Human Pose Estimation

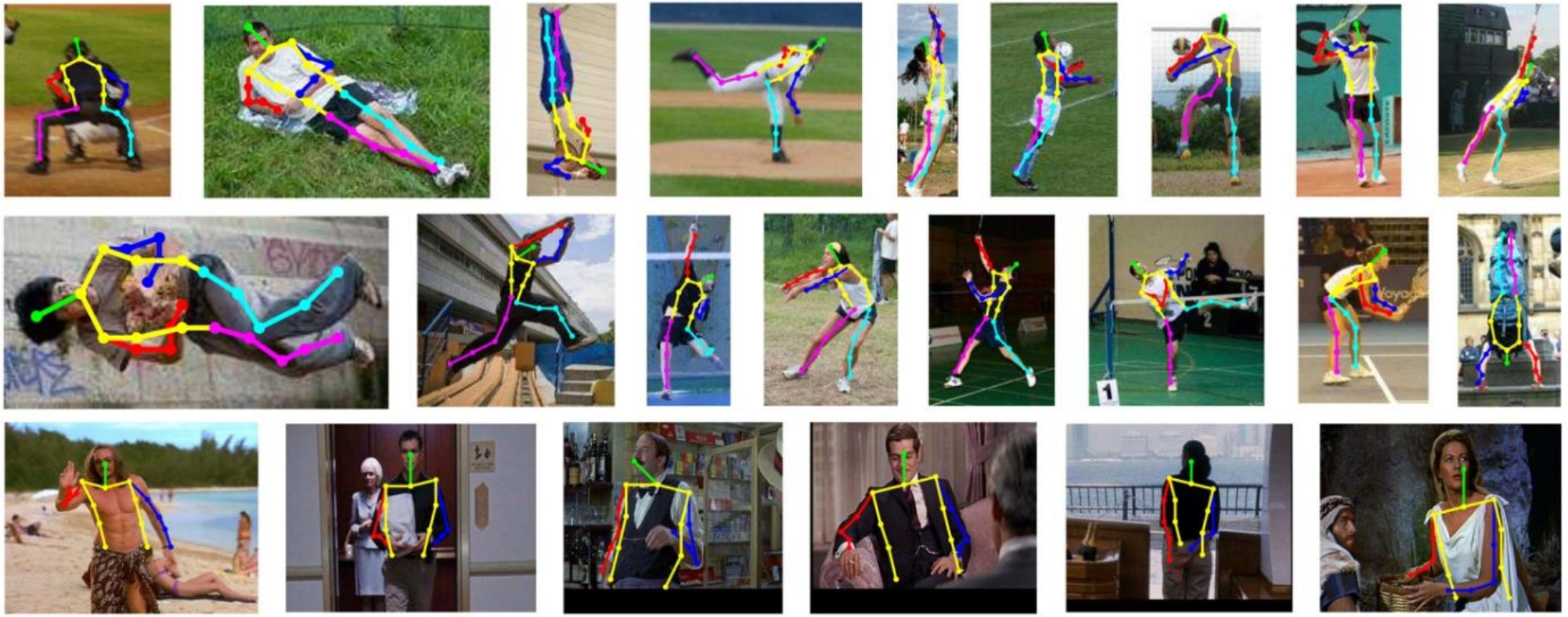
Assignment 2020

Computer Vision

Name	: Satyajit Ghana
USN	: 17ETCS002159
Course Code	: CSE308A
Department	: Computer Science

