Seeing Invisible Poses: Estimating 3D Body Poses from Egocentric Video

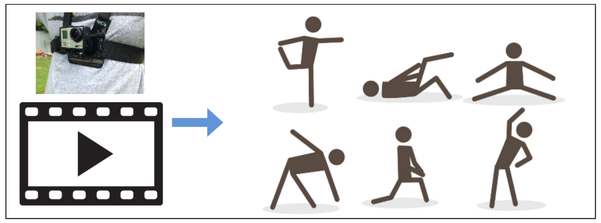
Hao Jiang
Boston College

Kristen Grauman
University of Texas at Austin

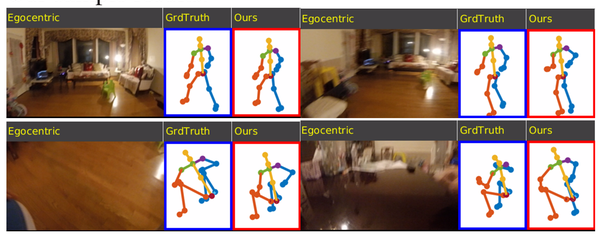
Introduction

We tackle a new problem of inferring 3D body poses from chest-mounted egocentric camera video.

Key challenges are:
(1) Body parts may not be visible at all.
(2) Scene is not limited to single environment.
(3) People have different motion styles.



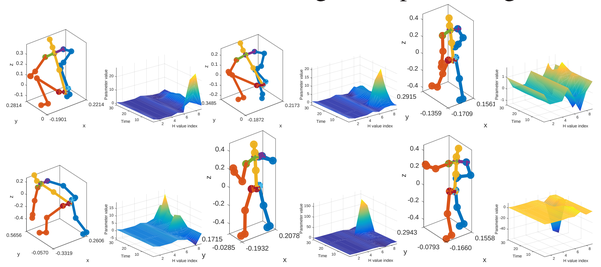
Our sample results:



We develop a learning approach that gives detailed 3D human poses and body part movements.

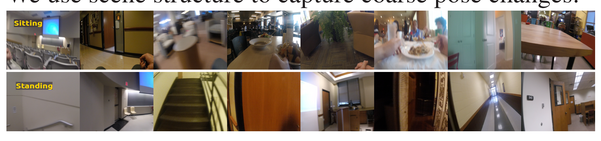
Instantaneous Pose Estimation

We use motion to reflect fine-grained pose changes:



Dynamic features: stack of homographies between successive frames in 1-second video to capture fine-grained pose cues.

We use scene structure to capture coarse pose changes:

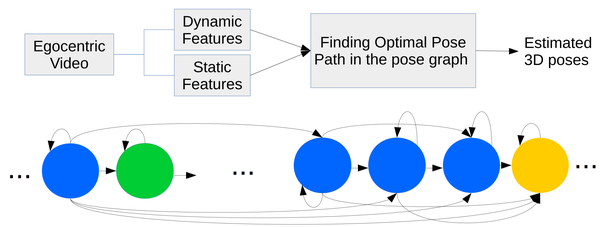


Static scene features: CNN classifier for sitting-like vs. standing-like.

Based on these two features, we initially classify each video frame into pose clusters.

Optimizing Pose Sequences

We propose a sequence-level joint optimization for the pose estimates, which overcomes flaws in instantaneous estimates.



$$\min_{\mathcal{X}}\{U(\mathcal{X})+T(\mathcal{X})+V(\mathcal{X})+S(\mathcal{X})\} \quad (1)$$

s.t. $\mathcal{X}$ represents poses drawn from exemplar sequence.

Unary cost U:
$U=\sum_{n=1..N,i\in P} e_{i,n}x_{i,n}$, where $e_{i,n}$ is determined by the motion and scene structure features and P is the pose sets in a long training sequence.

Step size term T (first order smoothness):
T enforces feasible and smooth pose transitions. $T=\sum_{i,j,n} w_{j,i}x_{j,n-1}x_{i,n}$, where $w_{j,i}=0$ if $i-j\leq 2, i\geq j$ and otherwise $w_{j,i}=\delta$ if pose i and j are in consecutive pose clusters, where δ is a positive constant penalty; if pose clusters are not consecutive $w_{j,i}=+\infty$.

The speed smoothness of the path V (second order smoothness):
V encourages uniform speed pose transitions.

$$V=\sum_{i,j,n} q(|s_{j,n-1}-(i-j)|)x_{j,n-1}x_{i,n}\ , \quad (2)$$

where q is a truncated linear function.

