3D Human Pose Estimation = 2D Pose Estimation + Matching

Ching-Hang Chen, Deva Ramanan
Robotics Institute, Carnegie Mellon University

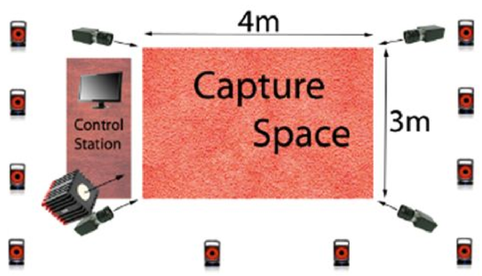
Brief Introduction:

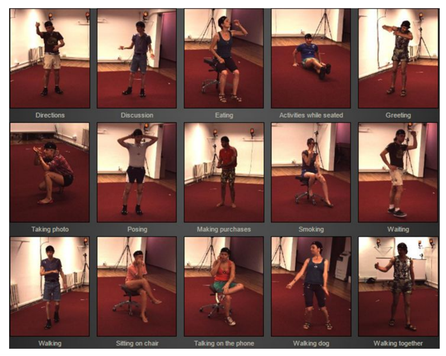
- We propose a simple and effective approach to predict 3D human pose by monocular RGB images.
- The proposed two-step framework leverages CNN based 2D pose estimation, and complete 3D prediction by a warping of nearest exemplar.
- Different matching metric were examined for extracting nearest exemplars.
- The approach could be applied to images in the wild.

Dataset:

Human3.6M

- A MoCap system capturing 15 action classes performed by 15 human subjects
- Each frame has corresponding 2D and 3D human pose annotations





Assumption

- Weak perspective
- Given 2D pose(landmarks), the 3D pose is independent of the input image:

$$p(\mathbf{X}, \mathrm{x}, I) = p(\mathbf{X}|\mathrm{x}, I) \cdot p(\mathrm{x}|I) \cdot p(I)$$

$$p(\mathbf{X}, \mathrm{x}, I) = \underbrace{p(\mathbf{X}|\mathrm{x})}_{\text{NN}} \cdot \underbrace{p(\mathrm{x}|I)}_{\text{CNN}} \cdot p(I)$$

Approach

- First use CNN-based approach to estimate 2D pose (Convolutional Pose Machines(CPM)), and then nearest exemplar from 3D pose library is found by Euclidean distance.

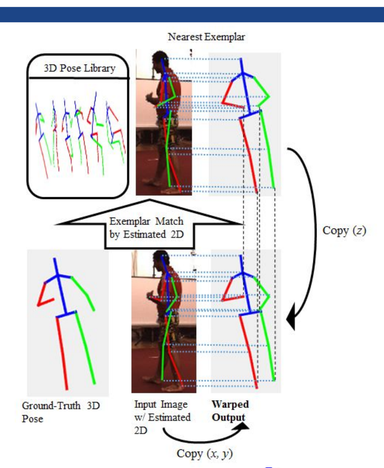
$$P(\mathbf{X} = i|\mathbf{x}) \propto e^{-\frac{1}{\sigma^2}||M_i(\mathbf{X}_i) - \mathbf{x}||^2}$$

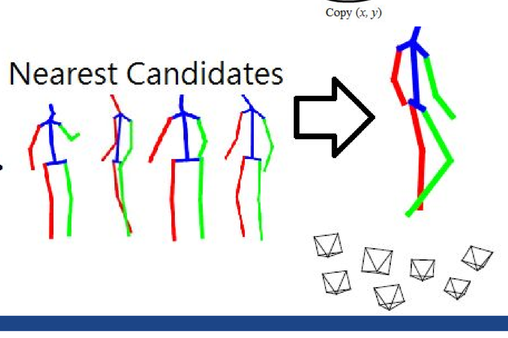
- The 3-rd dimension(depth) is predicted by the exemplar.
- We proposed to re-rank the nearest exemplars by camera resectioning

$$M_i^* = \underset{M}{\operatorname{argmin}} ||M(\mathbf{X}_i) - \mathbf{x}||^2$$