Introduction to HPE

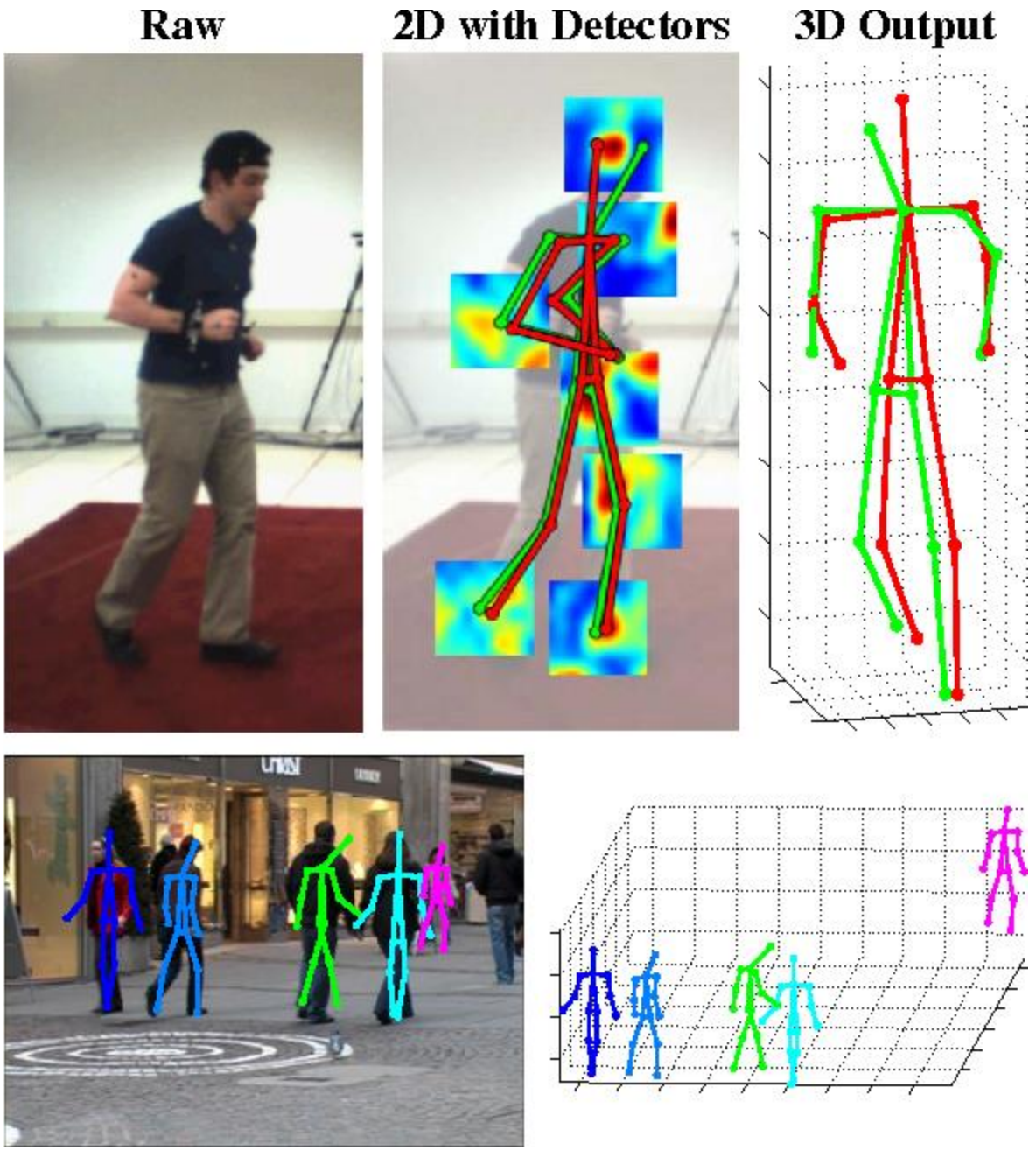
Refers to the task of recognizing the human body landmarks (Head, Shoulder, Wrist, Hip, Knee, Ankle) from a single monocular image.



But what are its Applications?

