the input 2D landmarks, the reconstructions from the proposed method and from the alternating minimization, respectively.

With Proposition 1 and the shape model in (2), the original problem in (1) is relaxed to:

$$\begin{aligned} \min_{c_1,\cdots,c_k,M_1,\cdots,M_k} \quad & \frac{1}{2}\left\| W - \sum_{i=1}^{k} M_i B_i \right\|_F^2 + \lambda \sum_{i=1}^{k} |c_i|, \\ \text{s.t.} \quad & \|M_i\|_2 \le |c_i|, \; \forall i \in [1,k] \end{aligned} \tag{3}$$

which is apparently equivalent to

$$\min_{M_1,\cdots,M_k} \; \frac{1}{2}\left\| W - \sum_{i=1}^{k} M_i B_i \right\|_F^2 + \lambda \sum_{i=1}^{k} \|M_i\|_2. \tag{4}$$

The problem in (4) is a penalized least-squares problem, where we estimate a set of orthogonal matrices by minimizing their spectral norms. We provide an efficient algorithm based on ADMM [1] to solve it.

Notice that $\|\cdot\|_2$ denotes the spectral norm of a matrix instead of the ℓ_2-norm of a vector. As we show in the paper, minimizing the spectral norm of a matrix is equivalent to minimizing the ℓ_∞-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 M_i to be orthogonal (an orthogonal matrix has equal singular values). Interestingly, the conditions for exact recovery using spectral-norm relaxation has been theoretically analyzed in [2]. We provide numerical results in the paper.

[1] 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.

[2] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

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

[4] Mohsen Hejrati and Deva Ramanan. Analyzing 3d objects in cluttered images. In *Advances in Neural Information Processing Systems*, 2012.

[5] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

[6] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European conference on Computer Vision*, 2012.