Detangling People: Individuating Multiple Close People and Their Body Parts via Region Assembly

Hao Jiang
Boston College

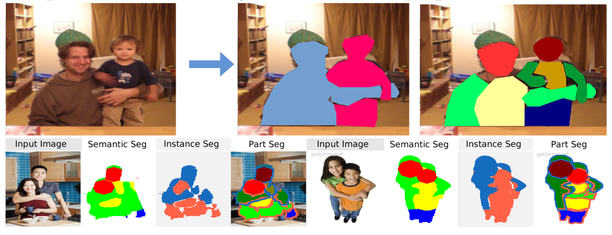
Kristen Grauman
University of Texas at Austin

Contribution

Problem: Parsing "entangled" people in crowded scenes

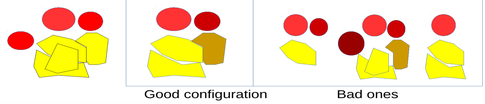
Our idea: Human segmentation as a region assembly problem and its main steps:

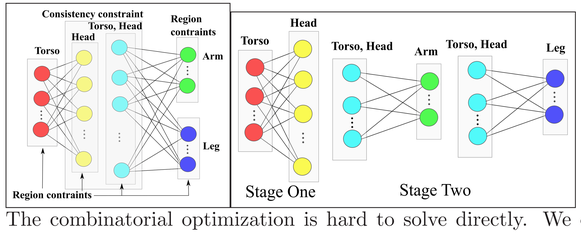
1. Detect regions and body part candidates
2. Optimize their assembly and individuate people



Model

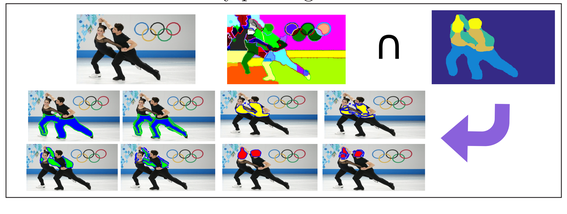
Human segmentation can be formulated as a grouping problem.



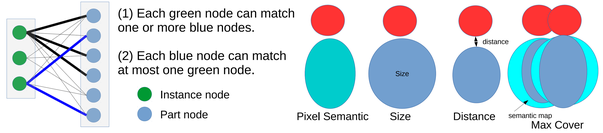


The combinatorial optimization is hard to solve directly. We decompose it into three easier ones.

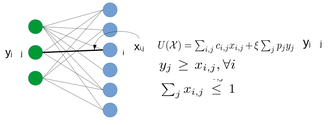
We first extract the body part region candidates:



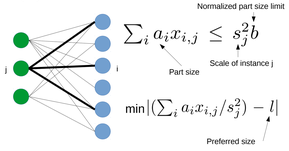
Optimizing Region Assembly



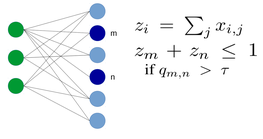
(1) Part-instance association cost



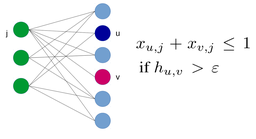
(2) Size constraint



(3) Region overlapping exclusion



(4) Color exclusion



(5) Max covering constraint

The optimization has the following special structure:

$$\min_{x,y,e} \{g^T x + w^T y + \phi 1^T e\} \quad (1)$$

$$\text{s.t. } Ax \leq 1, Bx + Ce + Dy \leq f,\ e \geq 0,\ x, y \text{ are binary.}$$

Its Lagrangian relaxation is:

$$\max_{\nu} \min_{x,y,e} \{g^T x + w^T y + \phi 1^T e + \nu^T (Bx + Ce + Dy - f)\}$$

$$\text{s.t. } Ax \leq 1, 0 \leq e \leq M,\quad x, y \text{ are binary}, \nu \geq 0, \quad (2)$$

For each ν, it is efficiently solved by decomposing into:

$$\text{[P1]: } \min_x (g^T + \nu^T B)x, \text{ s.t. } Ax \leq 1,\ x \text{ is binary.} \quad (3)$$

$$\text{[P2]: } \min_y (w^T + \nu^T D)y, \text{ s.t. } y \text{ is binary.} \quad (4)$$

$$\text{[P3]: } \min_e (\phi 1^T + \nu^T C)e, \text{ s.t. } 0 \leq e \leq M. \quad (5)$$

Experimental Results

Our branch and bound method gives *global optimal body part assembly*. It gives results superior to different competing approaches, and it also achieves the state of art on proxemics recognition.

Average person instance (Upper table) and part (Lower table) IoU ratio comparison (%) for the UCI and MPII dataset.

		Ours	Connected	Greedy	DPM	Poselet	R-I	R-II	R-III	NBest	CNN-D
UCI	F	**63.02**	41.62	46.88	57.64	53.50	56.04	54.01	36.32	61.81	48.58
	B	**63.45**	29.16	45.91	55.59	51.72	47.10	41.47	33.47	57.48	48.96
MPII	F	**57.48**	30.88	40.15	42.21	40.00	56.04	54.01	36.32	47.74	38.24
	B	**57.15**	18.85	39.88	47.91	48.43	47.10	41.47	33.47	48.66	45.48

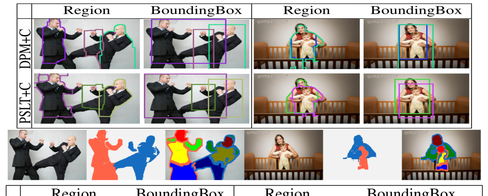
		Ours	C	G	NB	CD
UCI	F	**38.39**	24.75	27.29	37.98	26.49
	B	**38.56**	18.43	32.30	31.08	26.75
MPII	F	**35.48**	20.26	24.25	28.71	22.27
	B	**35.47**	12.54	29.80	29.16	28.91

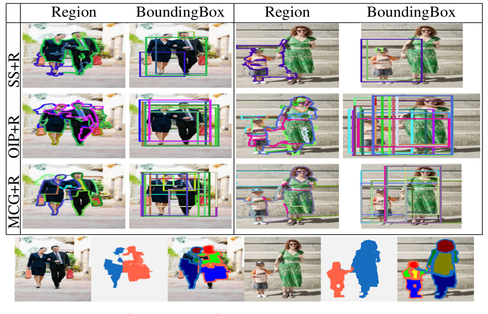
R-I: RCNN+OIP, R-II: RCNN+MCG, R-III: RCNN+SelectiveSearch, CNN-D: CNN pose detector [Chen NIPS14]. In part IoU: Connected component is C, Greedy method as G, Nbest as NB and CNN-D as CD.

Experiment results on proxemics recognition:

	HH	HS	SS	HT	HE	ES	Mean(a)	Mean(b)
Ours	**59.7**	**52.0**	53.9	33.2	36.1	36.2	**45.2**	**47.58**
Yang et al., CVPR12	37	29	50	**61**	38	34	42	38
Sadeghi et al., CVPR11	31	20	40	20	11	12	22	23
Chu et al., ICCV15	41.2	35.4	**62.2**	NA	**43.9**	**55.0**	NA	47.54

Comparison with people detector based approaches:





Comparison with approaches using pose estimation:

Experimental Results (Cont'd)

Sample results of our method:

Sample results for proxemics recognition:

Conclusion

Our novel method segments human instances and labels their body part regions. It is robust to complex human interactions, occlusions, and difficult poses, and it is rotation and scale invariant. Our results compare favorably to a wide array of alternative methods, and we improve the state of art on proxemics recognition.