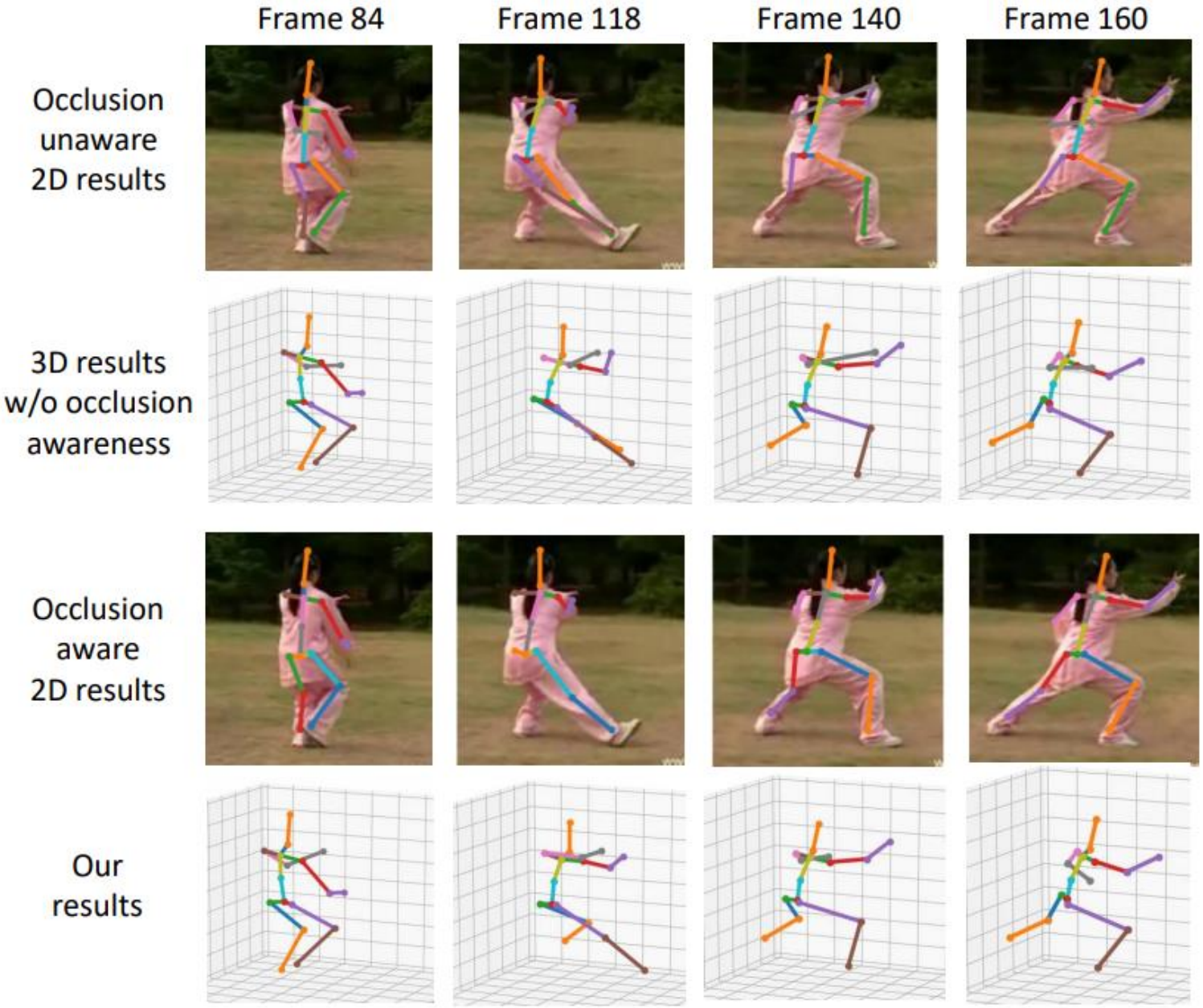
- Activity Recognition
- Human Tracking
- Movies and Animation Sequencing
- Virtual Reality
- Human-Computer Interaction
- Medical Surgery Assistance
- Military Training Assistant
- Self-Driving Cars
- Motion Analysis

*Basically a precursor to most computer vision tasks



Challenges in HPE

- Flexible body configuration indicates complex independent joints and degree-of-freedom limbs, which may cause self-occlusion or rare-complex poses
- Diverse body appearance includes different clothing and self-similar parts
- Complex environment may cause foreground occlusion or similar parts from nearby persons, various viewing angles, and truncation in the camera view

