# Estimating 3D Human Pose by A Few Clicks

Chen Kong (chenk)
Carnegie Mellon University
chenk@cs.cmu.edu

## Abstract

*We develop an application for mobile devices to estimate 3D human pose from a single image. Due to widely used touch screens in apple devices (*e.g. *iPhone, iPad), annotating key joints of a human body for a single image is no more than just a few touches/clicks. Our application leverages this feature such that users can take photos, annotate key joints rapidly and get an estimated 3D human pose at the end. We test our application using both images downloaded from Internet and taken by users in daily life.*

## 1. Introduction

Estimating 3D human pose from a single image, as an application of estimating the 3D shape of an object, has received more and more attention in computer vision community. A common strategy for this task involves two components: (i) key joints detection and (ii) 3D inversion. Key joints detection deals with the problems of determining the locations of key joints (*e.g*. head, elbows, knees, and *etc*.) in a single image. Once key joints have been established, the inversion problem of recovering the 3D structure from the 2D point projections must be solved, requiring a priori constraints on structure and camera matrix.

Even though the state-of-the-art key joints detection method [1] using convolutional neural network achieves an almost perfect performance, it consumes a huge amount of computation, memory and power *i.e*. a desktop tower with four high-performance GPUs. This is problematic for mobile device application where computation, memory and especially power are highly limited. However, annotating human joints by using a touch screen is trivial for users and no more than a few clicks. Therefore, in this project, we develop a GUI for users to rapidly annotate human joints manually instead of using deep neural networks.

As for 3D inversion, we learn a shape dictionary beforehand by leveraging the increasing availability of 3D human shapes. This saves power and computation dramatically as the dictionary can be saved into a file and the 3D inversion problem degenerates from dictionary learning problem to a sparse code estimation task. We solve this problem by employing the convex relaxation algorithm originally proposed by Zhou *et al*. [2], which can be optimized efficiently by alternating direction method of multipliers.

## 2. Background

The package `AVFoundation.framework` and the delegate `UIImagePickerControllerDelegate` are used in this project for the access to the camera. Since matrix multiplication is the major computation in this project, we utilize `armadillo` and `Accelerate.framework` to accelerate linear algebra.

## 3. Approach

For the completeness of this report, we visit the idea and show some basic equations from the paper [2].

In order to utilize the proposed convex relaxation, Zhou *et al*. proposed a shape-space model stating that the 3D shape of a human pose can be represented by a few rotated shape basis sparsely:

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

in which $\mathbf{S} \in \mathbb{R}^{3\times N}$ is the 3D human pose matrix, $\mathbf{B}_i \in \mathbb{R}^{3\times N}$ is the i-th shape basis, $\mathbf{R}_i$ is the related rotation for i-th basis, and $c_i$ is the combination weight.

In this regard, estimating an unknown shape can be achieved by minimizing the following objective:

$$\begin{aligned} \min_{\mathbf{c},\mathbf{M}} \quad & \|\mathbf{c}\|_0 \\ \text{s.t.} \quad & \mathbf{W} = \sum_{i=1}^{k} \mathbf{M}_i \mathbf{B}_i, \\ & \mathbf{M}_i \mathbf{M}_i^T = \mathbf{I}_2, \end{aligned} \tag{2}$$

where, by assuming weak perspective projection,

$$\mathbf{M}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} c_i \mathbf{R}_i. \tag{3}$$

Zhou *et al*. demonstrated in the paper [2] that the orthogonal constraint ($\mathbf{M}_i\mathbf{M}_i^T = \mathbf{I}_2$) can be relaxed to spectral-norm minimization such that the proposed objective can be equivalently expressed as

$$\min_{\mathbf{M}} \frac{1}{2}\left\|\mathbf{W} - \sum_{i=1}^{k}\mathbf{M}_i\mathbf{B}_i\right\|_F^2 + \lambda\sum_{i=1}^{k}\|\mathbf{M}_i\|_2. \qquad (4)$$

Due to the limited space, the alternating direction method of multipliers solution to the proposed objective is omitted here, which can be found in the paper [2]. We highly recommend readers to read the above paper.

## 4. Results

We show the performance of the developed iOS application using both images downloaded from Internet and taken by users in daily life. Figure 1 shows the estimated human pose for Messi and Figure 2 shows the estimated human pose for the author, where the photo is taken indoor using the build-in camera of iPad pro. One can see the developed application recovers the depth of each joints successfully.

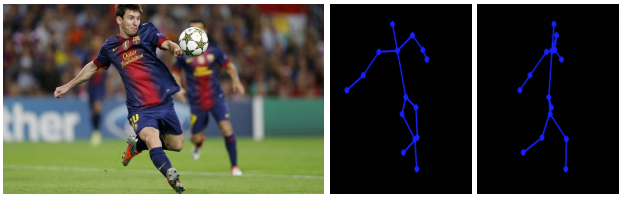

Figure 1. Estimated 3D human pose for the image of Messi.

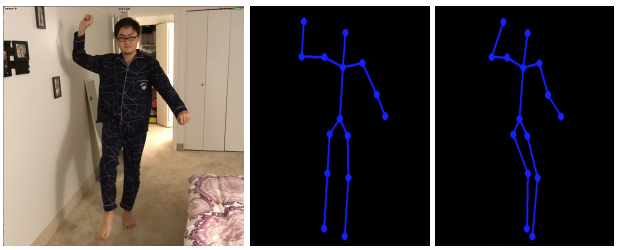

Figure 2. Estimated 3D human pose for the author.

## 5. Application

The application associated with this project is developed specifically for iPad pro. One can check the YouTube video[1] for learning instructions and view the results from more viewpoints. A GitHub page[2] has been setup where the application is released to publics for downloading.

[1] https://youtu.be/oikGjG9ExkQ

[2] https://github.com/kongchen1992/16623_final_project

## References


[1] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[2] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4447–4455, 2015.