Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, Kostas Daniilidis

University of Pennsylvania

CVPR July 21-26 HONOLULU 2017

Training Code
Testing Code
tinyurl.com/PoseVolumetric

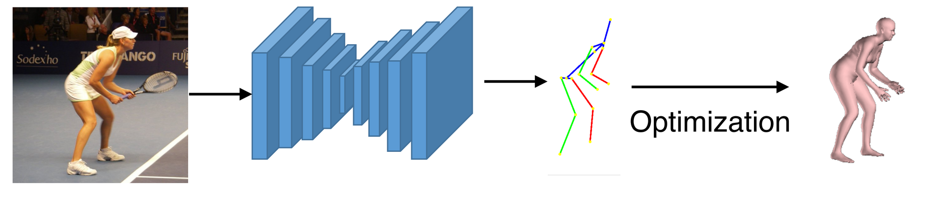
Goal: Estimate 3D human pose from a single color image

Two paradigms dominate this problem.
Reconstruction and discriminative approaches.

Two-step Reconstruction Approaches

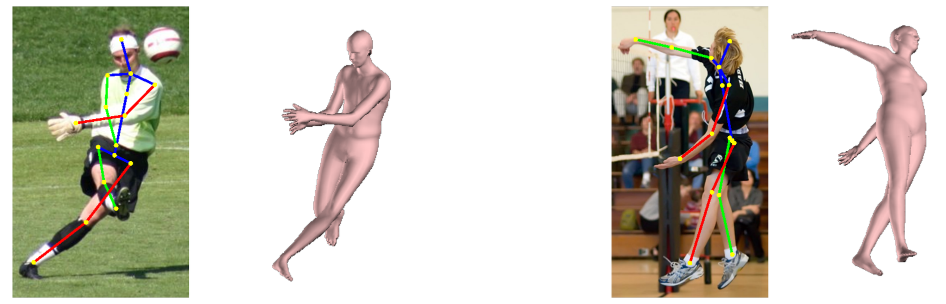
Approach

2D pose estimation + optimization lifting 2D-to-3D



Limitation

Reconstruction Ambiguity!

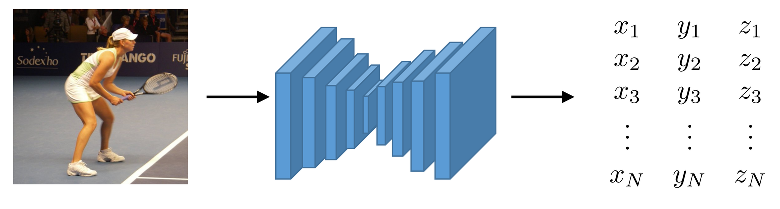


Discriminative Approaches

e.g.: coordinate regression

Approach

Estimate the 3D pose directly from the image.



Limitation

Problem is highly non-linear.
Mapping from images to 3D coordinates is hard to learn.
Underperforms compared to two-step approaches.

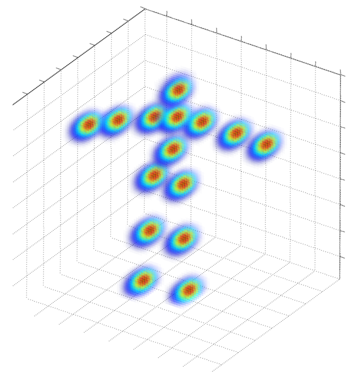
This work

We attempt to bridge the gap!

We introduce the volumetric representation for 3D human pose.

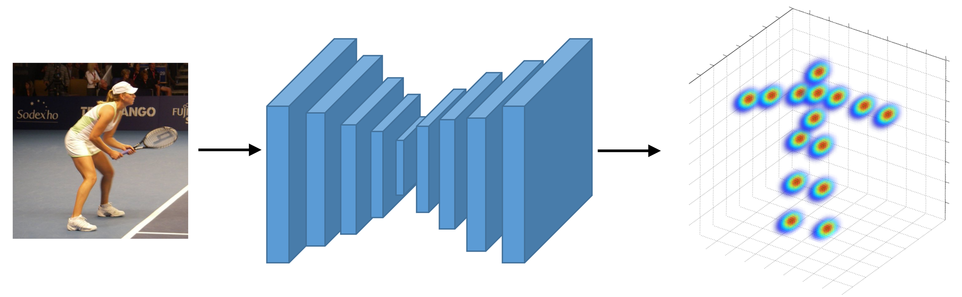
We use a coarse-to-fine prediction scheme to deal with the excessive dimensionality.

