DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation

Leonid Pishchulin[1], Eldar Insafutdinov[1], Siyu Tang[1], Bjoern Andres[1],
Mykhaylo Andriluka[1,3], Peter Gehler[2] and Bernt Schiele[1]

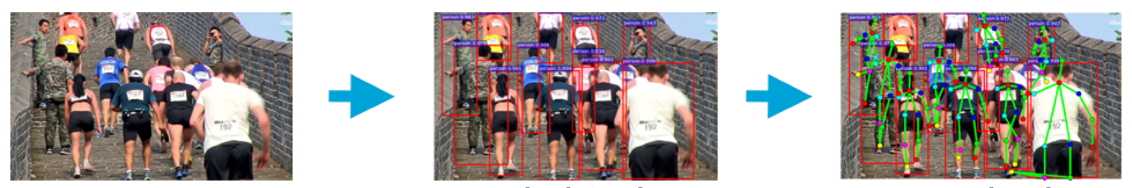
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]Stanford University, Stanford, USA

Goal

- Multi-person pose estimation in monocular RGB images

State of the Art

- single person + occl. reasoning [Chen&Yuille, CVPR'15]
  - no true multi-person reasoning
- two-stage approaches [Eichner&Ferrari, ECCV'10]



- reliable people detector required
- feed-forward approach prone to errors

Contributions

- Novel joint formulation

+ no people detector required
+ joint labeling and grouping of body part hypotheses
+ joint multi-person pose estimation

Code available!
https://pose.mpi-inf.mpg.de

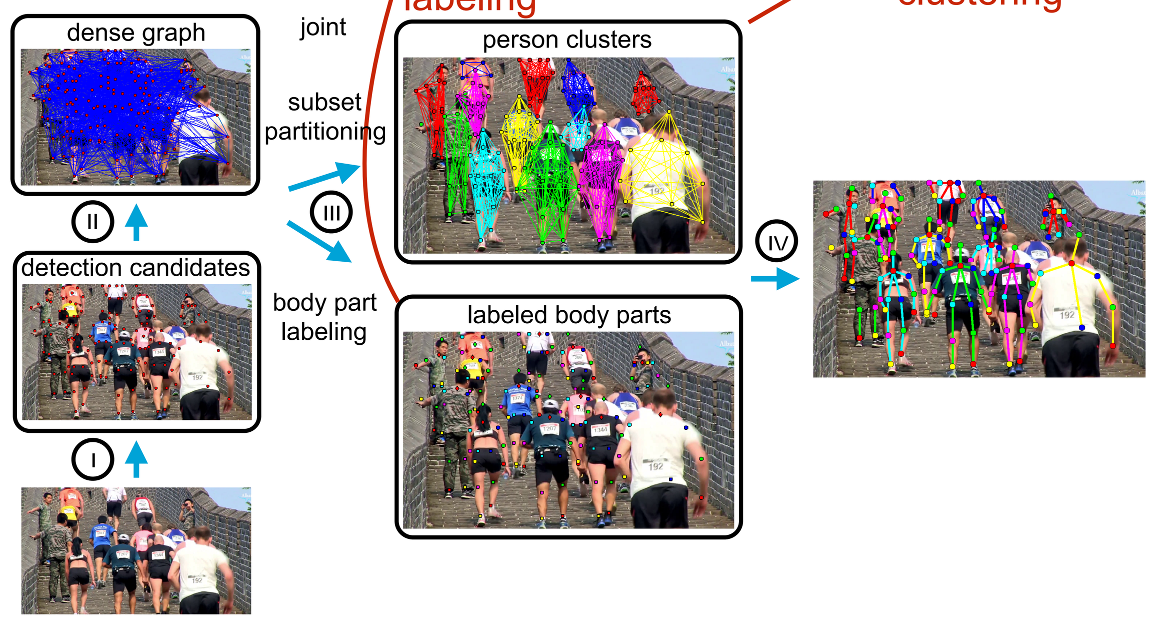
Qualitative results

fully connected graph | joint partitioning and labeling | jointly estimated poses

