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

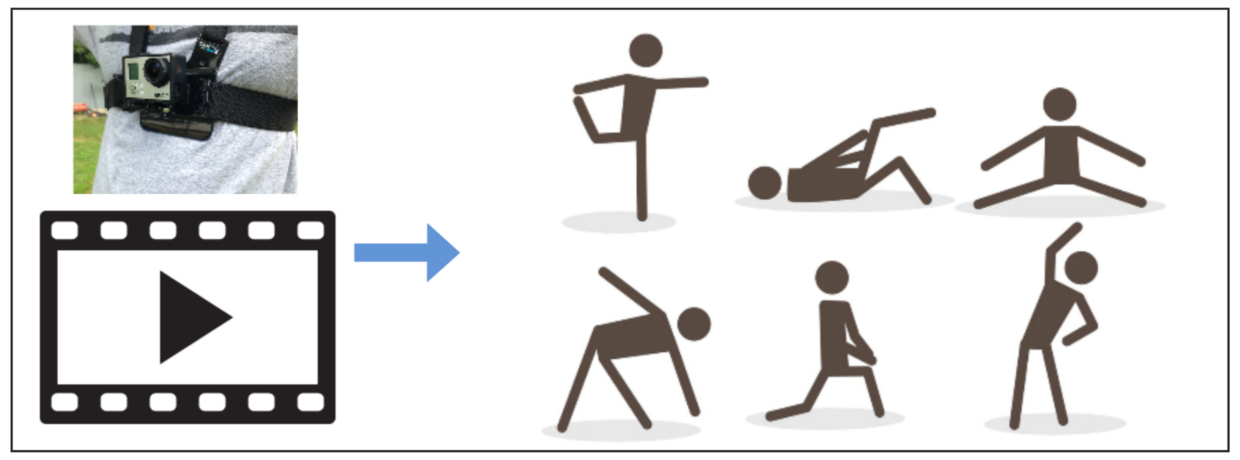
Hao Jiang — *Boston College*
Kristen Grauman — *University of Texas at Austin*

Introduction

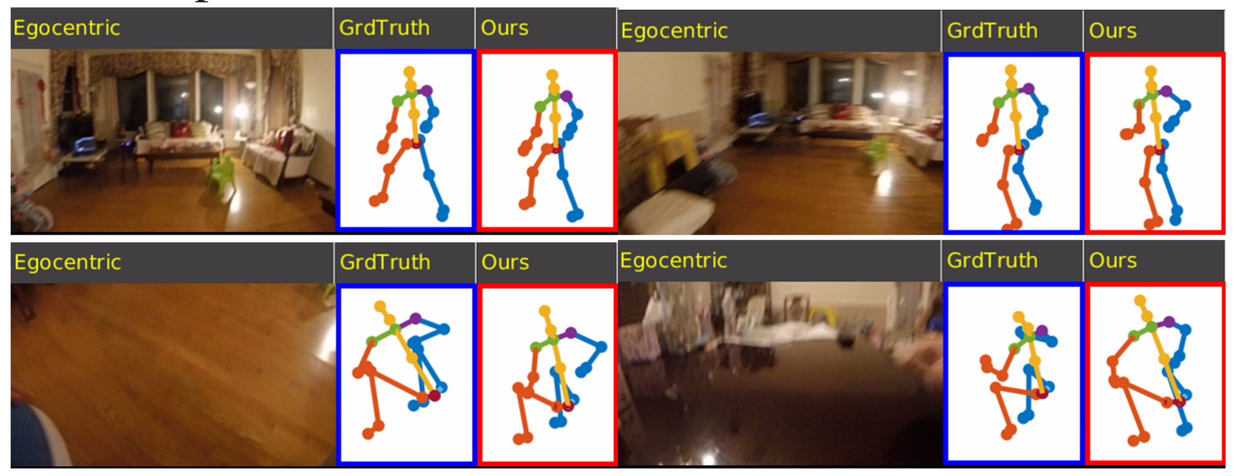
We tackle a new problem of inferring 3D body poses from chest-mounted egocentric camera.

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.



