3D Human Pose Reconstruction - A convex optimization approach

Chun-Han Yao
A53234899
chy235@eng.ucsd.edu

Tsunghan Lee
A53219632
tsl021@eng.ucsd.edu

Yuwei Wang
A53212966
yuw267@ucsd.edu

Abstract

As more and more sensing technologies is affordable nowadays, cameras are the most common and ubiquitous equipment in daily life. Considering that cameras nowadays are not only cheap and lightweight but precise with high resolutions, constructing 3D models from image sequence becomes a more viable alternative. In this paper, we review the state-of-the-art methods and implemented an optimization algorithm for 3D human pose reconstruction. We further use ADMM when computing. Our work shows that this algorithm can perform well on most of images with human poses.

1. Introduction

The desires to apply computer vision technologies on products has become popular in these years. With the growth of games, movie industry, and virtual/augmented reality applications, there has been increasing demands for geometry 3D models. However, existing solutions are not fully satisfying. User-driven modeling is time-consuming and error-prone, while 3D scanners are costly and cumbersome. As more and more sensing technologies is affordable nowadays, cameras are the most common and ubiquitous equipment in daily life. Considering that cameras nowadays are not only cheap and lightweight but precise with high resolutions, constructing 3D models from image sequence becomes a more viable alternative.

Also, although the computation of devices is much more powerful than in the past, application of computer vision still relies on the mathematical techniques, especially convex optimization method. Therefore, this paper, first, reviews the state-of-the-art models. Second, we implemented a optimization algorithm for 3D human pose reconstruction. Furthermore, we leverage a ADMM algorithm for optimization. Finally, we show our implementation result by applying it on multiple types of pictures. The program does well in most of the images.

2. Related Work

Though the video [5] shows that the 3D shape of an object can be reconstructed from 2D images, such technique is not sufficient enough to handle all the object reconstruction. In practice, many non-rigid objects, e.g. the human poses, face can deform with certain structures. Intuitively, the only difference between non-rigid and rigid situations is that the non-rigid shape is a weighted combination of certain shape bases. Thus, knowing these bases is important in non-rigid structure recovering.

This work is correlated to nonrigid structure from motion (NRSfM), i.e. a deformable shape can be recovered from multi-frame 2D-2D correspondences. Torresani et al. [4] proposed to recover time-varying shape and motion of nonrigid 3D objects from uncalibrated 2D point tracks. In NRSfM, the low-rank shape-space model has been frequently used, but the basis shapes are still unknown. Typically, the joint estimation of shape variables and basis shapes is solved via matrix factorization. Christoph et al. proposed the first model free approach that can recover non-rigid shape models from single-view video sequences [2]. Moreover, Jing Xiao et al. provided anther closed-form solution to this problem by introducing other two constraints, rotation constraints & basis constraints [6]. Some other works use iterative algorithm [3] or sequential process [1] for better performance.

3. Model

3.1. Problem Formulation

In this framework, the goal is to estimate the 3D shape of the object from a single 2D image. The unknown 3D model is defined by a set of landmarks and assumed to be a linear combination of some predefined basis shapes with sparse coefficients.

Let $S \in \mathbb{R}^{3\times p}$ denote the 3D locations of the p landmarks, which are the joints of the human bodies in the case shown in Fig.1, and $W \in \mathbb{R}^{2\times p}$ is their projection on a 2D image. The relation can be represented as follows:

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

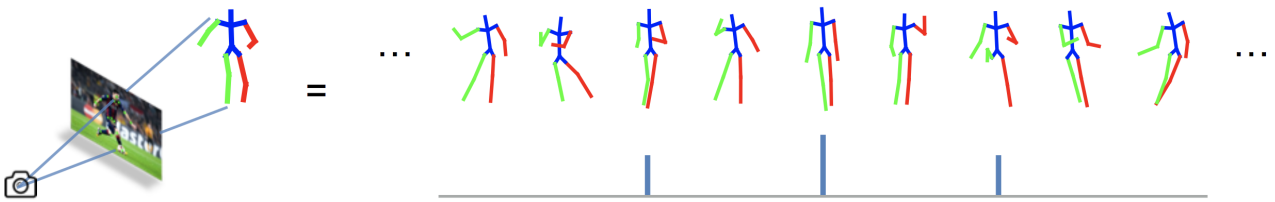

Figure 1. An overview of the framework, in which the 3D shape is defined by a set of landmarks and assumed to be a linear combination of some predefined basis shapes with sparse coefficients.

