UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

ECE CISE

POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE

Large-scale Intelligent Systems Laboratory
NSF I/UCRC Center for Big Learning
Department of Electrical and Computer Engineering
Department of Computer & Information Science & Engineering

DensePose: Dense Human Pose Estimation In The Wild

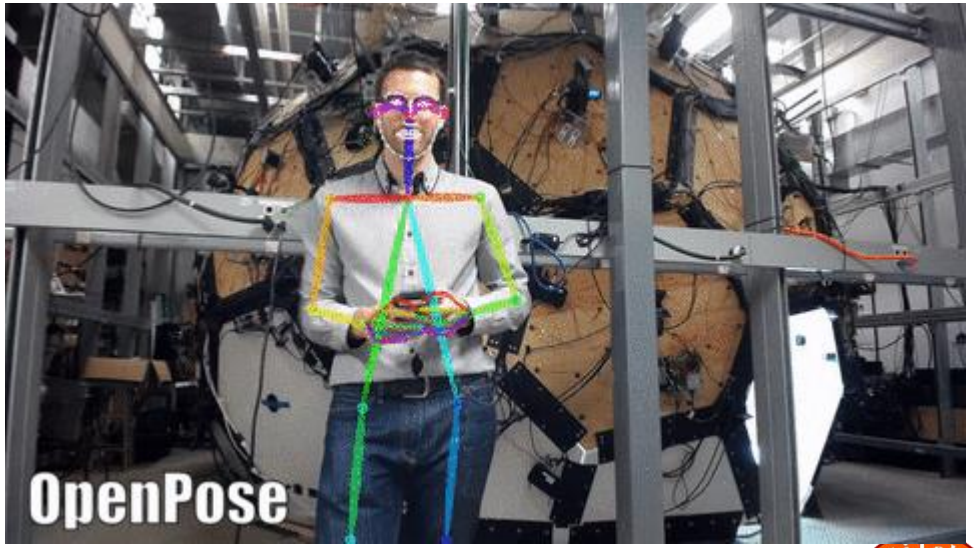
Facebook AI Research
CVPR 2018, Oral Paper
Presented by Chao Li

Dataset & Code: http://densepose.org/
(The dataset will soon be available on this website)

Background

Human 2D pose estimation-the problem of localizing anatomical keypoints or “parts”.

Single Person



Multiple Person

Background

Performance for Single Person on MPII dataset:

http://human-pose.mpi-inf.mpg.de/#results

PCKh evaluation measure

PCKh: PCK measure that uses the matching threshold as 50% of the head segment length.

PCKh @ 0.5

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCKh
Pishchulin et al., ICCV'13	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson et al., NIPS'14	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al., CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al., CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu&Ramanan., CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin et al., CVPR'16*	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al., ECCV'16	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al., ECCV'16	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al., BMVC'16	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis&Zisserman, FG'17**	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al., ECCV'16	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al., CVPR'16*	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat&Tzimiropoulos, ECCV'16	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al., ECCV'16	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ning et al., TMM'17	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon et al., arXiv'17	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu et al., CVPR'17	**98.5**	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al., arXiv'17	98.2	**96.8**	92.2	88.0	**91.3**	89.1	84.9	91.8
Chen et al., ICCV'17	98.1	96.5	92.5	88.5	90.2	**89.6**	86.0	91.9
Yang et al., ICCV'17	**98.5**	96.7	92.5	**88.7**	91.1	88.6	86.0	92.0
Ke et al., arXiv'18	**98.5**	**96.8**	**92.7**	88.4	90.6	89.3	**86.3**	**92.1**

Background

Multiple Person

Top-down approaches:

Employ a person detector and perform single-person pose estimation for each detection

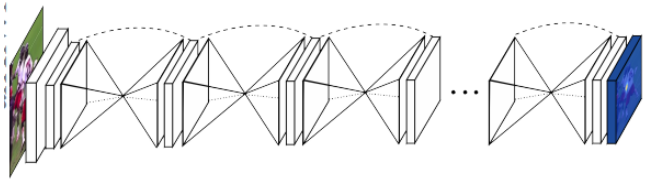
e.g. Stacked Hourglass Networks for Human Pose Estimation, Convolutional Pose Machines