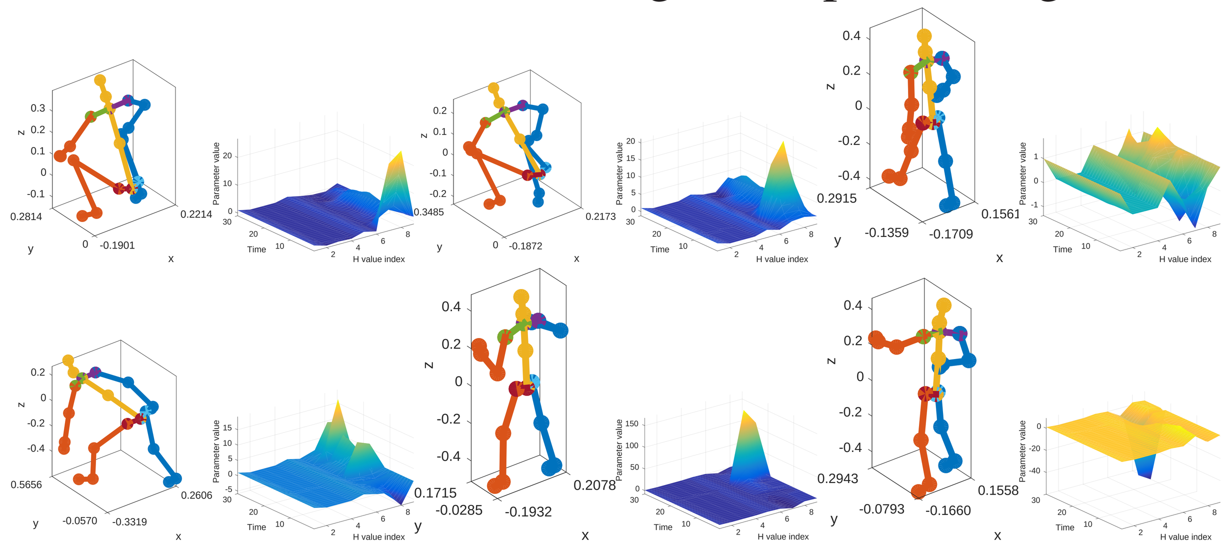
Our sample results:



Using a learning approach, our method gives detailed 3D human poses and body part movements.

Instantaneous Pose Estimation

We use motion to reflect fine-grained pose changes:



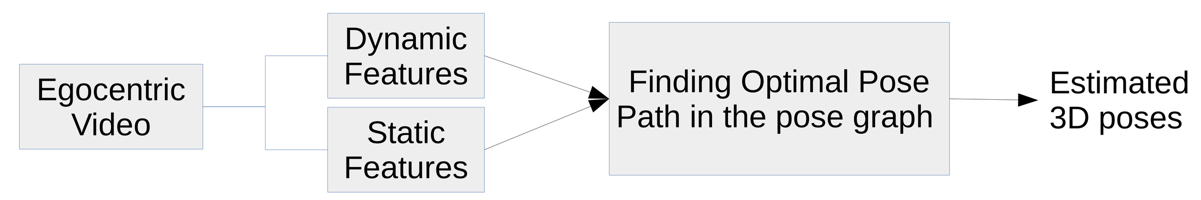
We use stack of homography between successive frames in one-second video as the dynamic feature. Similar poses often have similar dynamic features, and distinct poses have different motion patterns.

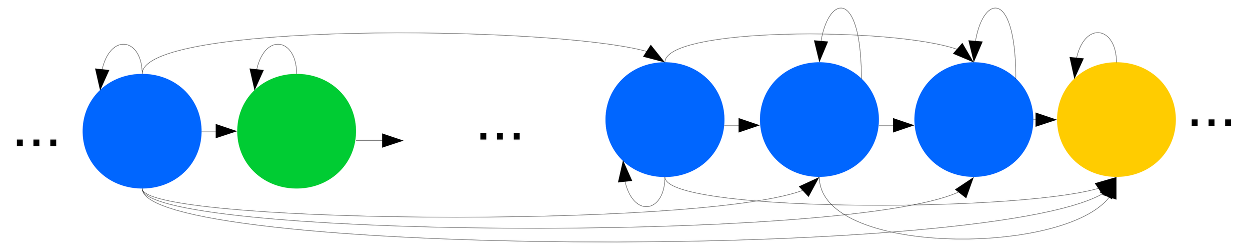
We also use scene structure as a clue:

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

Optimizing Pose Sequences

Instantaneous pose estimation is not sufficient. It has low temporal resolution and it is error prone due to lack of pose transition constraints.





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

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 exemplar 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 $i - j \le 2, i \ge j$ 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.

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

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.