, where S is represented as a linear combination of k basis shapes B.

$$S = \sum_{i=0}^{k} c_i B_i \tag{2}$$

For simplicity, the object is assumed to be far away from the camera. Therefore, the camera calibration matrix Π can be reduced to a simple form.

$$\Pi = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix} \tag{3}$$

We further introduce $R \in \mathbb{R}^{3\times3}$ and $T \in \mathbb{R}^3$ for the rotation and translation of the object, respectively.

$$W = \Pi(R\sum_{i=0}^{k} c_i B_i + T1^T) \tag{4}$$

The projection matrix Π can be combined with the rotation R as $\tilde{R}$, which is the first two rows of R multiplied by a scalar s. The translation T can also be eliminated by centralizing the data.

$$W = \tilde{R}\sum_{i=0}^{k} c_i B_i \tag{5}$$

With the representation in Equation.5, 3D shape reconstruction can be formulated as the following optimization problem:

$$\min_{c,\tilde{R}} \frac{1}{2} \left\| W - \tilde{R}\sum_{i=0}^{k} c_i B_i \right\|_2^F + \alpha \|c\|_1 \tag{6}$$

$$subject\ to \quad \tilde{R}\tilde{R}^T = I_2 \tag{7}$$

The first term of the objective function stands for the reprojection error, and the second term is the l-1 norm of the linear coefficients, enforcing the representation to be sparse. Note that the rotation matrix $\tilde{R}$ is constrained to be orthonormal, which in general is not a convex set.

To address the problem, one can use the method of alternating minimization. That is, we fix $\tilde{R}$ and update c, then fix c and update $\tilde{R}$, until convergence. However, due to the non-convexity of the problem, we may be stuck at local minimum easily if the initialization is far away from the true solution.

3.2. Convex Relaxation

To obtain the optimal 3D reconstruction, a convex formulation is proposed to approximate the original one, which is called convex relaxation.

First, we replace the original projection model by a shape-space model, in which there is a rotation for each basis shape. With the merit of this change, we can combine c and R into M and get rid of the bilinear form in the previous formulation.

$$W = \Pi\sum_{i=0}^{k} c_i R_i B_i = \sum_{i=0}^{k} M_i B_i \tag{8}$$

,where $M_i \in \mathbb{R}^{2\times3}$ is the product of c_i and the first two rows of R_i. Now the formulation can be written as:

$$\min_{c_1,\ldots,c_k,M_i,\ldots,M_k} \frac{1}{2} \left\| W - \sum_{i=0}^{k} M_i B_i \right\|_2^F + \alpha \|c\|_1 \tag{9}$$

$$subject\ to \quad M_i M_i^T = c_i I_2 \tag{10}$$

Subsequently, the following proposition is used to relax the orthogonality constraint on M, which makes the problem non-convex.

Proposition:
Given a set of orthogonal matrices $S = \{M \in R^{m\times n} \mid M^T M = c^2 I_n\}$, its convex hull $conv(S) = \{M \in R^{m\times n} \mid \|M\|_2 \leq |c|\}$

The orthogonality constraint is then turned into a constraint on the spectral norm of M.

$$\min_{c_1,\ldots,c_k,M_i,\ldots,M_k} \frac{1}{2} \left\| W - \sum_{i=0}^{k} M_i B_i \right\|_2^F + \alpha \|c\|_1 \tag{11}$$