We employ a decoupled architecture for 3D human pose estimation "in-the-wild".



We achieve more than **30% relative error reduction** for standard benchmarks!

Volumetric representation for 3D human pose



We cast the problem as 3D keypoint localization.

We regress 3D heat maps of dimensions 64x64x64 for each joint.

Major advantages

ConvNets can naturally map from 2D images to 3D volumes.

The mapping can be achieved with a Fully Convolutional Network.

Rich output (64x64x64 for each joint). Useful for other tasks/postprocessing.

Average Error Human3.6M (mm)

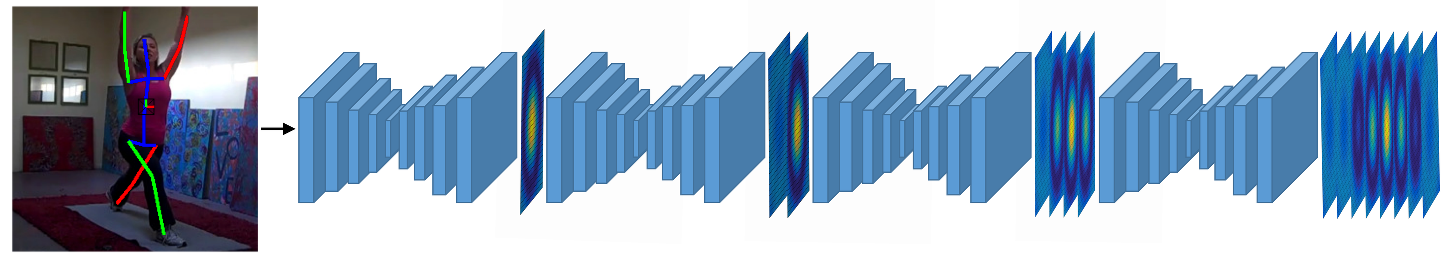
Coordinate Regression	112.41
Volume Regression (depth = 32)	92.23
Volume Regression (depth = 64) 162	85.82

Coarse-to-Fine prediction

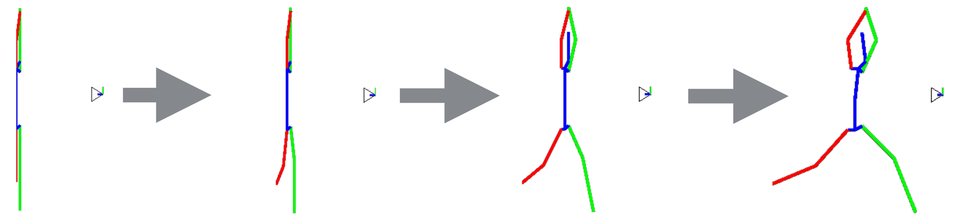
Iterative estimation offers diminishing returns because of the excessive dimensionality of our representation.

We do it in a coarse-to-fine way!

The resolution of the supervision volume increases gradually for the most challenging z-dimension.



z-dimension:



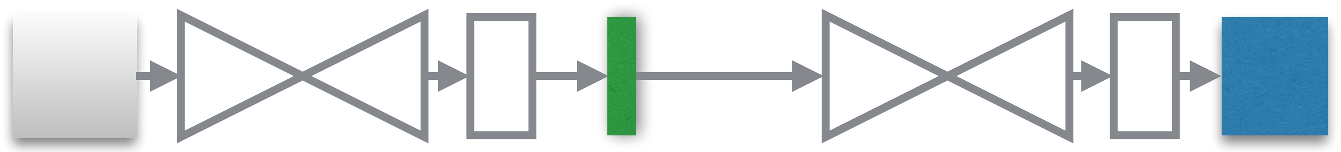
Average Error Human3.6M (mm)

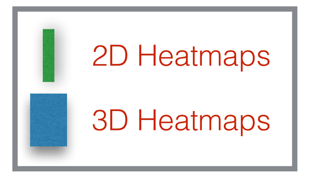
	Naive Stacking	Coarse-to-Fine
Two Hourglasses	80.14	69.77
Three Hourglasses	78.17	68.49
Four Hourglasses	75.06	64.76

Versatility of volumetric representation

Regress 3D heatmaps using 2D heatmaps as input.