$$\min_{R,t} ||k[R\,|\,t]X - x|| \quad (1)$$

$$\min_{S,R,t} ||sRX + t - x|| \quad (2)$$

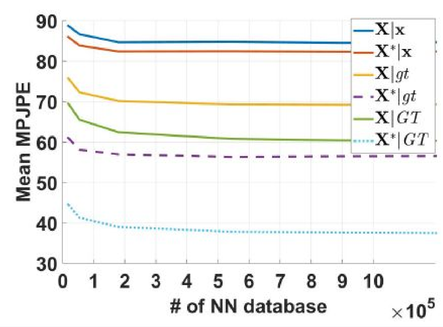




Experiments Setting

- Trainset: S1, S5, S6, S7, S8, S9(1.5M exemplars); Testset: S11
- Evaluation Metric: Mean Per Joint Position Error in mm (MPJPE)
- Evaluation Protocol A: MPJPE up to a rigid transformation
- Evaluation Protocol B: MPJPE aligning the reference (pelvis) joint
- Re-rank k exemplars, for k = 1, 10, 100

Trainset Size

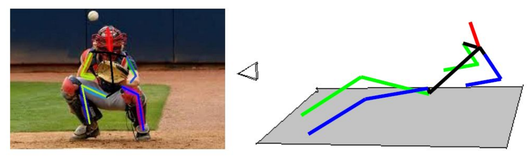


- The performance vs trainset size

$\mathbf{X}\|\mathbf{x}$	Unwarped exemplar queried with CNN 2D
$\mathbf{X}^*\|\mathbf{x}$	Warped exemplar queried with CNN 2D
$\mathbf{X}\|gt$	Unwarped exemplar queried with GT 2D
$\mathbf{X}^*\|gt$	Warped exemplar queried with GT 2D
$\mathbf{X}\|GT$	Unwarped exemplar queried with GT 3D
$\mathbf{X}^*\|GT$	Warped exemplar queried with GT 3D

Fail case

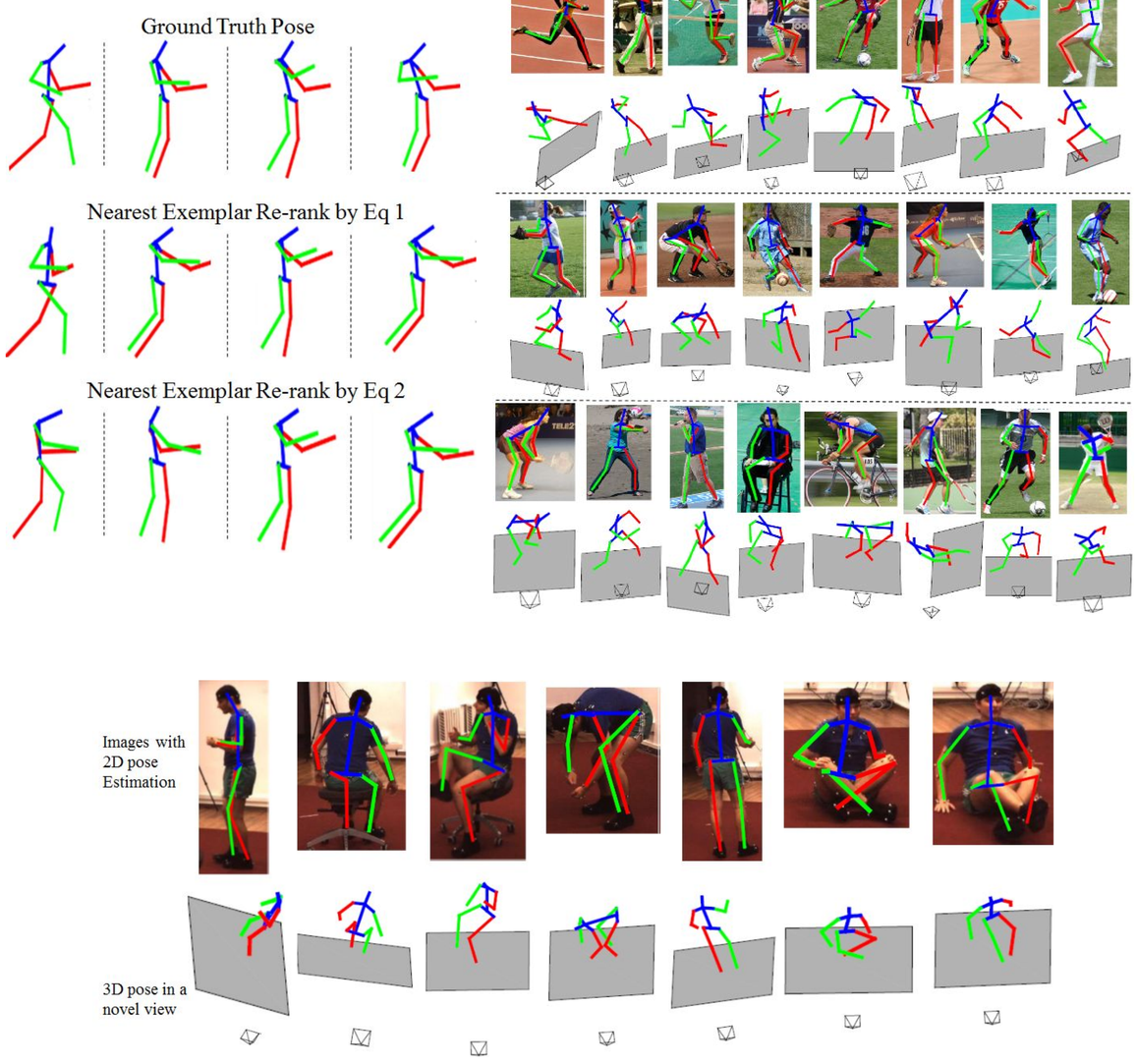
- When the assumption doesn't hold



Input image with the 2D pose estimation by CPM

The retrieved 3D pose using the 2D pose as query

Qualitative evaluation



Quantitative evaluation

Protocol A	k = 10	k = 100
Eq 1	86.32	89.83
Eq 2	**83.70**	84.10

Table 1: Comparison between Eq 1 and Eq 2 for different k under **Protocol A**

Protocol B	k = 10	k = 100
Eq 1	111.42	115.86
Eq 2	**109.00**	109.56

Table 2: Comparison between Eq 1 and Eq 2 for different k under **Protocol B**

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Yasin [1]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	**170.8**
Rogez [2]	-	-	-	-	-	-	-	-	-
k = 1	71.63	66.60	74.74	79.09	**70.05**	**67.56**	89.30	**90.74**	195.62
k = 10	**71.11**	**58.18**	**74.61**	**78.71**	70.23	67.59	**89.10**	91.29	195.86
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg	Median	-
Yasin [1]	108.2	142.5	86.9	92.1	165.7	102.0	108.3	-	-
Rogez [2]	-	-	-	-	-	-	88.1	-	-
k = 1	83.46	**93.26**	71.15	55.74	85.86	**62.51**	82.72	69.05	-
k = 10	**82.29**	93.58	**70.54**	**55.43**	**85.85**	63.38	**82.42**	**68.03**	-

Table 3: Comparison to [1], [2] by **Protocol A**. k = 10 performs better than k = 1, and the proposed approach outperforms recent state-of-the-art.

	Walk	Jog	Throw Catch	Gestures	Box	Avg.
Warped	64.46	69.88	59.99	67.89	79.22	68.29
Unwarped	90.17	95.27	82.74	88.82	103.85	92.17

Table 4: We evaluate a Human3.6M-trained model on HumanEva[3] (under **Protocol A**). To isolate the impact of 3D matching, we use ground-truth 2D keypoints. As a point of comparison, average error on Human3.6M test is 70.93 (unwarped) and 57.5 (warped)

Reference

[1] H. Yasin et al,. A dual source approach for 3d pose estimation from a single image., CVPR 2016.
[2] G. Rogez et al., Mocap-guided data augmentation for 3d pose estimation in the wild, NIPS, 2016.
[3] L. Sigal et al., Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV, 2010.