Bottom-up approaches:

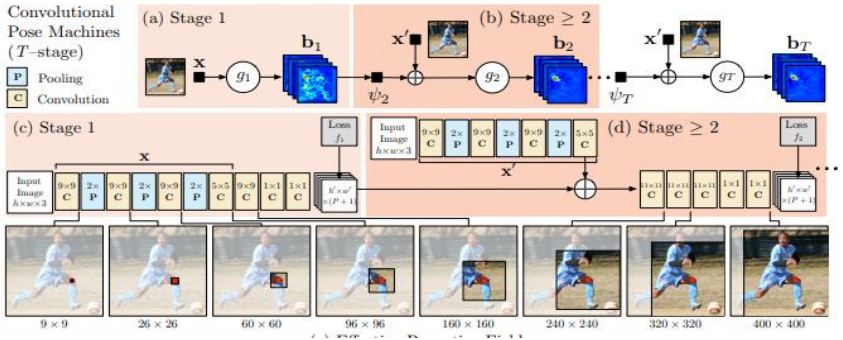
Predict all the point of the image and then decide each point belong to which person

e.g. Openpose

Background

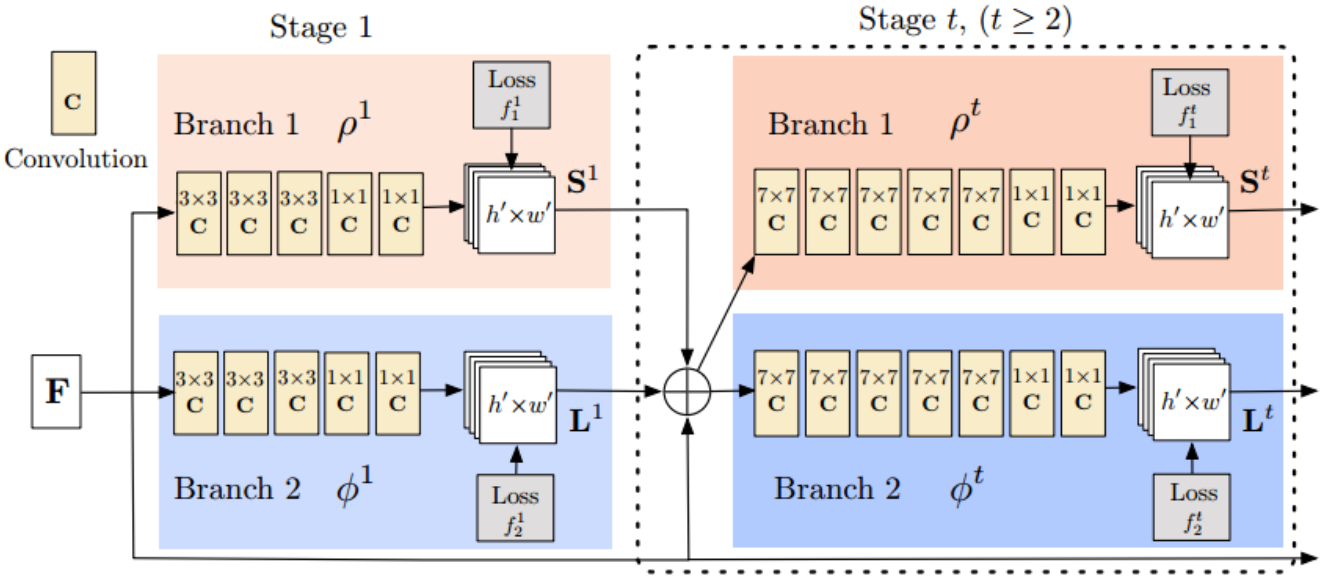


Stacked Hourglass Networks for Human Pose Estimation



Convolutional Pose Machines

Background

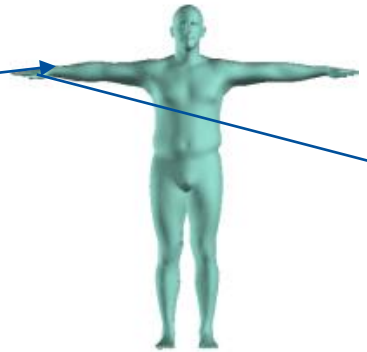


Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

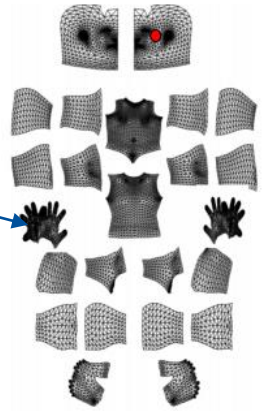
Task Motivation: Motivation and goals

This work aims at pushing further the envelope of human understanding in images by establishing dense correspondences from a 2D image to a 3D, surface-based representation of the human body

RGB Image
(Input)



Template 3d model (SMPL)
(Intermediate Steps)



U-V Coordinate
(Output)