$$subject\ to \quad \|M_i\|_2 \leq |c_i| \tag{12}$$

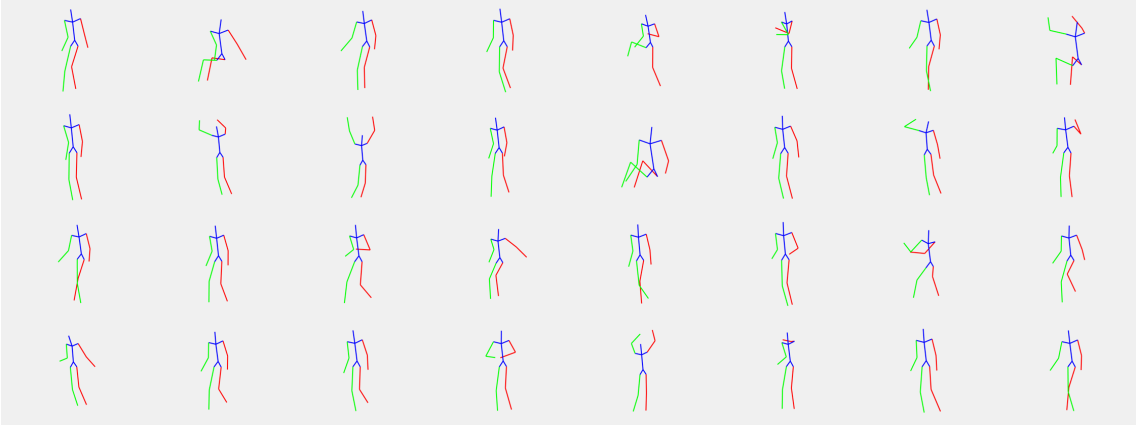

Figure 2. Examples of the basic shapes. Each shape is composed by 15 points.

Moreover, we can replace the second term of the objective function by the spectral norm of M and get rid of the constraint since the optimal value occurs when equality holds.

Consequently, the final formulation can be written as follows:

$$\min_{M_i,...,M_k} \frac{1}{2} \left\| W - \sum_{i=0}^{k} M_i B_i \right\|_2^F + \alpha \sum_{i=1}^{k} \|M_i\|_2 \quad (13)$$

Note that minimizing the spectral norm $\|.\|_2$ of a matrix is equivalent to minimizing the $l-\infty$ norm of the vector of its singular values. Therefore, by spectral-norm minimization, we can not only minimize the number of activated basis shapes but also enforce M_i to be orthogonal, since an orthogonal matrix has equal singular values.

The final formulation is convex, yet it is not trivial to optimize the objective function directly. In Sec.3.3 we will briefly introduce several optimization methods used in the framework.

3.3. Optimization

First, the Alternating Direction Method of Multipliers (ADMM) is applied. Suppose the original objective function have multiple terms that are not easy to solve jointly. For instance, the original problem is represented as:

$$\min_x f(x) + g(x) \quad (14)$$

The ADMM method introduces an auxiliary variable along with an additional constraint so that the different terms can be updated separately.

$$\min_{x,y} f(x) + g(y) \quad (15)$$

$$subject\ to \quad x = y \quad (16)$$

In our problem, an auxiliary variable Z is introduced in order to optimize the two terms of the objective function alternatively.

$$\min_{\tilde{M},Z} \frac{1}{2} \left\| W - Z\tilde{B} \right\|_2^F + \alpha \sum_{i=0}^{k} \|M_i\|_2 \quad (17)$$

$$subject\ to \quad \tilde{M} = Z \quad (18)$$

, where $\tilde{M} = [M_1, M_2, ..., M_k]$, and $\tilde{B} = [B_1, B_2, ..., B_k]^T$.

The second approach of optimization is the augmented Lagragian method, which is similar to the penalty method. Suppose that the original problem can be formulated as:

$$\min_x f(x) \quad (19)$$

$$subject\ to \quad c_i(x) = 0 \quad (20)$$

The penalty method adds a penalty term to the objective function:

$$\min_x \Phi_k(x) = f(x) + \mu_k \sum_i c_i(x)^2 \quad (21)$$

In each iteration of optimization, the problem is resolved with a larger μ_k. It is obvious that the original constraint holds as μ_k approaches infinity. However, it is not computationally reasonable to solve the problem with μ_k equals to infinity.