Decoupled

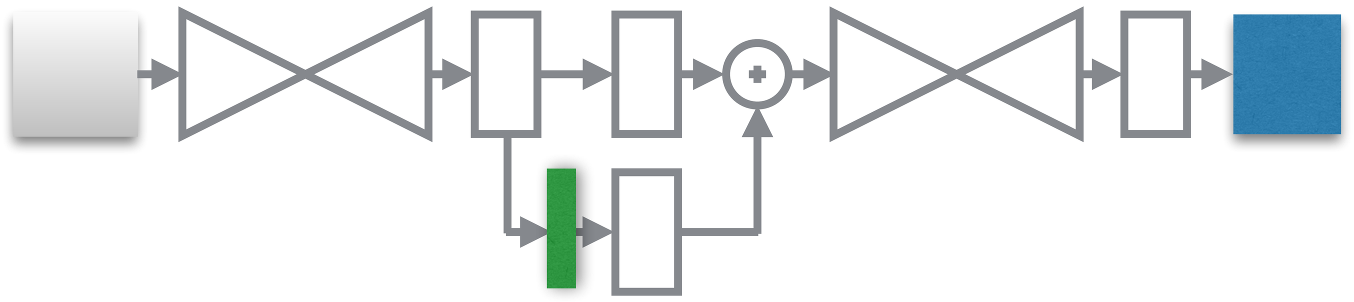




+ Allows us to train independently for the 2D and the 3D task.
+ We present compelling results for in-the-wild images.

- Uses only 2D joint locations and discards additional image evidence.
- When 2D estimates are wrong, 3D prediction can be lead astray.
- Underperforms compared to end-to-end approach.

Coarse-to-Fine

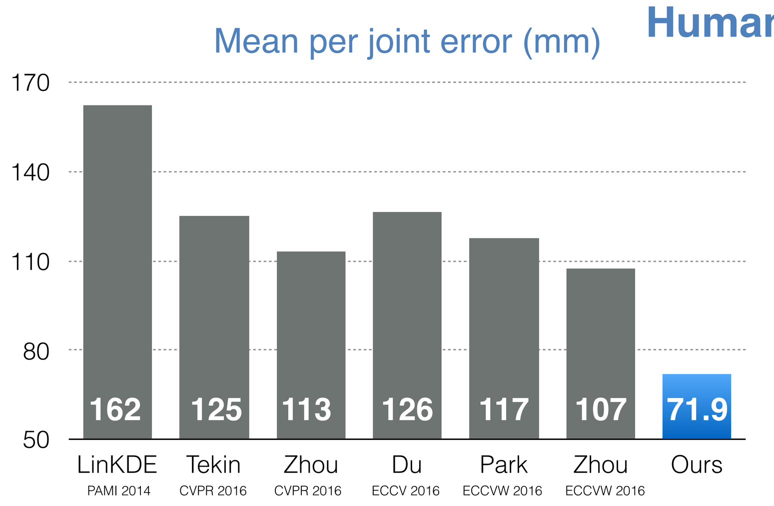


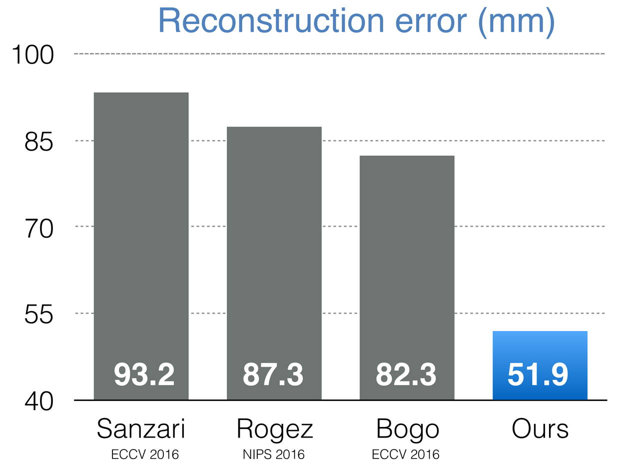
Average Error Human3.6M (mm)