Contribution

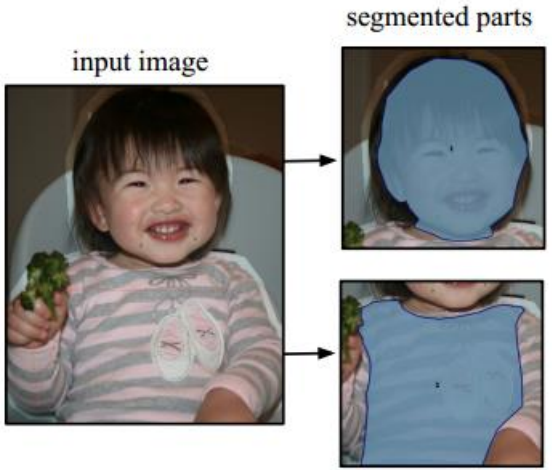
- They introduce the first manually-collected ground truth dataset for the task, by gathering dense correspondences between the SMPL model and persons appearing in the COCO dataset.

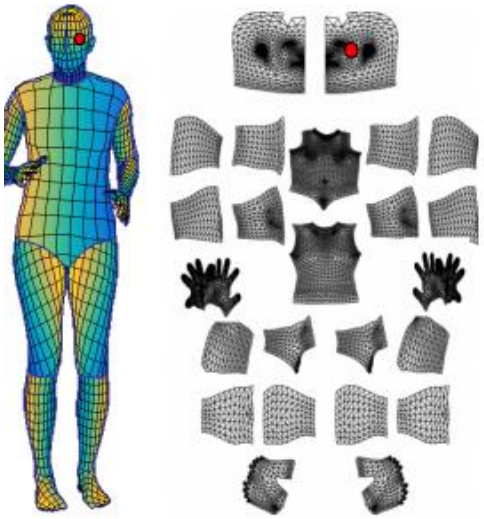
- They use the resulting dataset to train CNN-based systems that deliver dense correspondence ‘in the wild’, by regressing body surface coordinates at any image pixel, observing a superiority of Mask RCNN and cascading networks.

- They explore different ways of exploiting the constructed ground truth information and find that using these sparse correspondences to train a ‘teacher’ network can ‘inpaint’ the supervision signal and improve the performance.

COCO-DensePose Dataset

Step 1:

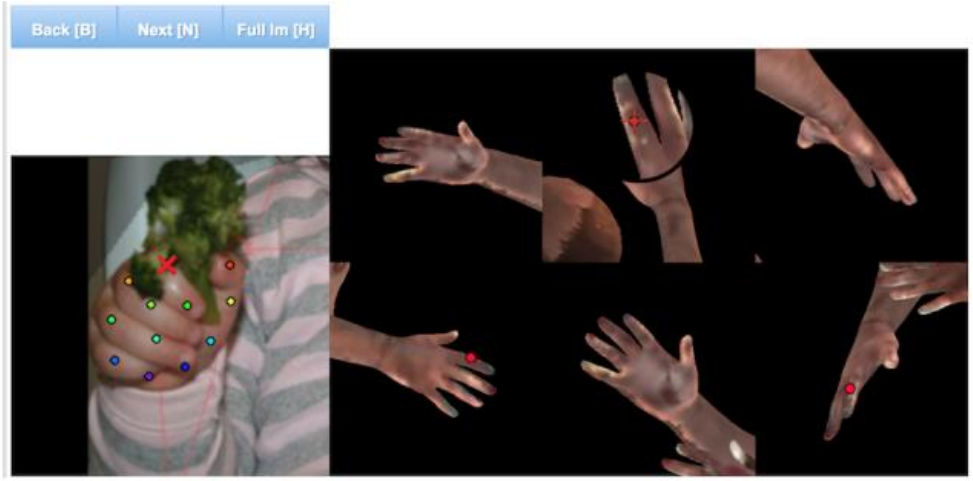




Ask annotators to segment the body into 24 parts as shown in the right figure.

COCO-DensePose Dataset

Step 2:

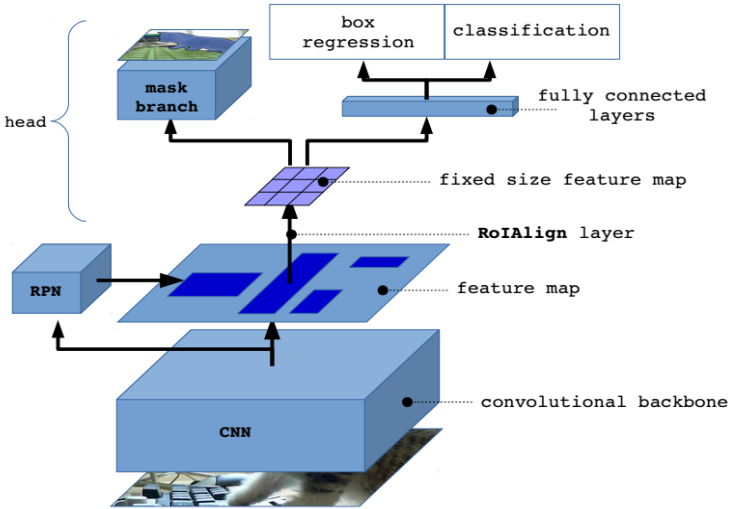


- They sample every part region with a set of roughly equidistant points obtained via k-means and request the annotators to bring these points in correspondence with the surface.

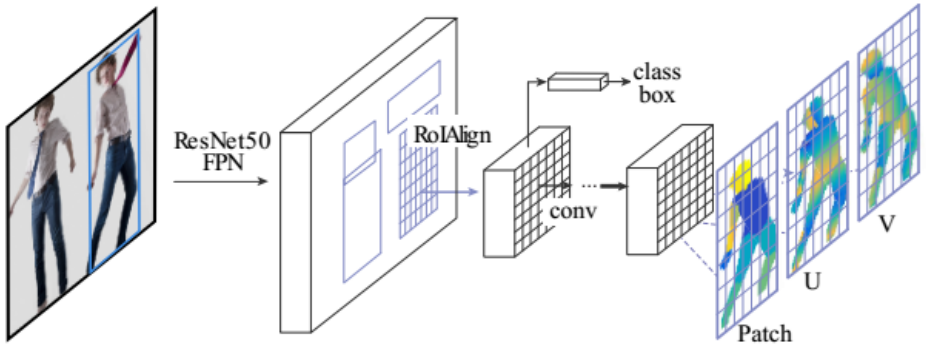
- In order to simplify this task they ‘unfold’ the part surface by providing six pre-rendered views of the same body part and allow the user to place landmarks on any of them.

Proposed Method

Basic Model:



Mask RCNN

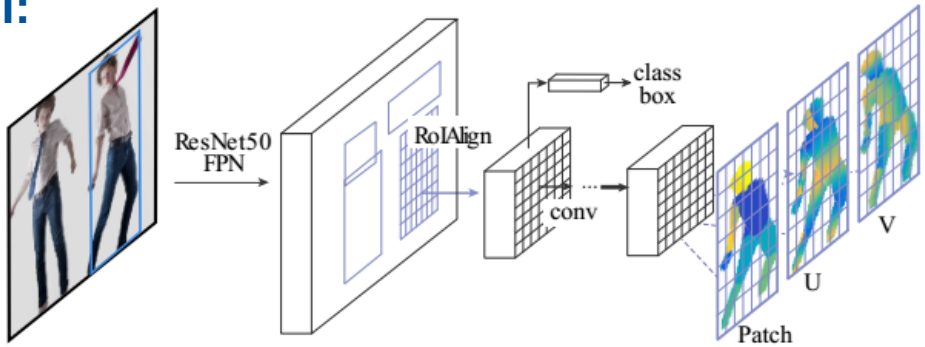


Region-based Dense Pose Regression

Replace the mask head with dense pose head. Such architectures decompose the complexity of the task into controllable modules.