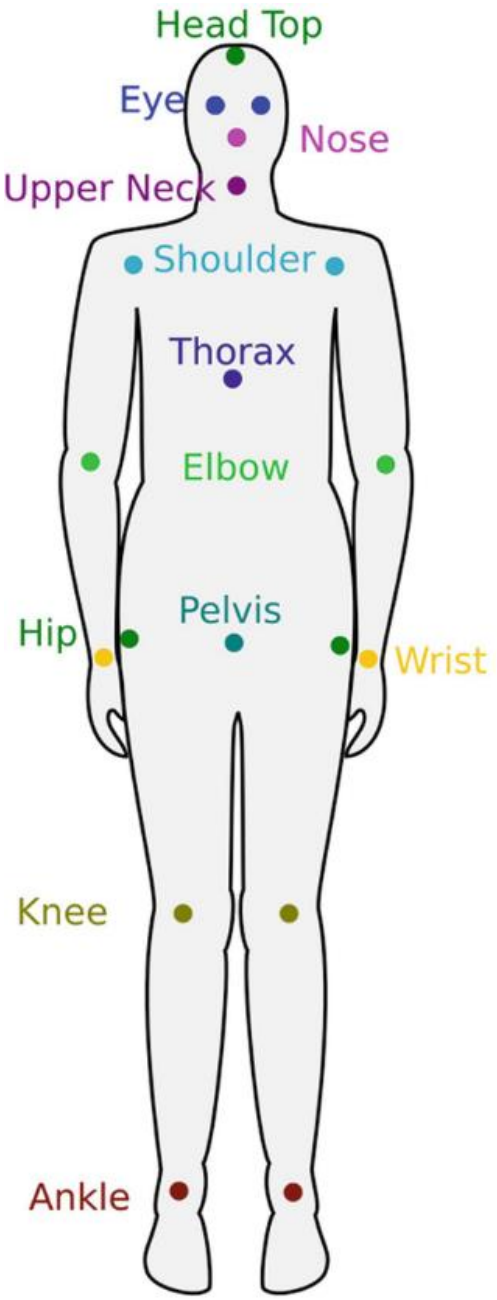


Methods for HPE

- Generative vs Discriminative
- Top-Down vs Bottom-Up
- Regression vs Detection
- One-Stage vs Multi-Stage

Dataset

- MPII is the state-of-the-art benchmark for evaluation of articulated human pose estimation.
- The dataset includes 25K images containing over 40K people with annotated body joints.
- Each image was extracted from a YouTube video and provided with preceding and annotated frames



MPII

Ankle	Upper Neck
Knee	Head Top
Hip	Wrist
Pelvis	Elbow
Thorax	Shoulder

LSP

Ankle	Wrist
Knee	Elbow
Hip	Shoulder
Upper Neck	
Head Top	

Synch

Thorax	Wrist
Upper Neck	Elbow
Head Top	Shoulder

BBC

Head Top	Elbow
Wrist	Shoulder

Flic

Hip	Shoulder
Wrist	Eye
Elbow	Nose

Comparison of SOTA Models

Methods	Backbone	Input size	Highlights	PCKh (%)
Regression-based				
(Toshev and Szegedy, 2014)	AlexNet	220×220	Direct regression, multi-stage refinement	-
(Carreira et al., 2016)	GoogleNet	224×224	Iterative error feedback refinement from initial pose.	81.3
(Sun et al., 2017)	ResNet-50	224×224	Bone based representation as additional constraint, general for both 2D/3D HPE	86.4
(Luvizon et al., 2017)	Inception-v4+ Hourglass	256×256	Multi-stage architecture, proposed soft-argmax function to convert heatmaps into joint locations	91.2

Detection-based				
(Tompson et al., 2014)	AlexNet	320×240	Heatmap representation, multi-scale input, MRF-like Spatial-Model	79.6
(Yang et al., 2016)	VGG	112×112	Jointly learning DCNNs with deformable mixture of parts models	-
(Newell et al., 2016)	Hourglass	256×256	Proposed stacked Hourglass architecture with intermediate supervision.	90.9
(Wei et al., 2016)	CPM	368×368	Proposed Convolutional Pose Machines (CPM) with intermediate input and supervision, learn spatial correlations among body parts	88.5
(Chu et al., 2017)	Hourglass	256×256	Multi-resolution attention maps from multi-scale features, proposed micro hourglass residual units to increase the receptive field	91.5
(Yang et al., 2017)	Hourglass	256×256	Proposed Pyramid Residual Module (PRM) learns filters for input features with different resolutions	92.0
(Chen et al., 2017)	conv-deconv	256×256	GAN, stacked conv-deconv architecture, multi-task for pose and occlusion, two discriminators for distinguishing whether the pose is 'real' and the confidence is strong	91.9
(Peng et al., 2018)	Hourglass	256×256	GAN, proposed augmentation network to generate data augmentations without looking for more data	91.5
(Ke et al., 2018)	Hourglass	256×256	Improved Hourglass network with multi-scale intermediate supervision, multi-scale feature combination, structure-aware loss and data augmentation of joints masking	92.1
(Tang et al., 2018a)	Hourglass	256×256	Compositional model, hierarchical representation of body parts for intermediate supervision	92.3
(Sun et al., 2019)	HRNet	256×256	high-resolution representations of features across the whole network, multi-scale fusion.	92.3
(Tang and Wu, 2019)	Hourglass	256×256	data-driven joint grouping, proposed part-based branching network (PBN) to learn representations specific to each part group.	92.7
(Zhihui et.al, 2019)	Hourglass-Multistage	256x256	Cascaded ResNet-50 and ResNet-101, and multi stage network	93.9
(Bruno and Andreas, 2020)	Hourglass	256x256	WASP Module, LSTM like architecture for videos	92.7
(Adrian et al, 2020)	Hourglass	256x256	Toward fast and accurate human pose estimation via soft-gated skip connections	94.1

So what are the trends ?