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

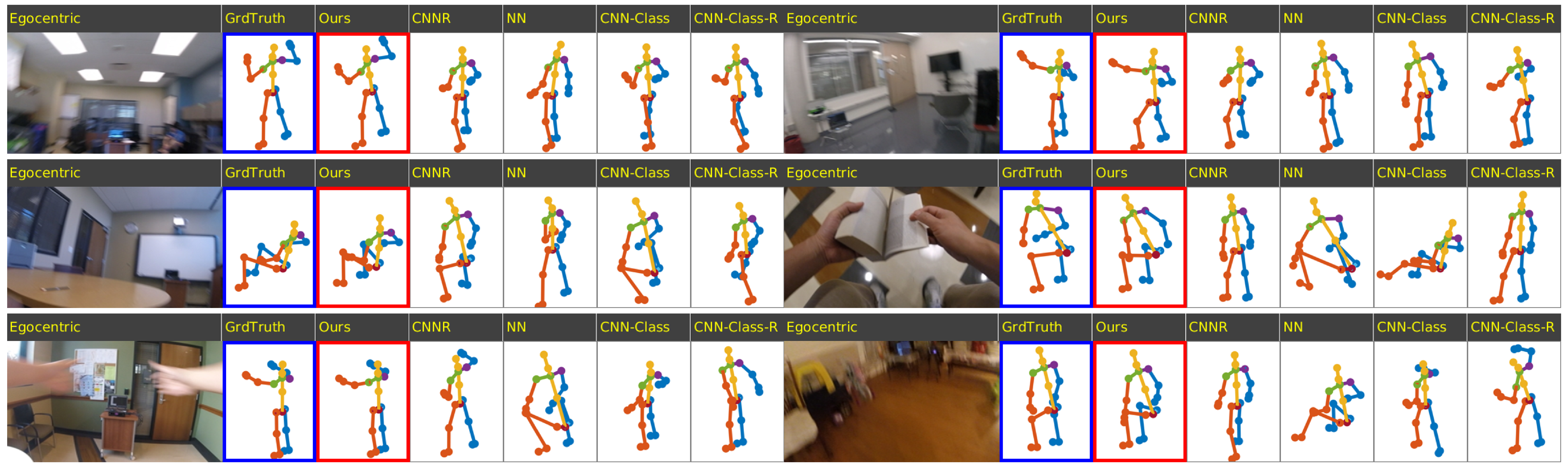
where $r(u(j,n-1),i) = 0$ if $i \ne j$, otherwise $r(u(j,n-1),i) = t(u(j,n-1)+1)$, and $t(.)$ is a truncated linear function.

Dynamic programming: we globally optimize the pose sequence estimation using dynamic programming. We further take advantage of the sparse property of the trellis to speed up the computation.

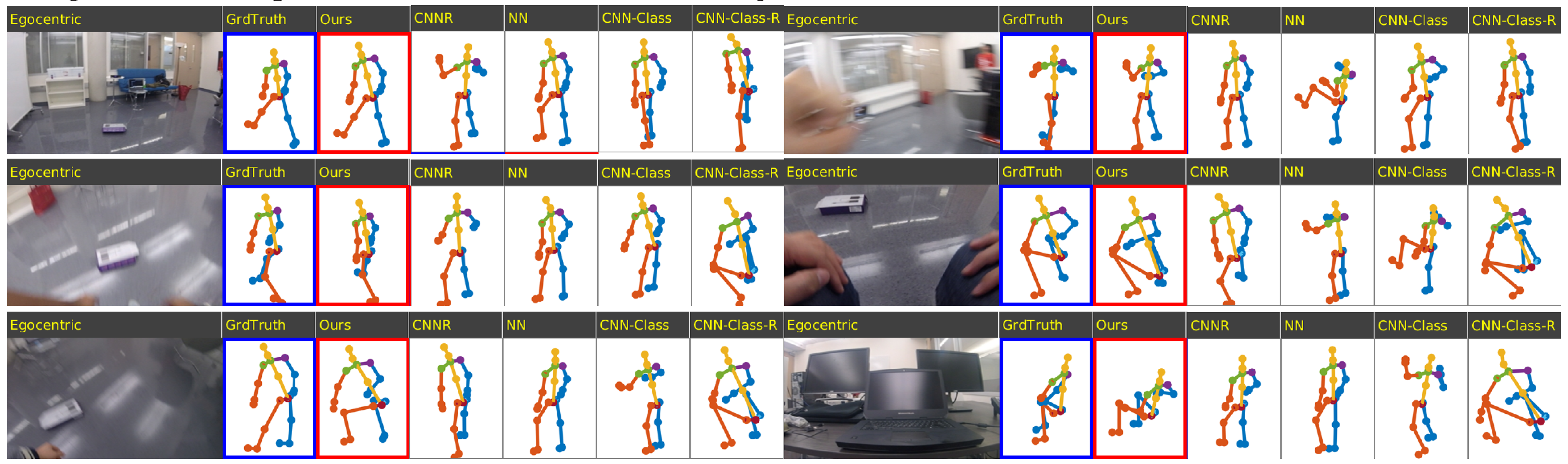
Ground Truth Data Test

We test our methods on about 40-minute long ground truth video. The comparison results are as follows:

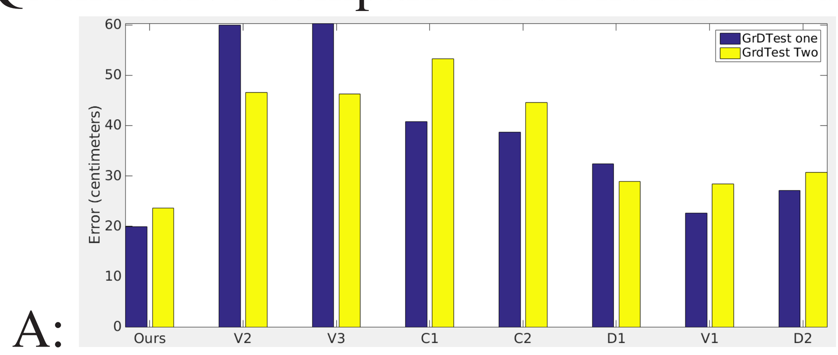
Sample results for ground truth test one (same subject, different environments):

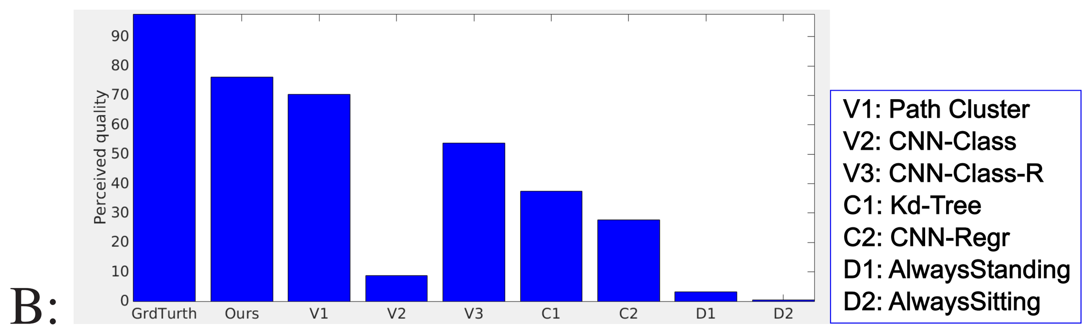


Sample results for ground truth test two (different subjects, different environments):

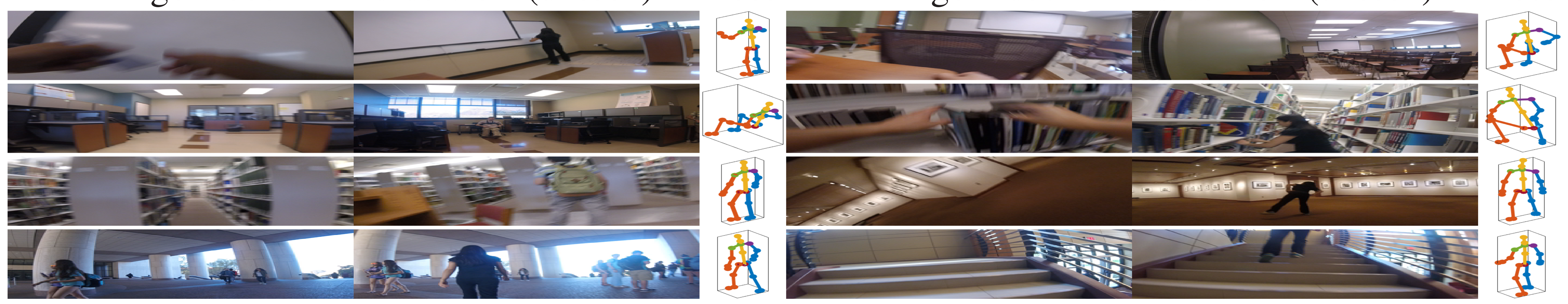


Quantitative comparison on normalized joint errors (A) and perceived quality scores (B):

A: 

B: 

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`.