Proposed Method

Basic Model:

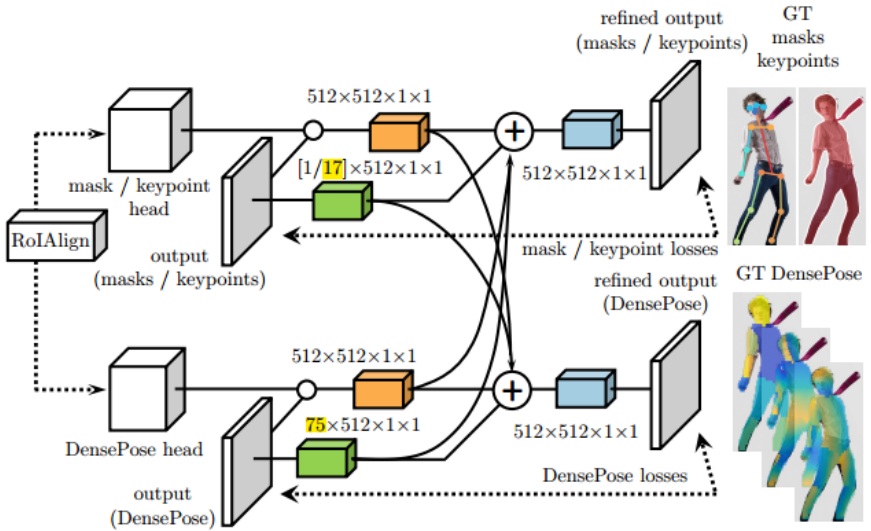


- Patch: a classification that provide the part assignment. (25*H*W)
  They classify a pixel as belonging to either background, or one among several body parts which provide a coarse estimate of surface coordinates.

- (U, V): a regression head that provide part coordinate predictions in each part. (25*H*W*2)
  Indicates the exact coordinates of the pixel within the part.

Proposed Method

Modification 1:



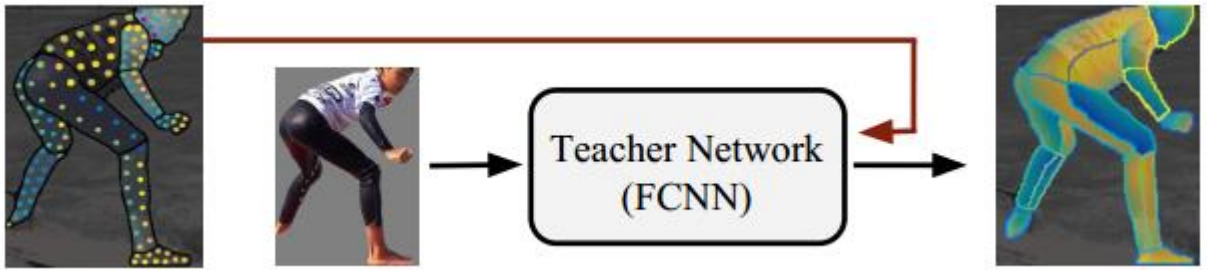
Multi-task cascaded architectures:

- Inspired by the success of recent pose estimation models based on iterative refinement , so they provide the output of previous stage as the input of the next stage.

- exploit information from related tasks, such as keypoint estimation and instance segmentation, which have successfully been addressed by the Mask-RCNN architecture.

Proposed Method

Modification 2:



Multi-task cascaded architectures:

- Even though they aim at dense pose estimation at test time, in every training sample we annotate only a sparse subset of the pixels, approximately 100-150 per human

- They first train a 'teacher network' with their sparse, manually-collected supervision signal, and then use the network to 'inpaint' a dense supervision signal (Output: H*W). Finnally, they used the predicted dense point to train our region-based system.

