Harvesting Multiple Views for Marker-less 3D Human Pose Annotations

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

University of Pennsylvania

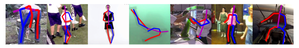
CVPR July 21-26 2017 Honolulu

Project Page
Testing Code
tinyurl.com/PoseHarvesting

Motivation

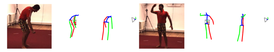
ConvNets have achieved impressive results on large scale human pose estimation benchmarks.

2D human pose estimation
MPII & FLIC



Multi-person pose estimation
MPII Multi-Person & CoCo Keypoints

3D human pose estimation
Human3.6M & HumanEva



But, ground truth data is not always readily available!

Some tasks live in the small-data regime.

3D human pose estimation "in-the-wild"

Limitation 1

Human3.6M HumanEva

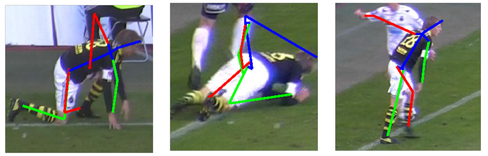
MoCap systems for capturing ground truth work only under constrained settings.

Limitation 2

LSP H3D

Humans cannot annotate metric 3D information.

"Personalizing" 2D human pose

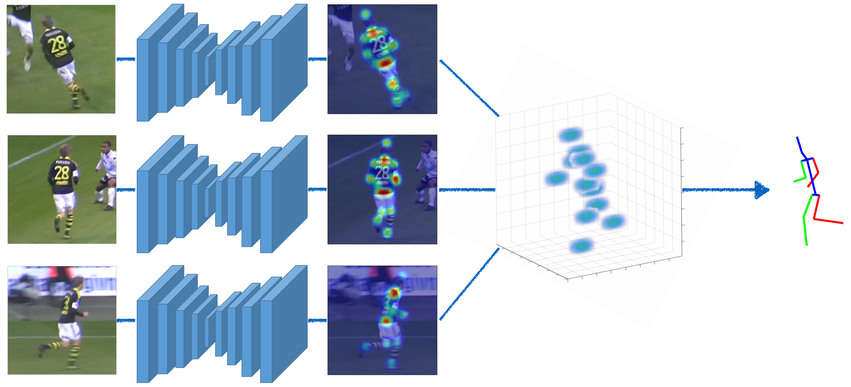


2D pose detectors are still not perfect out-of-the-box.

Can we automatically refine a generic ConvNet for a specific task?

How can multi-view geometry help us?

We propose to produce automatic 3D human pose annotations by harvesting multiple camera views of a scene!

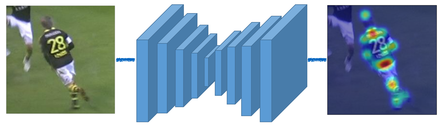


Multi-view Setup | Generic 2D pose ConvNet | Heatmaps | 3D Pictorial Structures | 3D Annotations

Step 1

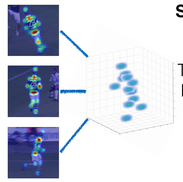
The input is a set of videos from a calibrated multi-view setup.

Step 2



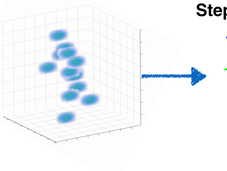
A generic ConvNet produces 2D pose estimates in the form of heatmaps for each view.

Step 3



The single-view estimates are back projected in a common discretized 3D space.

Step 4



A 3D Pictorial Structures model estimates the globally optimal 3D pose.

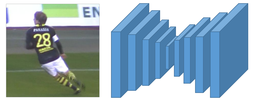
$$p(\mathbf{S}|\mathbf{I}) \propto \prod_{k=1}^{M}\prod_{i=1}^{N} p(\mathbf{I}_k|\pi_k(\mathbf{s}_i)) \prod_{(i,j)\in\mathcal{E}} p(\mathbf{s}_i,\mathbf{s}_j)$$

Posterior | Data likelihood | Pairwise prior

Annotation Leveraging

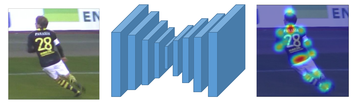
The harvested 3D pose estimates can be used as high quality annotations for human pose estimation tasks.

3D Annotations



We train a ConvNet that takes a single color image as input, and predicts the 3D pose.

"Personalized" 2D Annotations

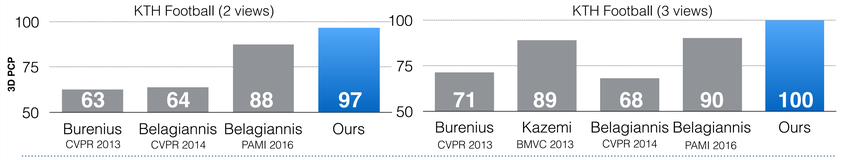


We refine a generic ConvNet for 2D pose by using the automatic 3D annotations projected to the 2D image.

Results

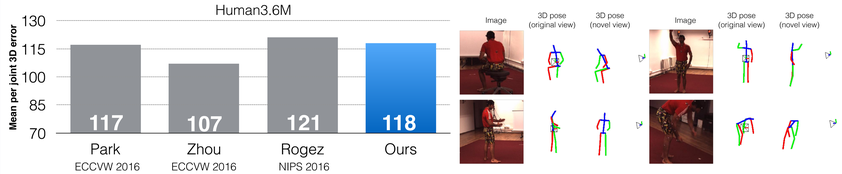
Multi-view pose estimation

State-of-the-art **using only generic 2D pose detector - no retraining.**



Single-view 3D human pose estimation

On par with the state-of-the-art **without using 3D ground truth for training.**



Automatic refinement of generic 2D pose detector

Consistent benefit over all body parts