Decoupled	Coarse-to-Fine
78.10	69.77

Quantitative results

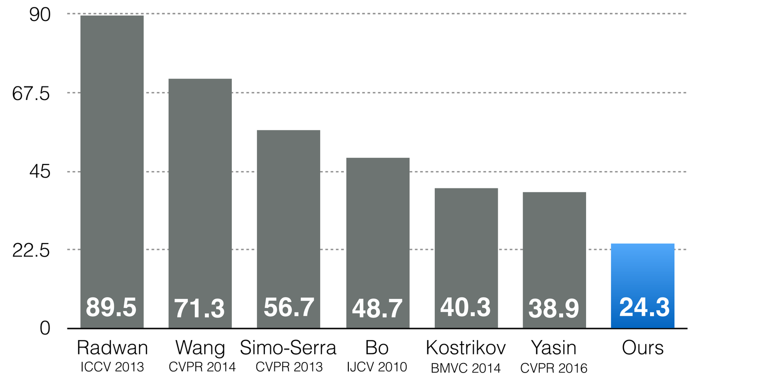
Human3.6M





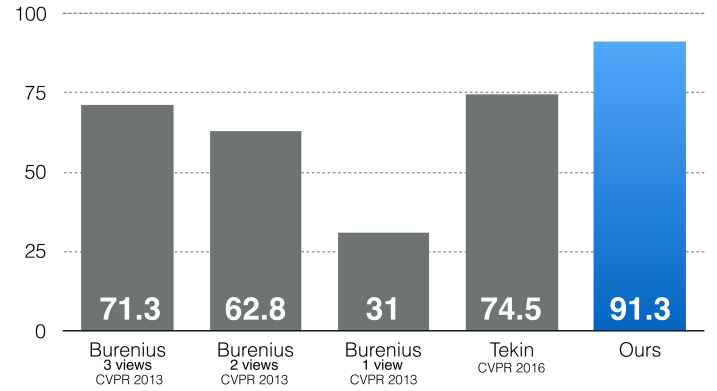
HumanEva

Reconstruction error (mm)