Alternatively, the augmented Lagrangian method relaxes the constraint by some value λ/μ, thus we can achieve the optimal value with a finite value of μ.

$$\min_x f(x) \quad (22)$$

$$subject\ to \quad c_i(x) = 0 \quad (23)$$

After the relaxed constraint is plugged into the penalty term, a new objective function, also called the augmented Lagrangian, can be derived as:

$$\min_x \Phi_k(x) = f(x) + \frac{\mu_k}{2}\sum_i c_i(x)^2 - \sum_i \lambda_i c_i(x) \quad (24)$$

As described above, the augmented Lagrangian $\mathcal{L}_\mu(\tilde{M}, Z, Y)$ of our problem can be written as:

$$\frac{1}{2}\left\|W - Z\tilde{B}\right\|_2^F + \alpha\sum_{i=0}^{k}\|M_i\|_2 + \langle Y, \tilde{M} - Z\rangle + \frac{\mu}{2}\left\|\tilde{M} - Z\right\|_2^F \quad (25)$$

, where Y is the dual variable and μ is the parameter controlling the step size in optimization. To minimize the augmented Lagrangian, we iteratively update $\tilde{M}$, Z, and Y until convergence:

$$\tilde{M}^{t+1} = \text{argmin}_{\tilde{M}}\ \mathcal{L}_\mu(\tilde{M}, Z^t, Y^t) \quad (26)$$

$$Z^{t+1} = \text{argmin}_{Z}\ \mathcal{L}_\mu(\tilde{M}^{t+1}, Z, Y^t) \quad (27)$$

$$Y^{t+1} = Y^t + \mu(\tilde{M}^{t+1} - Z^{t+1}) \quad (28)$$

Note that Y can be seen as an estimate of the Lagrange multiplier, and the accuracy of this estimate improves at every iteration.

3.4. Reconstruction

Finally, with the optimal M, we can directly derive c and R and reconstruct the 3D shape of the object, as described in Algorithm 1.

Algorithm 1 Direct Reconstruction.

Input: M_1, M_2, ..., M_k;
Output: S;

1: **for** $i = 1$ to k **do**
2: $\quad c_i = \|M_i\|_2$;
3: $\quad r_i^{(1)} = m_i^{(1)}/c_i$;
4: $\quad r_i^{(2)} = m_i^{(2)}/c_i$;
5: $\quad r_i^{(3)} = m_i^{(3)}/c_i$;
6: $\quad R_i = [\, r_i^{(1)}, r_i^{(2)}, r_i^{(3)}]^T$;
7: **end for**
8: $S = \sum_{i=0}^{k} c_i R_i B_i$;
9: **return** S;

4. Experiments

We implemented the algorithm and tested it on several photos. We used 128 human poses as basic shapes. Each shape is composed by 3D co-ordinations of 15 joint point of human. Some of the basic shapes are shown in Figure 2. The basic shapes contains human poses of standing, jumping, running, etc. Our algorithm should compute a linear combination of the basic shapes to minimize the difference to the 2D projection of the joint points in the origin photos. For each photo, we manually marked 15 joint points of the person. Then, we ran the program to reconstruct the 3D pose of the photos.

Our program is implemented in MATLAB. The optimization algorithm converged in 1,000 iterations. Finally, we obtained a linear combination of the basic shapes. The results are shown in Figure 3. We used some photos of running, standing and jumping people. We also used some photos of weird poses as well as some photos of gorillas. The results reveal that the algorithm did well in reconstructing the 3D poses for the origin images. It could even reconstruct the poses of gorilla climbing a tree. The results in for move poster and an upside-down person are not as good as others. This is because the poses in the two photos are far away from the basic shapes. Thus, the error in the results are larger than others.

5. Conclusion

In conclusion, our project implemented a optimization algorithm for 3D human pose reconstruction. We used a ADMM algorithm to do the optimization. We did the experiments on multiple types of images. The program did well in most of the images.

A drawback of the algorithm is that it relies on the basic shapes. Thus, the algorithm can not perform well in some weird poses. In addition, it needs to manually mark the joint points of a human photo. In the future, we can use computer vision algorithms to automatically capture the landmark points of a human so that the algorithm can be widely used.

References

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

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

[3] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear modeling via augmented lagrange multipliers (balm), 2012.

[4] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors, 2008.

[5] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc, 2014.

[6] J. Xiao, J.-x. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery, 2004.

Origin images | Images with marked joint points | Results

Figure 3. Results of the experiments.