DeepCut

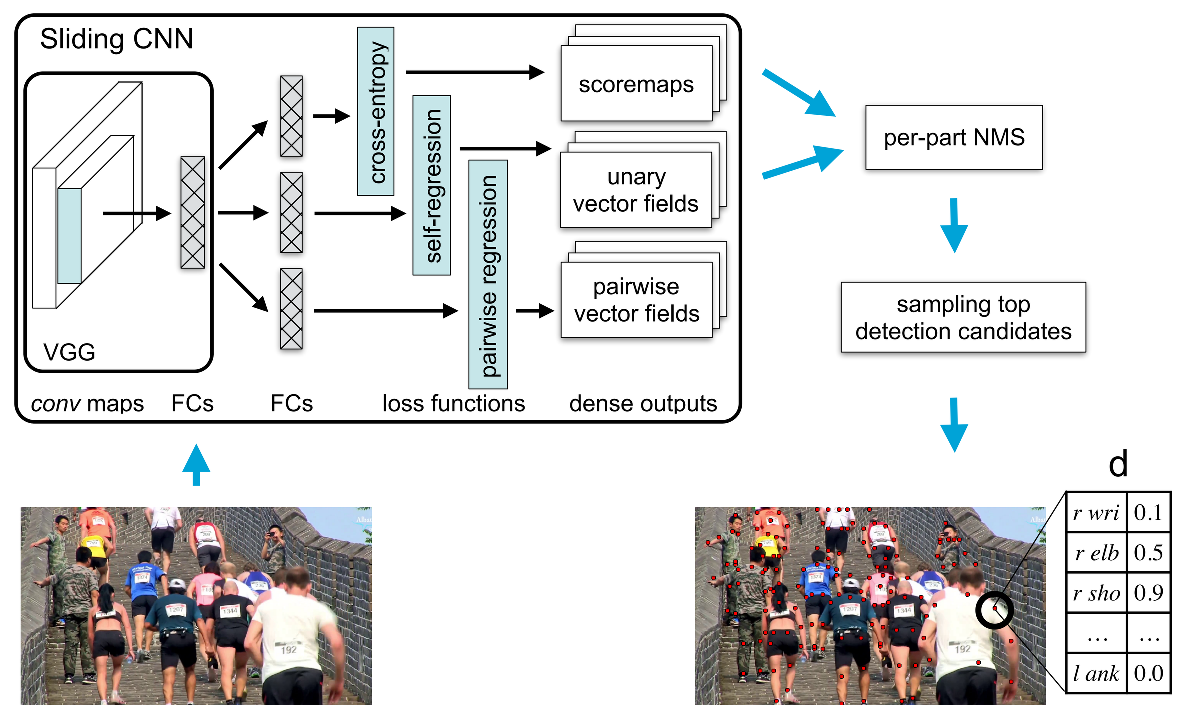
- **Joint labeling and grouping** of parts via 0/1 variables

$$\min_{(x,y)\in X_{DC}} \sum_{d\in D} \sum_{c\in C} \alpha_{dc} \underbrace{x_{dc}}_{\in\{0,1\}} + \sum_{dd'\in\binom{D}{2}} \sum_{c,c'\in C} \beta_{dd'cc'} \underbrace{x_{dc}x_{d'c'}y_{dd'}}$$



I. Unary probabilities

- fully-convolutional CNN architecture based on VGG [7]



II. Pairwise probabilities

- **Proximity**
  - same body part class ($c = c'$)
  - probability $\propto$ distance^{-1}

- **Kinematic relations**
  - different body part classes ($c != c'$)
  - probability via logistic regression from spatial relationship

- **Capture part relationships within/across people**

DeepCut (contd.)

III. Integer Linear Program (ILP)

- Substitute $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$ to convert objective to ILP
- **NP-Hard** problem solved via branch-and-cut (1% gap)
- **Linear constraints** on 0/1 labelings: plausible poses
  - **uniqueness**

$$\forall d \in D: \quad \sum_{c\in C} x_{dc} \le 1$$

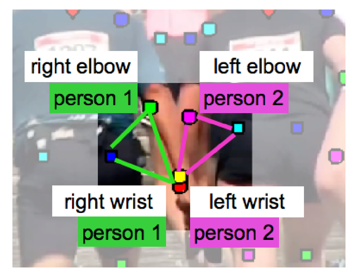
  - **consistency**

$$\forall dd' \in \binom{D}{2}: \quad y_{dd'} \le \sum_{c\in C} x_{dc}$$

$$\forall dd' \in \binom{D}{2}: \quad y_{dd'} \le \sum_{c\in C} x_{d'c}$$

  - **transitivity**

$$\forall dd'd'' \in \binom{D}{3}: \quad y_{dd'} + y_{d'd''} - 1 \le y_{dd''}$$



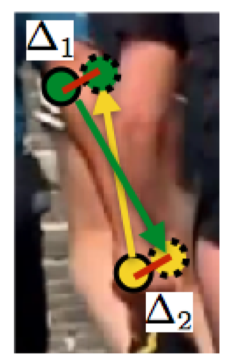
Improvements: DeeperCut (arXiv'16) [4]