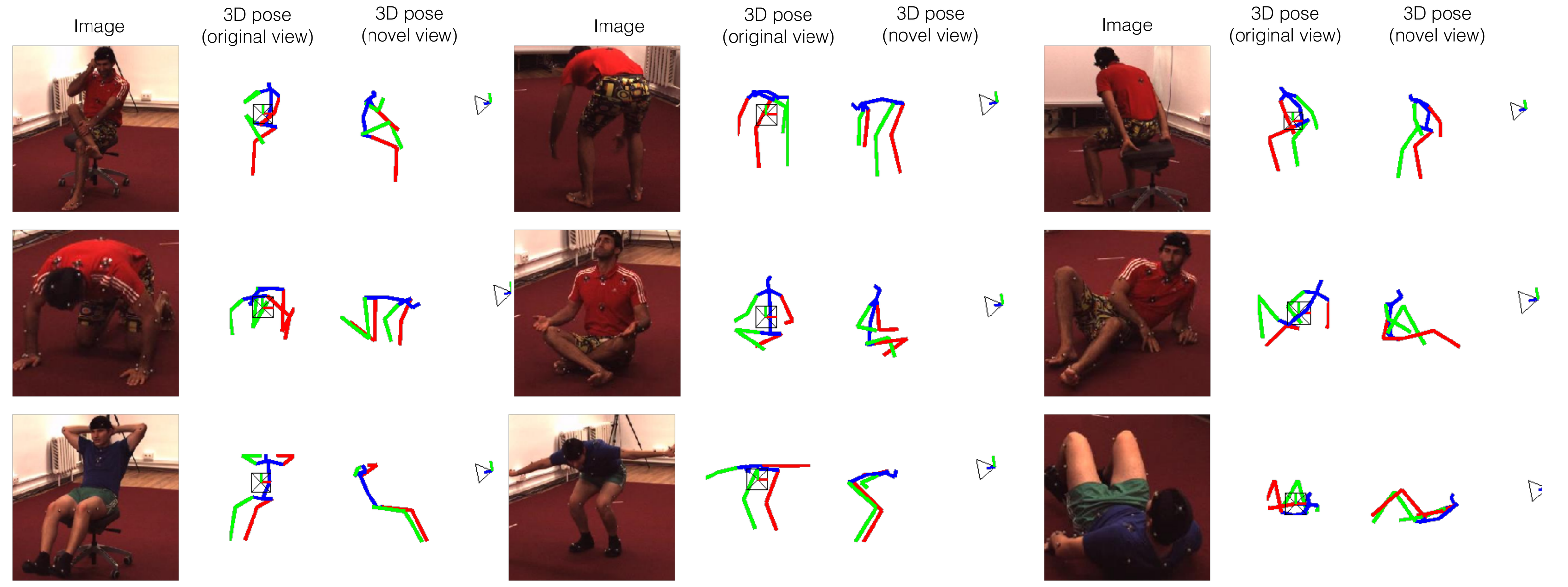
KTH Football

3D PCP

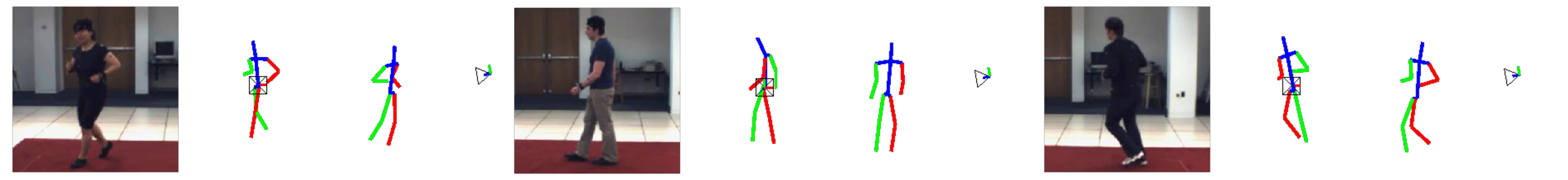


Qualitative results

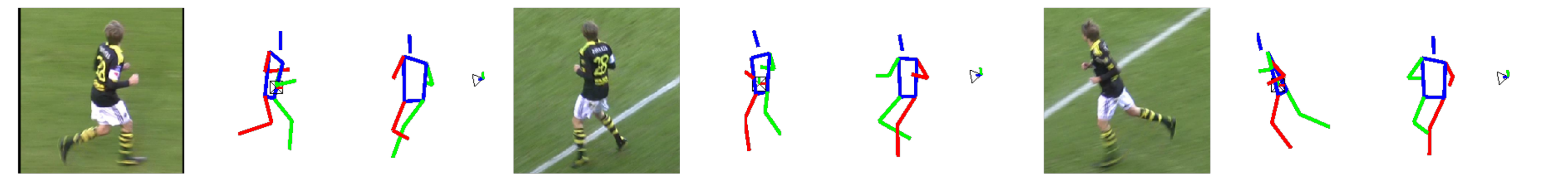
Human3.6M



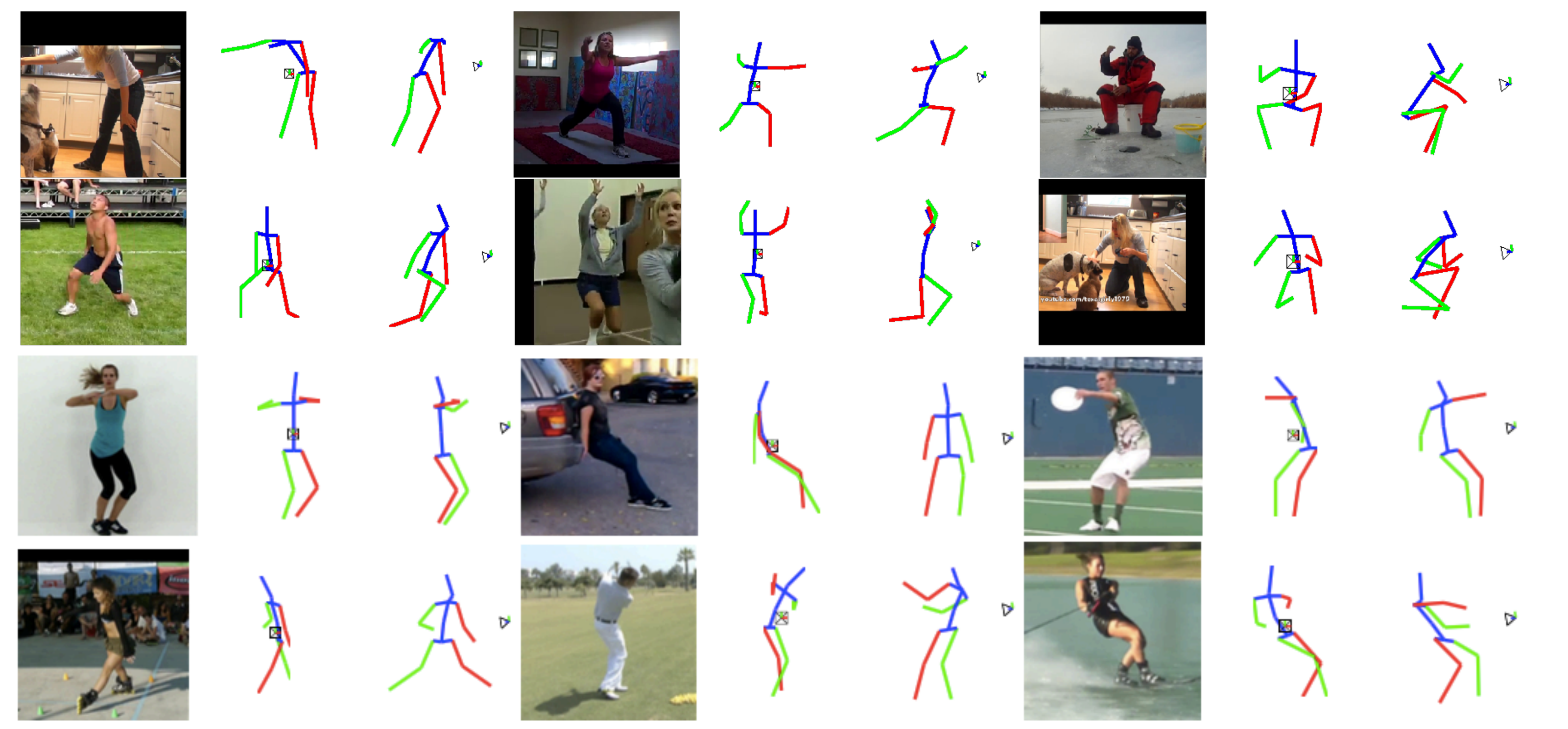
HumanEva



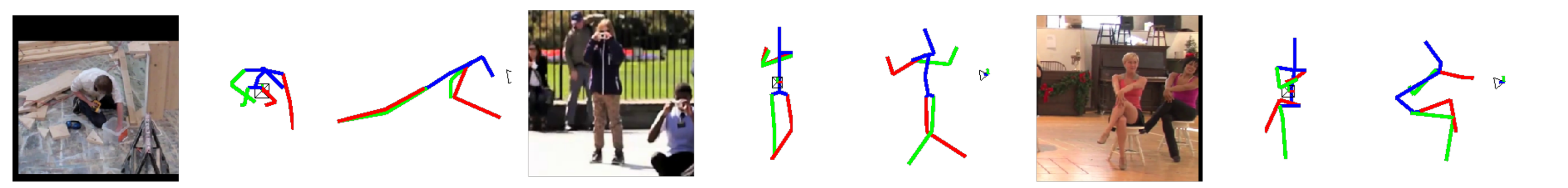
KTH Football



MPII



Failure cases



Decoupled vs Coarse-to-Fine