Evaluation Measures

1. Pointwise evaluation

The prediction is declared correct if the geodesic distance is below a certain threshold (*t*).

As the threshold t varies, we obtain a curve *f(t)* of Ratio of Correct Point (RCP) , and evaluate the area under the curve (AUC):

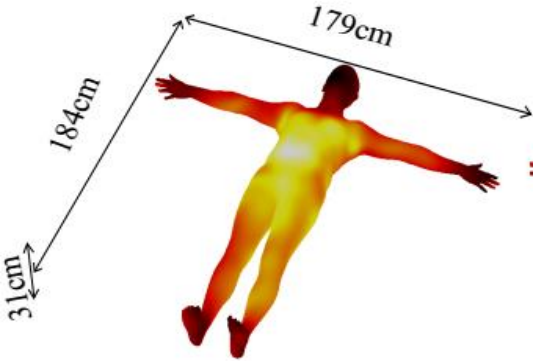
$$\mathrm{AUC}_a = \frac{1}{a}\int_0^a f(t)dt$$

geodesic distance: the distance of two vertices on the surface of 3D human model

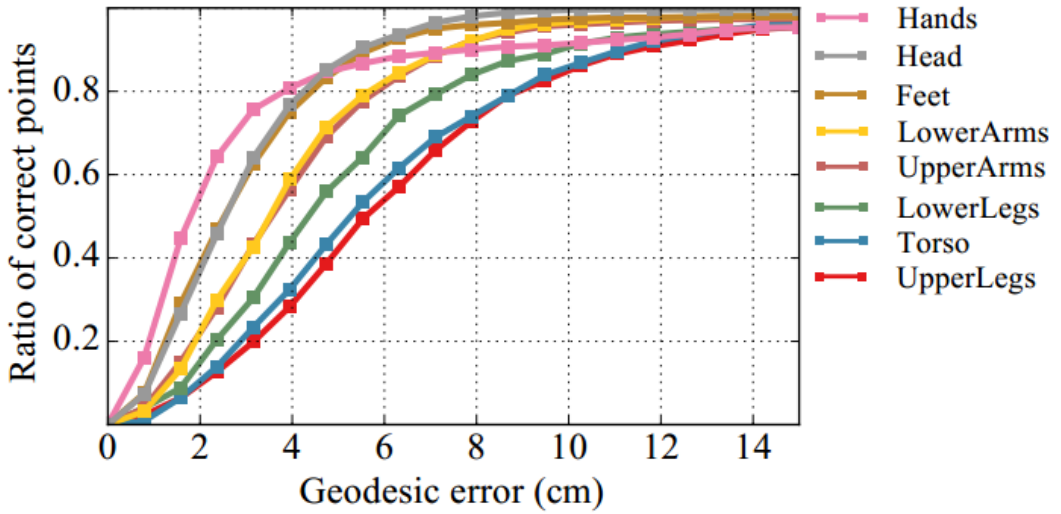
Usually choose two different values of *a = 10cm; 30cm* yielding AUC10 and AUC30 respectively .

Evaluation Measures

1. Pointwise evaluation example



Template Human Model



The curve of pointwise evaluation Example

Evaluation Measures

2. Per-instance evaluation

It's similar with object keypoint similarity (OKS) measure (http://cocodataset.org/#keypoints-eval).

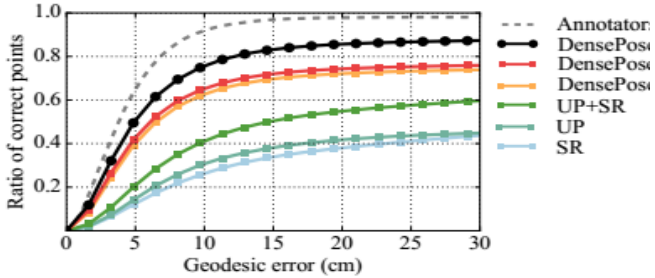
$$\mathrm{GPS}_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2\kappa^2} \right)$$

$$AP@s = \frac{\sum_p \delta\left(OKS_p > s\right)}{\sum_p 1}$$, *OKS = GPS*

- Average Precision (AP) at a number of GPS thresholds ranging from 0.5 to 0.95. (They set κ=0.255m so that a single point has a GPS value of 0.5 if the distance is approximately 0.3 m)).