I. Unary probabilities

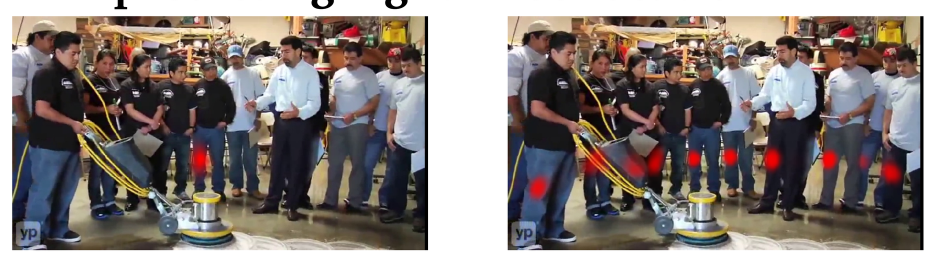
- deeper architectures based on Residual Networks [3]

II. Pairwise probabilities

- **image conditioned** CNN pairwise regression
  - probability via logistic regression from both spatial relationship and appearance



regression from left shoulder predicting right knee location



regression from all parts | unary only

III. Multi-stage optimization

- optimize for reliable parts first, add less reliable later

Stage 1	Stage 2	Stage 3
head, shoulders	elbows, wrists	hips, knees, ankles

Results

Multi-person pose estimation

- MPII Multi-Person dataset [1]
  - Mean Average Precision (mAP) metric

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	mAP	time (s)
subset of 288 images									
DeepCut	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
DeeperCut									
+image cond. pairwise	83.1	75.8	64.6	54.0	60.6	52.0	44.9	62.6	2336
+deeper architecture	83.3	79.4	66.1	57.9	63.5	60.5	49.9	66.2	1333
+multi-stage optim.	**87.5**	**82.8**	**70.2**	**61.6**	**66.0**	**60.6**	**56.5**	**69.7**	**230**
full set									
DeeperCut (1-stage optim.)	73.7	65.4	54.9	45.2	52.3	47.8	40.7	54.7	2785
DeeperCut	**79.1**	**72.2**	**59.7**	**50.0**	**56.0**	**51.0**	**44.6**	**59.4**	485
Faster R-CNN [6] + unary	64.9	62.9	53.4	44.1	50.7	43.1	35.2	51.0	**1**

- We are Family (WAF) [2]
  - Percentage of Correct Parts (PCP) metric

Setting	Head	U Arms	L Arms	Torso	*m*PCP	AOP	(time (s))
DeepCut	**99.3**	81.5	79.5	**87.1**	84.7	86.5	22000
DeeperCut	**99.3**	**83.8**	**81.9**	**87.1**	**86.3**	**88.1**	**13**
Ghiasi et al., CVPR'14	-	-	-	-	63.6	74.0	-
Eichner&Ferrari, ECCV'10	97.6	68.2	48.1	86.1	69.4	80.0	-
Chen&Yuille, CVPR'15	98.5	77.2	71.3	88.5	80.7	84.9	-

- Failure cases

Single person pose estimation

- MPII Single Person dataset [1]

Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCKh	AUC
DeepCut (unary)	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5
DeeperCut (unary)	96.6	94.6	88.5	84.4	87.6	83.9	79.4	88.3	60.7
Newell et al., arXiv'16	97.6	**95.4**	**90.0**	**85.2**	**88.7**	**85.0**	**80.6**	**89.4**	59.6
Wei et al., CVPR'16	**97.8**	95.0	88.7	84.0	88.4	82.8	79.4	88.5	**61.4**
Gkioxary et al., arXiv'16	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1	57.3
Lifshitz et al., arXiv'16	**97.8**	93.3	85.7	80.4	85.3	76.6	70.2	85.0	56.8

- Leeds Sports Poses (LSP) [5]

Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCK	AUC
DeepCut (unary)	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1	63.5
DeeperCut (unary)	97.4	**92.7**	**87.5**	**84.4**	**91.5**	89.9	87.2	90.1	**66.1**
Wei et al., CVPR'16	**97.8**	92.5	87.0	83.9	**91.5**	**90.8**	**89.9**	**90.5**	65.4
Lifshitz et al., arXiV'16	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7	61.1

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR'14*.
[2] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV'10*.
[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv'15*.
[4] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *arXiv'16*.
[5] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*.
[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS'15*.
[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv'14*.