The stationary step penalty S in the path:
To prevent poses from staying the same for a long time, we introduce a penalty term S.

$$S=\sum_{i,j,n} r(u(j,n-1),i)x_{j,n-1}x_{i,n}\ , \quad (3)$$

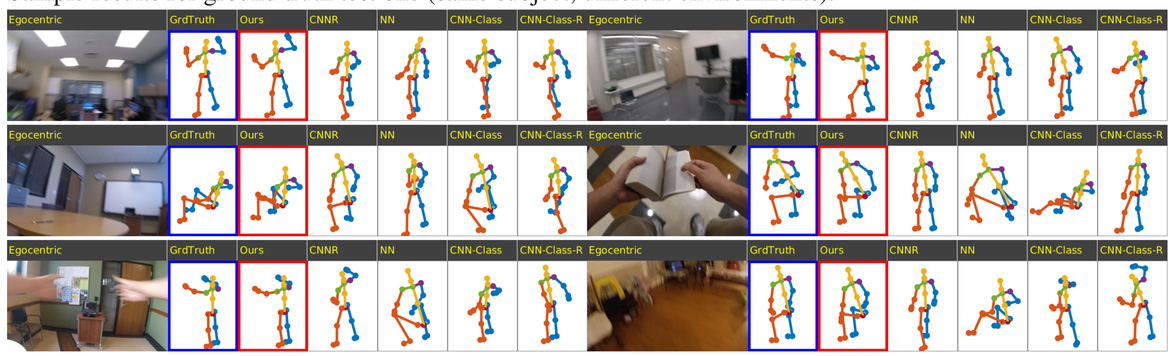
where $r(u(j,n-1),i)=0$ if $i\neq j$, otherwise $r(u(j,n-1),i)=t(u(j,n-1)+1)$, and $t(.)$ is a truncated linear function.

Dynamic programming: We show how to globally optimize the pose inference using dynamic programming. We further take advantage of the sparse property of the trellis to speed up the computation.

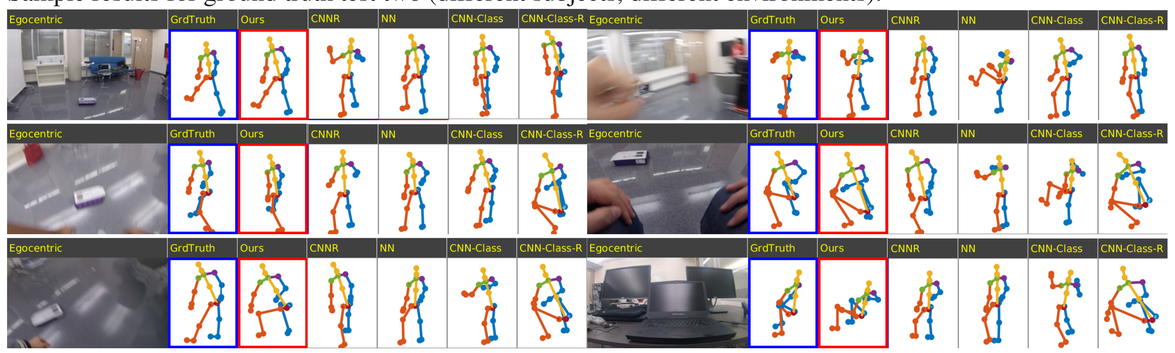
Ground Truth Data Test

We test our methods on about 40-minute long test video with ground truth.

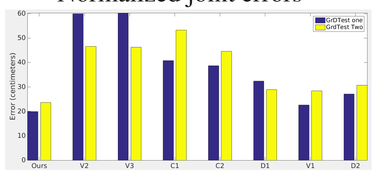
Sample results for ground truth test one (same subject, different environments):



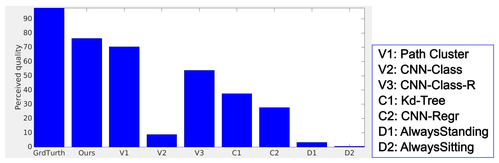
Sample results for ground truth test two (different subjects, different environments):



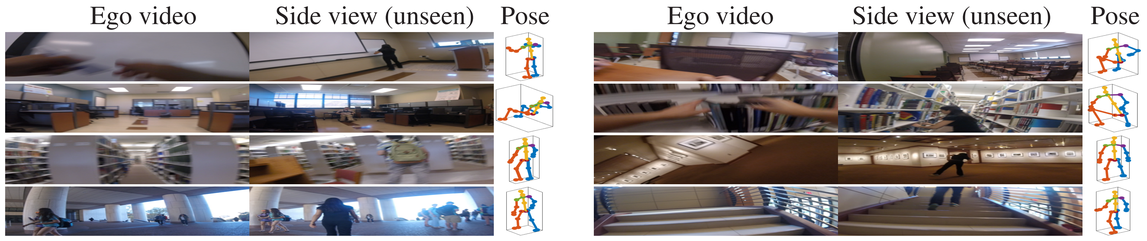
Normalized joint errors



Perceived quality scores



Unconstrained Data Test



Our method can reliably estimate 3D human poses in unconstrained settings.

Summary

Our proposed method infers 3D body poses from egocentric video even without seeing body parts. For more details see the project website: `www.cs.bc.edu/~hjiang/egopose/index.html`.