- There are lot more Detection-based models than Regression based models.
- Accuracy of Detection based models is way more that Regression based.
- Hourglass models are heavily used
- ResNet is preferred as the backbone architecture for most of the models

Why are Detection models more popular?

- Direct regression learning of only one single point is a difficulty since it is a highly nonlinear problem and lacks robustness.
- Toshev and Szegedy (2014) firstly attempted to train an AlexNet-like network to learn join coordinated from full images in a straightforward manner without using any body model or part detectors.

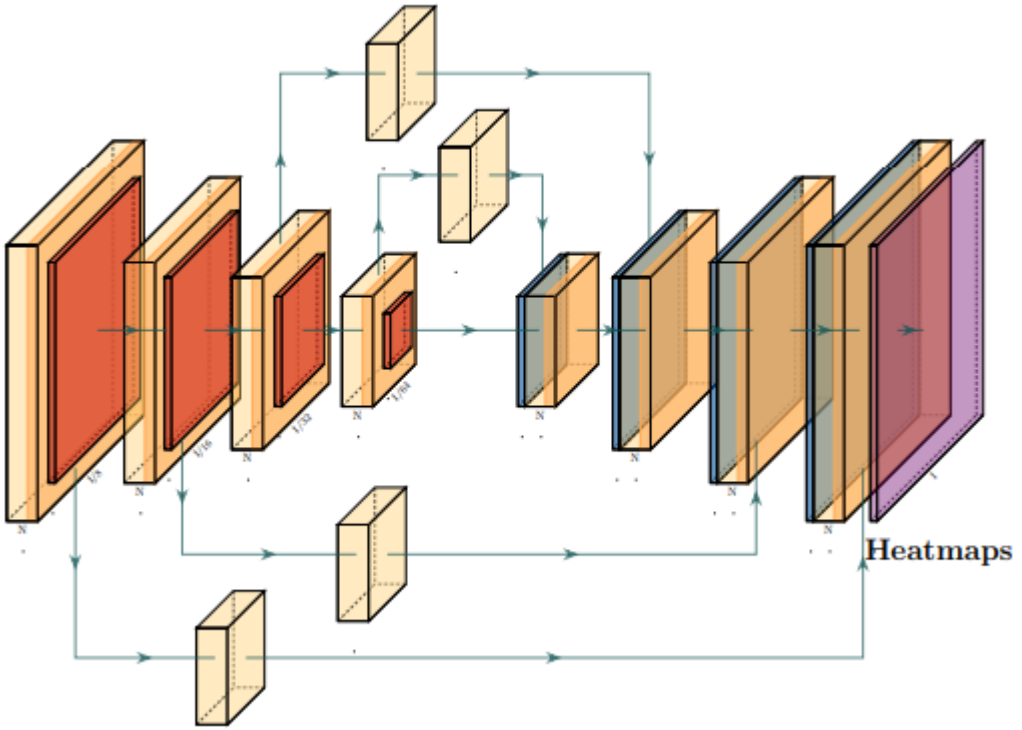
DeepPose Framework

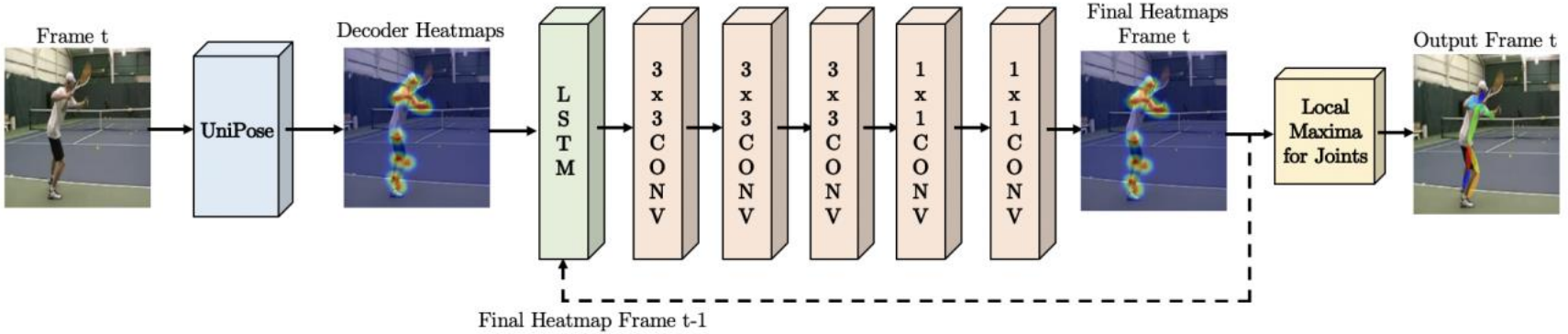
- Gkioxari et al. (2014) used a R-CNN architecture to detect a person, estimate pose, and classify action.
- Each of the joint is shown in a 2D Gaussian distribution centered at the target joint location. Since heatmap representation is more robust than coordinate representation, most of the recent research is based on heatmap representation

Why are HourGlass and ResNet popular?

- First and foremost, they are SIMPLE !
- Do not have vanishing gradient problem or gradient explosion problem
- Skip Connections preserve a lot of information !
- They make use of the input much better than other networks

An HourGlass-ResNet Model





Unipose LSTM based model for pose estimation in videos

Evaluation Metric

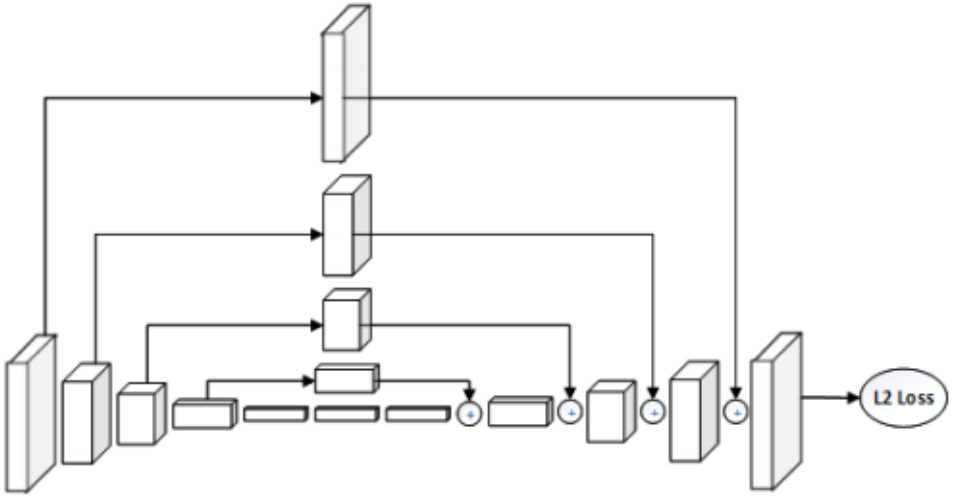
PCK (Percentage of Correct Keypoints)

- A candidate body joint is considered as correct if it falls within the threshold pixels of the ground-truth joint.
- The threshold can be a fraction of the person bounding box size, pixel radius
- 50% of the head segment length of each test image (denoted as PCKh@0.5)

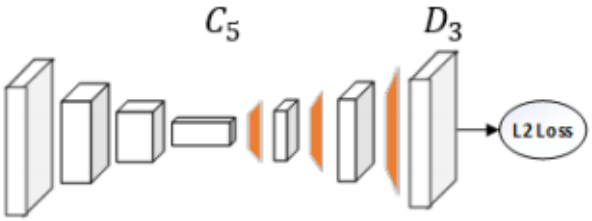
Example !

Simple Baselines for HPE – Microsoft Research

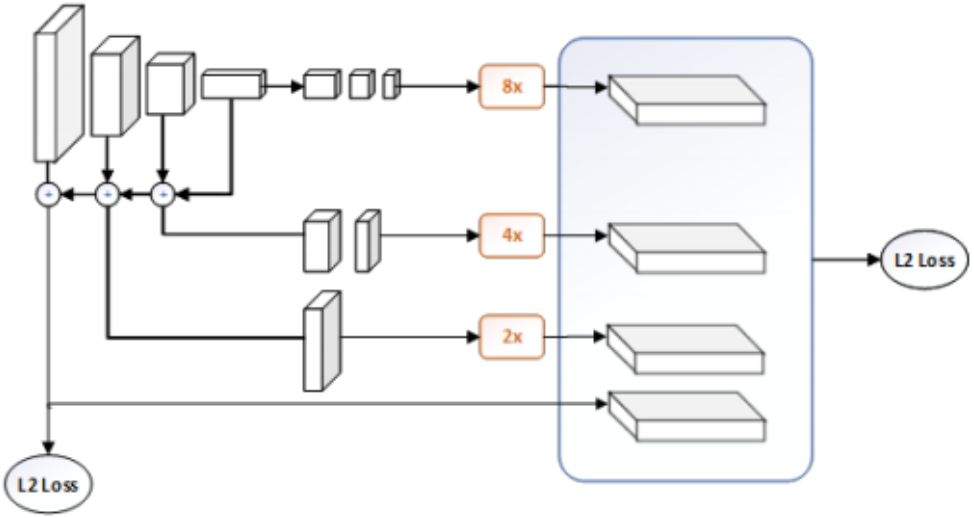
- Proposed a simple model, i.e. a DeConv Head Attached to a ResNet backbone
- This architecture is arguably the simples to generate heatmaps from deep and low-resolution features and also adopted in SOTA Mask R-CNN



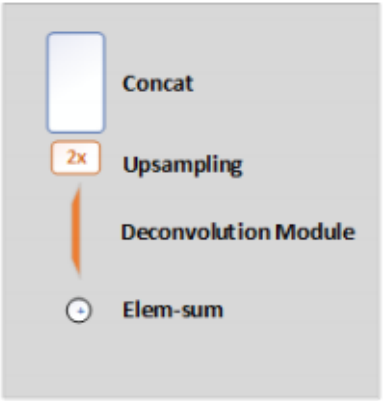
(a) Hourglass



(c) Our Network

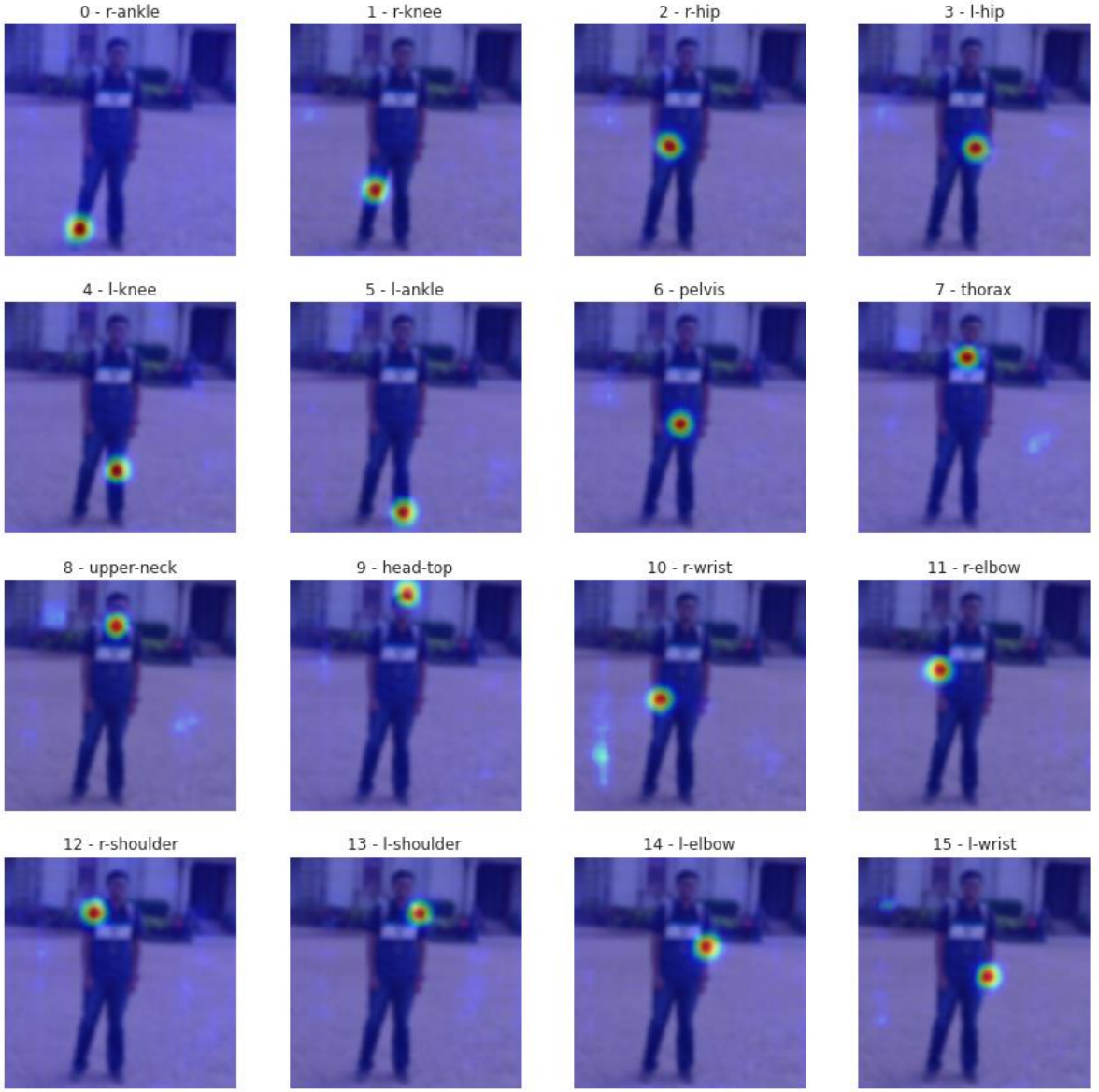


(b) CPN

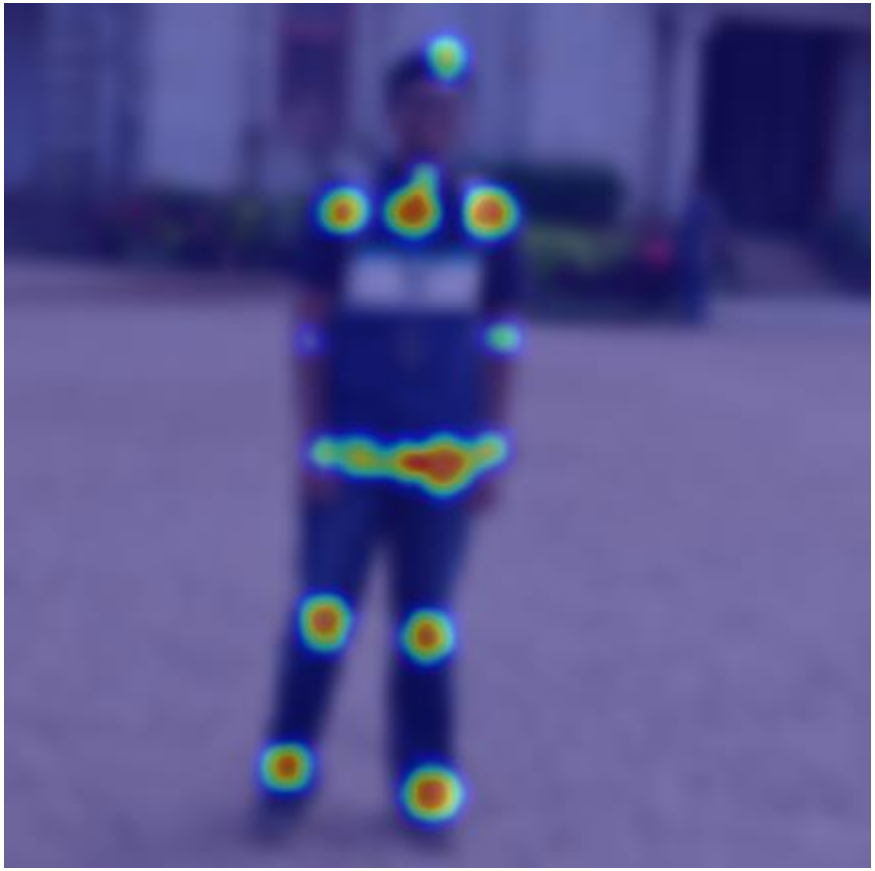


Input Image

Joints HeatMaps



Estimated Pose



Demo !

https://tensorclan.tech/human-pose-estimation

References

1. Bulat, A., Kossaifi, J., Tzimiropoulos, G. and Pantic, M., 2020. Toward fast and accurate human pose estimation via soft-gated skip connections. arXiv preprint arXiv:2002.11098.

2. Chen, Y., Tian, Y. and He, M., 2020. Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 192, p.102897.

3. Sigal L. (2014) Human Pose Estimation. In: Ikeuchi K. (eds) Computer Vision. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-31439-6_584

4. Satyajit G., 2020. Human Pose Estimation and Quantization of PyTorch to ONNX Models - A Detailed Guide. Satyajit Ghana. Available at: https://satyajit.tensorclan.tech/2020/08/pose-estimation-onnx.html [Accessed December 1, 2020].

5. Xiao, B., Wu, H. and Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481).

Thank You