- AP is consistent with keypoint detection (http://cocodataset.org/#keypoints-eval)

Experiments

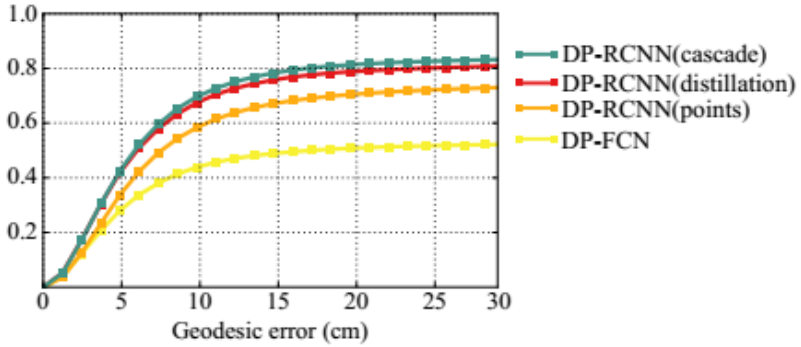


Method	AUC_{10}	AUC_{30}
SR	0.124	0.289
UP	0.146	0.319
SR + UP	0.201	0.424
DensePose + SR	0.357	0.592
DensePose	0.378	0.614
DensePose*	0.445	0.711
Human Performance	0.563	0.835

Single Person:

- Cropped around ground-truth boxes to out the effects of detection performance
- SR: SURREAL dataset
- UP: Unite the People' (UP) dataset
- FCN method is used on all the different datasets to assess the usefulness of the COCODensePose dataset.
- DensePose* use the ground truth mask to out the effects of background.

Experiments



Method	AUC_{10}	AUC_{30}	IoU
DP-FCN	0.253	0.418	0.66
DP-RCNN (points only)	0.315	0.567	0.75
DP-RCNN (distillations)	0.381	0.645	0.79
DP-RCNN (cascade)	0.390	0.664	0.81
DP*	0.417	0.683	–
Human Performance	0.563	0.835	–

Multi-person:

- Distillations: use the “teacher network” to inpaint a dense supervision signal
- Cascade: use multi-task cascaded architectures
- DP* combine all the modifications together.

Experiments

Per-instance evaluation of DensePose-RCNN

Method	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_M$	**AP**$_L$	**AR**	**AR**$_{50}$	**AP**$_{75}$	**AR**$_M$	**AR**$_L$
DensePose (ResNet-50)	51.0	83.5	54.2	39.4	53.1	60.1	88.5	64.5	42.0	61.3
DensePose (ResNet-101)	51.8	83.7	56.3	42.2	53.8	61.1	88.9	66.4	45.3	62.1
Multi-task learning										
DensePose + masks	51.9	85.5	54.7	39.4	53.9	61.1	89.7	65.5	42.0	62.4
DensePose + keypoints	52.8	85.6	56.2	42.2	54.7	62.6	89.8	67.7	45.4	63.7
Multi-task learning with cascading										
DensePose-cascade	51.6	83.9	55.2	41.9	53.4	60.4	88.9	65.3	43.3	61.6
DensePose + masks	52.8	85.5	56.1	40.3	54.6	62.0	89.7	67.0	42.4	63.3
DensePose + keypoints	55.8	87.5	61.2	48.4	57.1	63.9	91.0	69.7	50.3	64.8

Experiments

Qualitative evaluation of DensePose-RCNN:
We observe that their system successfully estimates body pose regardless of skirts or dresses, while handling a large variability of scales, poses, and occlusions.

Experiments

Qualitative results for texture transfer :
The whole video can be seen at http://densepose.org

Thank You!


NSF


NIH
National Institutes
of Health


U.S. DEPARTMENT OF
HOMELAND SECURITY


DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA