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

Hao Jiang
Boston College, USA
hjiang@cs.bc.edu

Kristen Grauman
University of Texas at Austin, USA
grauman@cs.utexas.edu

Abstract

Today's person detection methods work best when people are in common upright poses and appear reasonably well spaced out in the image. However, in many real images, that's not what people do. People often appear quite close to each other, e.g., with limbs linked or heads touching, and their poses are often not pedestrian-like. We propose an approach to detangle people in multi-person images. We formulate the task as a region assembly problem. Starting from a large set of overlapping regions from body part semantic segmentation and generic object proposals, our optimization approach reassembles those pieces together into multiple person instances. Since optimal region assembly is a challenging combinatorial problem, we present a Lagrangian relaxation method to accelerate the lower bound estimation, thereby enabling a fast branch and bound solution for the global optimum. As output, our method produces a pixel-level map indicating both 1) the body part labels (arm, leg, torso, and head), and 2) which parts belong to which individual person. Our results on challenging datasets show our method is robust to clutter, occlusion, and complex poses. It outperforms a variety of competing methods, including existing detector CRF methods and region CNN approaches. In addition, we demonstrate its impact on a proxemics recognition task, which demands a precise representation of "whose body part is where" in crowded images.

1. Introduction

Person detection has made tremendous progress over the last decade [1]. Standard methods work best on pedestrians: upright people in fairly simple, predictable poses, and with minimal interaction and occlusion between the person instances. Unfortunately, people in real images are not always so well-behaved! Plenty of in-the-wild images contain multiple people close together, perhaps with their limbs intertwined, faces close, bodies partially occluded, and in a variety of poses. A number of computer vision applications

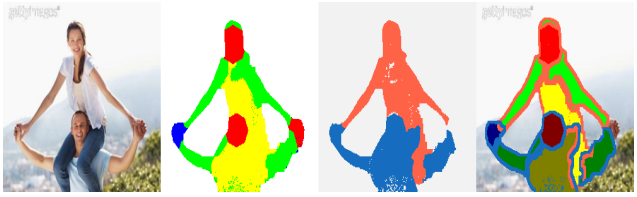


Figure 1. Our method finds human instances and the body part regions (arms, legs, torso, and head). From left to right: input image, semantic body part segmentation, person instance segmentation, final person individuation and part labeling.

demand the ability to parse such natural images into individual people and their respective body parts—for example, fashion [2], consumer photo analysis, predicting interperson interactions [31], or as a stepping stone towards activity recognition, gesture, and pose analysis.

Current methods for segmenting person instances [9, 10, 4, 26, 27, 23, 24] take a top-down approach. First they use a holistic person detector to localize each person, and then they perform pixel level segmentation. Limited by the efficiency and performance of person detectors, such methods are slow when dealing with people at unknown scales and orientations. Furthermore, they suffer when presented with close or overlapping people, or people in unusual non-pedestrian-like body poses [31].

We propose a new approach to detangle people and their body parts in multi-person images. Reversing the traditional top-down pipeline, we pose the task as a region assembly problem and develop a bottom-up, purely region-based approach. Given an input image containing an unknown number of people, we first compute a pool of regions using both body-part semantic segmentations and object proposals. Regions in this pool are often fragmented body parts and often overlap. Despite their imperfections, our method automatically selects the best subset and groups them into human instances. To solve this difficult jigsaw puzzle, we formulate an optimization problem in which parts are assigned to people, with constraints preferring small overlap, correct sizes and spatial relationships between body parts, and a low-energy association of body part regions to their person instance. We show that this problem can be solved

efficiently using decomposition and a branch and bound method.

Fig. 1 shows an example result from the proposed method. Note that we not only estimate pixel-level body part maps, but we also indicate "which part belongs to whom", even in a crowded scene with occluding people.

Experiments on three datasets show our method strongly outperforms an array of existing approaches, including bounding box detectors, CNN region proposals, and human pose detectors. Furthermore, we show the advantage of the proposed optimization scheme as compared to simpler inference techniques. Finally, we demonstrate our person detangler applied to *proxemics recognition* [31, 39, 41], where fine-grained estimation of body parts and body part owners is valuable to describe subtle human interactions (e.g., is he holding her hand or her elbow?).

1.1. Related work

Most previous methods for human instance segmentation require a person detector [11, 9, 10, 4, 6, 7, 8]. Multiple people instance segmentation in TV shows has been studied in [26, 27] using the detector CRF scheme, which combines a person detector and a pixel-level CRF to achieve accurate results. Sequential assignment is used to fit the human instance masks to image data. From instance masks, detailed human segmentation and body part regions are further estimated using a CRF. Hypercolumn [46] is a CNN approach that can be used for people parsing by classifying pixels in the initial person detection bounding boxes.

Whereas existing methods largely take the strategy of first detecting people and then segmenting their parts, we propose a reversal of this conventional pipeline. In particular, we propose to start with a pool of regions that are segments or sub-regions of body parts on multiple people, and then jointly assemble them into individuated person segments. The advantage of not depending on a holistic person detector is not only because these detectors have high computational complexity, but also because it is still a difficult problem for person detectors to deal with complex human poses, inter-person interactions, and large occlusions. Compared to previous detector-based methods, our approach is more efficient and gives better results.

Deep learning approaches have been studied in the joint detection and segmentation scheme [22, 46], related to RCNN [15], though the authors target generic PASCAL object detection as opposed to person individuation and body part labeling. Their method starts from object region proposals such as [16, 17, 18], and each region is classified as a target, such as a human subject, by using features on both color images and binary image masks. Potentially, such a method can be scale and rotation invariant and fast. The challenge is how to propose complete whole object regions, such as the whole mask of a person. This is often a difficult task due to the thin structure of human limbs, and arbitrary human poses. Our proposed method also uses region proposals, but our method allows fragmented sub-regions and can reassemble the broken regions back to human body parts.

Part voting approaches have been intensively studied for human or object instance segmentation. In [3], boundary shape units vote for the centers of human subjects. In [23, 24], the poselets vote for the centers of people instances. The poselets that cast the votes are then identified to obtain the object segmentation. In [25] the object boundary is obtained by reversely finding the activation parts used in the voting. Similar to the Hough Transform, such a voting approach is more suitable to targets that have relatively fixed shape. Our proposed method finds the optimal part assembly using articulation invariant constraints instead of simply voting for the person center; it therefore can be used to segment highly articulated human subjects.

Our method is also related to human region parsing, in that we segment and label each person's body part regions. Human region parsing has been mostly studied for analyzing body part regions of a single person [12, 13, 14, 5]. To handle multiple people, in [4] a pedestrian detector is used to find the bounding box of each single person. Finding people with arbitrary poses using a bounding box detector is still a hard problem, whereas our method naturally handles multiple people with complex interactions and poses. Part segmentation has recently been used to improve semantic segmentation of animals [42], but the pairwise CRF method cannot individuate multiple animal instances. In contrast, our method is able to individuate tangled people with complex poses.

Our work is also distantly related to human pose estimation, which has been intensively studied on depth images [35] and on color images using pictorial structure methods [36, 37, 38] and CNNs [33, 43, 40, 44, 45]. However, unlike our approach, human pose estimation methods usually do not directly give the instance and body part region segmentation. Deepercut [45] optimizes multiple people stick figure representations using integer programming. Different from our approach, Deepercut's body part candidates are body joint candidates from CNN and thus the method does not infer region assembly and it does not deal with region splitting and merging as our approach does. Our method produces multiple human segmentations without extracting human poses (stick figures).

In summary, the main contributions of this paper are: (1) We tackle the new problem of multiple person instance individuation and body part segmentation from region assembly. (2) We propose a novel linear formulation. (3) We propose a Lagrangian relaxation method to speed up lower bound estimation, with which we solve the optimization using fast branch and bound. Our experiments show that our method is fast and effective, outperforming an array of alter-

native methods, and improving the state-of-the-art on proxemics recognition.

2. Method

We first overview our approach (Sec. 2.1), then present the big picture formulation of region assembly as a graph labeling problem (Sec. 2.2). We describe in detail how we implement the components of that formulation (Sec. 2.3). We introduce our efficient optimization approach (Sec. 2.4). Finally, we discuss optimization details (Sec. 2.5).

2.1. Overview

Region proposals may already give body part regions of separate human instances, or more likely they are partial sub-regions of body parts. Many proposal regions do not correspond to body part regions, or may be the union of two individuals' body parts. Our goal is to select a subset of regions from these proposals and reassemble them to individuate human instances and the associated body parts. Intuitively, a good configuration should have arm, leg, torso, and head regions in proportional sizes, and part regions should follow correct neighborhood relations.

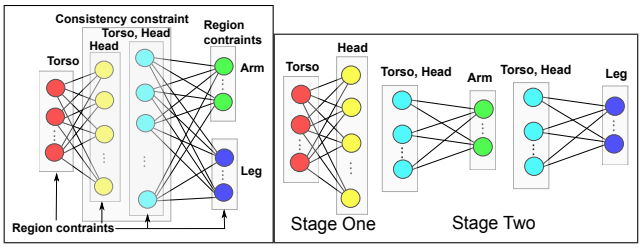
We denote P be the set of overlapping regions or sub-regions of different body parts. Let $\mathcal{X}$ be a vector of integers that indicate a specific region in P is assigned to a person i from $1, \ldots, N$ and N is the number of human instance candidates determined by the algorithm during the optimization (details below). The element of $\mathcal{X}$ is zero if the corresponding region candidate does not belong to any person and a natural number otherwise. We find the optimal $\mathcal{X}$ by jointly optimizing over all potential people instances:

$$\mathcal{X}^* = \text{argmin}_{\mathcal{X}}\{U(\mathcal{X}) - R(\mathcal{X}) + S(\mathcal{X})\} \quad (1)$$
$$\text{s.t. } I(\mathcal{X}) \leq 0, G(\mathcal{X}) \leq 0, W(\mathcal{X}) \leq 0.$$

Here U is the cost of assigning part regions to specific human instances. R is a term that encourages the selected region candidates to cover corresponding body part regions. S is a term that enforces the assembled body regions in each detected human instance to have correct sizes. Apart from these terms, we also introduce constraint I to limit the intersection area between the selected regions, and G to constrain the color histogram between specific region pairs. We also use constraint W to enforce the total body part area of each instance person to be within an upper bound. All these terms are defined in detail below.

2.2. Region assembly as a graph labeling problem

Fig. 2(a) illustrates the region assembly problem as graph labeling. The nodes correspond to the regions or sub-regions of different body parts. Head nodes and head-torso nodes in Fig. 2(a) are also denoted as human instance nodes. The head-torso nodes represent the head-torso region combinations. The binary edges correspond to possible region-to-human instance assignments, and the hyper edges constrain the region coupling and assignment consistency. The



(a) (b)

Figure 2. (a) To optimize region assembly we find the node and edge 0-1 assignment which minimizes the objective in Eq. 1 while satisfying different region constraints on body part assembly. (b) We decompose the optimization into three optimizations in two stages. See text for details.

binary edges and nodes have weights. We essentially need to find an optimal node-edge labeling to minimize the total weight. The optimization is combinatorial. It is hard to solve due to the large number of edges, loopy structure and high order constraints. Instead of directly solving the hard problem, we decompose it into three optimizations on three simpler graphs in two stages, as shown in Fig. 2(b).

The optimization finds the pairing of nodes in each augmented bipartite graph in Fig. 2(b). The nodes on one side are the regions of torsos, arms, or legs. The nodes on the other side are the human instance representations using head regions (stage one) or a head-torso region combination (stage two). Each part region (arm, leg, torso) node can be used at most once, and each human instance node may receive zero or multiple region matches. When selecting multiple part nodes, we assemble corresponding body part regions using "broken" region pieces. The nodes of part regions are coupled by the size and exclusion constraints. The optimizations for torso, arms, and legs in fact share the same structure and therefore we can discuss them at the same time as follows.

2.3. Detailed formulation

Now we flesh out how we instantiate the general formulation presented above. We start from a semantic segmentation map in which each pixel is classified as one of the four part types (arms, legs, torso, and head) or the background. ("Background" = "not any person"; all person pixels are "foreground".) The map is obtained by first computing a stack of probability maps from a CNN (a modified AlexNet) for each part at different scales. Max-pooling is then applied to compute the body part soft semantic map. We use graph cuts with alpha-expansion to generate the final semantic segmentation map.

Overall, the goal is to have a large pool of part candidates with high recall, but possibly low precision; that way, there is a high chance that we can correctly use them to assemble and separate multiple human instances. With this in mind, regions and sub-regions of body parts (torso, arm, and legs) are generated as follows. Apart from using the connected components of body part regions from the CNN-derived semantic segmentation map, we use region proposals from

[16] to "chop" possibly merged part regions into smaller pieces by intersecting the region proposals with each part region. The regions therefore include both whole body parts and fragments of body parts.

Head regions are generated differently because the above method may not always be able to separate close head regions. The head regions are circular regions whose radii are determined by the max-response scale at each head point; the head points are detected by finding peaks in the soft semantic head map using non-maximum suppression. While our framework allows multiple head candidates with different scales at the same point, in practice we find selecting the single most likely head candidate at each point is sufficient. The head candidate regions are further intersected with the person foreground in the semantic map. The number of head candidates tells us the maximum number of people in the image. The head detections automatically tell our system the candidate people number in the image.

We introduce a binary variable $x_{i,j}$, the binarized version of $\mathcal{X}$ in Eq. 1, to label edges in Fig. 2(b): $x_{i,j} = 1$ if region i is selected to be part of person j, otherwise $x_{i,j} = 0$. We have the following constraint on x: $\sum_j x_{i,j} \leq 1$, which means each region can only be assigned to at most one human instance. Each person instance may connect to multiple regions to handle the region splitting case. We also introduce variable y_j to indicate whether person/head candidate j is selected. y is the human instance node variable. We enforce $y_j \geq x_{i,j}, \forall i$. In stage two, the head-torso regions come from the solution of stage one, and y is all one.

2.3.1 Region assignment costs U:

There is a cost $c_{i,j}$ to associate region i to person instance candidate j, and a cost p_j to select instance candidate j. The total assignment cost is $U(\mathcal{X}) = \sum_{i,j} c_{i,j}x_{i,j} + \xi \sum_j p_j y_j$. We optimize y only in stage one. In stage two, y is fixed to be all ones and can be removed from the optimization. p_j equals one minus the head region's peak probability on the head map, so as to emphasize costs incurred on more confident heads. ξ is a constant weight balancing the region association cost against instance selection cost. The cost $c_{i,j}$ aims to associate a person instance with regions that "look like" part of the corresponding body parts based on CNN soft semantic segmentation, and close to the anchor part (head in stage one and torso in stage two).

2.3.2 Size term S and constraint W:

When composing a human instance's body part, the total area of the selected regions is limited by the body part's size: $\sum_i a_i x_{i,j} \leq s_j^2 b$, where a_i is the area of the region i, s_j is the scale of the head candidate j and b is largest possible area of a body part for the reference person (150-pixels tall). In stage one b is the max area of the torso, and in stage two b limits the area of arms or legs. Apart from the hard constraint, a soft one encourages the total area of a region assembly to approach a target size of the corresponding body part. We minimize $|(\sum_i a_i x_{i,j}/s_j^2) - l|$, which can be converted to a linear form: $\min e_j$, s.t. $-e_j \leq (\sum_i a_i x_{i,j}/s_j^2) - l \leq e_j, e_j \geq 0$. Here l is the average body part size of the reference person from different view points. It corresponds to the torso in stage one and arms or legs in stage two.

2.3.3 Exclusion and color consistency constraints I and G:

We also prefer to select regions that are mostly non-overlapping to form each body part region. Thus, we introduce an exclusion constraint I to discourage overlap. Let $z_i = \sum_j x_{i,j}$ indicate whether region i is associated to a human instance. To construct constraint I, we let $z_m + z_n \leq 1$, if $q_{m,n} > \tau$, where $q_{m,n}$ is the area intersection to union ratio between region m and n and τ is a constant.

Apart from intersection exclusion, we also prefer that the color histograms should match if two regions are selected to form the same body part. We thus enforce the constraint G that $x_{u,j} + x_{v,j} \leq 1$, if $h_{u,v} > \varepsilon$, where $h_{u,v}$ is the L_1-distance between the normalized color histogram of region u and v and ε is a constant threshold.

2.3.4 Max covering term R:

If we simply minimize the above terms, x, y will be all zero since all the coefficients in the objective are non-negative. We introduce an extra covering term to encourage the chosen regions to cover the corresponding body part regions in the semantic segmentation map. We maximize the total region size $R = \sum_i r_i z_i$, where $r_i = a_i/m_{t_i}$ and t_i is the part type of candidate i, and m_{t_i} is the total area of part t_i in the semantic map. R is proportional to the total region size. This encourages region covering because we enforce the regions to be mostly disjoint.

Combining the above terms, we have our optimization objective:

$$\min \left\{ \sum_{i,j} c_{i,j}x_{i,j} + \xi \sum_j p_j y_j + \phi \sum_j e_j - \pi \sum_i r_i z_i \right\} \quad (2)$$

$$\text{s.t. } \sum_j x_{i,j} \leq 1, \quad z_i = \sum_j x_{i,j}, y_j \geq x_{i,j}, \forall i, j$$

$$z_m + z_n \leq 1, \text{ if } q_{m,n} > \tau, \quad x_{u,j} + x_{v,j} \leq 1, \text{ if } h_{u,v} > \varepsilon$$

$$\sum_i a_i x_{i,j} \leq s_j^2 b, \; -e_j \leq (\sum_i a_i x_{i,j}/s_j^2) - l \leq e_j, e_j \geq 0,$$

where ϕ and π are coefficients that serve to control the weights of the size and cover terms. If we vectorize variables x, y, e and substitute z by x terms, the optimization has the following format:

$$\min_{x,y,e} \{g^T x + w^T y + \phi 1^T e\} \quad (3)$$

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

Here, the vector x includes the edge variables and the vector y includes the human instance node variables. The dimension of x is the number of torso regions in stage one (or number of arm or leg regions in stage two) times the number of candidate head regions. The dimension of y equals the number of head regions. e is an auxiliary variable vector. g, w are constant coefficient vectors. ϕ is a constant. 1 is an all-one vector. $Ax \leq 1$ is the assignment constraint and $Bx + Ce + Dy \leq f$ represents the region coupling constraints.

2.4. The lower bound

The direct linear relaxation of the integer program has high complexity. With 1000 candidates and 2 human instances, the simplex method takes around 4 seconds to complete, while using the following speedup the time can be reduced to 0.1 seconds using the same CPU.

We obtain the lower bound using the Lagrangian dual. The size constraints and the exclusion constraints complicate the problem. We move them into the objective function. To simplify notation we use the compact format of Eq. 3:

$$\max_{\nu} \min_{x,y,e} \{g^T x + w^T y + \phi 1^T e + \nu^T (Bx + Ce + Dy - f)\}$$
$$\text{s.t. } Ax \leq 1, 0 \leq e \leq M, \quad x, y \text{ are binary}, \nu \geq 0, \tag{4}$$

where ν is the Lagrangian multiplier vector. We introduce an upper bound M for e to avoid unbounded solutions. Since the extra term in the objective is non-positive for all the feasible solutions of the original problem, the Lagrangian dual gives a lower bound.

The internal part of the dual is easy to solve because it can be decomposed into three simple problems (no P2 in stage two):

$$\text{[P1]: } \min_x (g^T + \nu^T B)x, \text{ s.t. } Ax \leq 1, \ x \text{ is binary.} \tag{5}$$
$$\text{[P2]: } \min_y (w^T + \nu^T D)y, \text{ s.t. } y \text{ is binary.} \tag{6}$$
$$\text{[P3]: } \min_e (\phi 1^T + \nu^T C)e, \text{ s.t. } 0 \leq e \leq M. \tag{7}$$

P1 can be solved by sequential assignment: in an assignment graph such as Fig. 2(b), for each body part region node, we check all the links to the human instance node and find the most negative link and let the corresponding x variable to be 1. If there is no negative link, no matching is made and the corresponding x is 0. In P2 and P3, y is set to 0 or 1 and e is set to 0 or M according to the positiveness of their coefficient.

Each set of Lagrangian multipliers corresponds to a lower bound of the original problem. We are interested in the largest lower bound. The bound with respect to the multipliers is a concave function and can be solved using the subgradient method. The iteration alternates between solving for x, y, e and updating ν by $\nu \leftarrow \max(0, \nu + \delta(Bx + Ce + Dy - f))$. Here δ is a small constant 10^{-6}. The initial values of these coefficients in ν are set to zero.

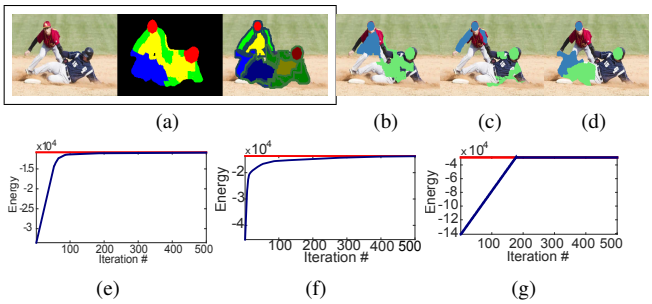


Figure 3. (a): From left to right: input image, semantic segmentation from CNN, and region assembly result of the proposed branch and bound method (shading and boundary color show the instance segmentation). In (a), note how the CNN output does *not* individuate parts into person instances (center), whereas our output does (right). (b-d): Part selection using Lagrangian dual for the torso, arms, and legs. For clarity, torso is not shown in the stage two optimization. Color indicates instance group. (e-g) show the energy of the Lagrangian dual approaches the solution (red line) of the linear program relaxation.

For this problem, the Lagrangian relaxation bound is the same as that of the linear program relaxation. This is due to total unimodularity of the internal problem of the Lagrangian dual [28].

Example. Fig. 3 shows an example of using the Lagrangian relaxation to obtain the lower bound. The Lagrangian relaxation is applied to three optimizations in two stages. As shown in Fig. 3(e-g), the result converges quickly to the linear program relaxation result (the red line) in a few hundred iterations. We see the relaxation assignment is indeed very similar to the globally optimal solution.

The complexity of finding the lower bound using the Lagrangian relaxation is $O(n)$, where n is the number of region proposals times the number of human instance candidates, and we use a fixed number of iterations in the subgradient method. In contrast, the average complexity of a linear relaxation [34] using the simplex method is $O(n \log(n))$. The above dual approach can be extended to estimate the lower bound at each node of the search tree. With the lower bounds, we use the branch and bound method to find the global optimum quickly.

We set the thresholds $\tau = 0.2, \varepsilon = 0.5$ and the weights for the energy terms as $\xi = 500, \phi = 1, \pi = 2 \times 10^5$. We fixed all parameters for all experiments after manually inspecting a few examples. With more labeled data, we can optimize these parameters for even better performance.

2.5. Branch and bound optimization

We use a branch and bound method to globally optimize the three sub-problems. We branch on x with the coefficient that is the median of the undetermined x because it is likely the most ambiguous. If an element of x is forced to be 0, it is equivalent to removing the variable from the optimization. If an element of x is forced to be 1, we can still remove it from the optimization, but we have to change the corresponding coefficients in the optimization. In either case, the Lagrangian relaxation method can still be used to further

Figure 4. Sample results on the UCI and MPII datasets. Each result contains four columns: (1) the input images, (2) our input semantic segmentation body part map, (3) final instance segmentation, and (4) final body part segmentation using the proposed method. We use both shading and different boundary colors to show the segmentation. The same body parts have the same chromaticity (arm: green, leg: blue, torso: yellow, head: red) but have different brightness if they belong to a different person. All figures best viewed on pdf.

obtain the lower bound in each branch. For each branch, if the dual solution is primal feasible and satisfies complementary slackness, it is the global optimal solution. For each node in the search tree, we obtain a primal feasible solution using a simple greedy assignment method and update the upper bound if the feasible solution has a smaller objective. One branch is pruned, if the lower bound is greater than the lowest upper bound or it is infeasible. We always branch on the node with the lowest lower bound. Due to the tight lower bound, the branch and bound terminates quickly. We also use a relaxed tolerance gap to speed up the procedure. The tolerance gap $(u-l)/|u+l|$, where u is the lowest upper bound and l is lowest lower bound in active branches, can be set to 20% and the method still gives good results. The lower bound can be found efficiently using the Lagrangian relaxation method in section 2.4. For most problems in the experiments, where n averages around 500, the branch and bound procedure terminates in a few seconds.

3. Experimental results

Overview: In the following, we compare our approach to 1) simpler inference methods, to show the value added over the initial CNN body part maps; 2) bounding box detector methods; 3) CNN methods using region proposals; 4) human pose detection based methods. Having established our method's accuracy, we then demonstrate its applicability for a downstream task: proxemics recognition.

Datasets and evaluation metrics: We evaluate the proposed method on 3 datasets: UCI [31], which contains 589 images, 100 images from the MPII dataset [32] that contain multiple tangled people, and Buffy [27]. The images include complex human poses, interactions, and occlusions among subjects. The person scales and orientations are unknown. These are the most comprehensive datasets available for locating people and parts. Nearly all test images have touching entangled people, whereas in generic recognition datasets like PASCAL or COCO, only 10% to 30% of the images even *have* multiple people. We manually label the human instances and four part regions in UCI and MPII datasets for ground truth evaluation only (not to train the CNN). The pixel level part semantic segmentation CNN is an AlexNet with the fully connected layers converted to convolutional layers and trained on the LSP dataset [20].

We use the standard area intersection to union (IoU) ratio against the ground truth labeling to quantify performance. We report the IoU for the human instances and mean IoU over all body part labels within each instance. To compute forward (F) scores, we match each ground truth segment to the best segmentation result. For the backward (B) scores, the matching is the other way around. The forward score is affected by missing detections and the backward score by the false alarms.

Fig. 4 shows sample results of our method on UCI and MPII.

Are our initial CNN body part maps enough? Would a simpler inference method on top of the CNN maps be sufficient? First, we stress that the CNN body part maps are not enough by definition, as they do not individuate which body part blobs go to which person. The person and part segmentations merge when people are close. For example, if their arms touch, that yields one connected component in the CNN output; see Fig 4, second column in each set.

Our CNN semantic segmentation itself is reasonable. On UCI and MPII, the average foreground pixel accuracy and part pixel accuracy are 73.13% and 42.41% respectively. However, this does not easily transfer to a good human instance segmentation. To confirm this quantitatively, we test 1) a baseline that returns connected components in the CNN map for the body part labels (**Connected**), and 2) a base-

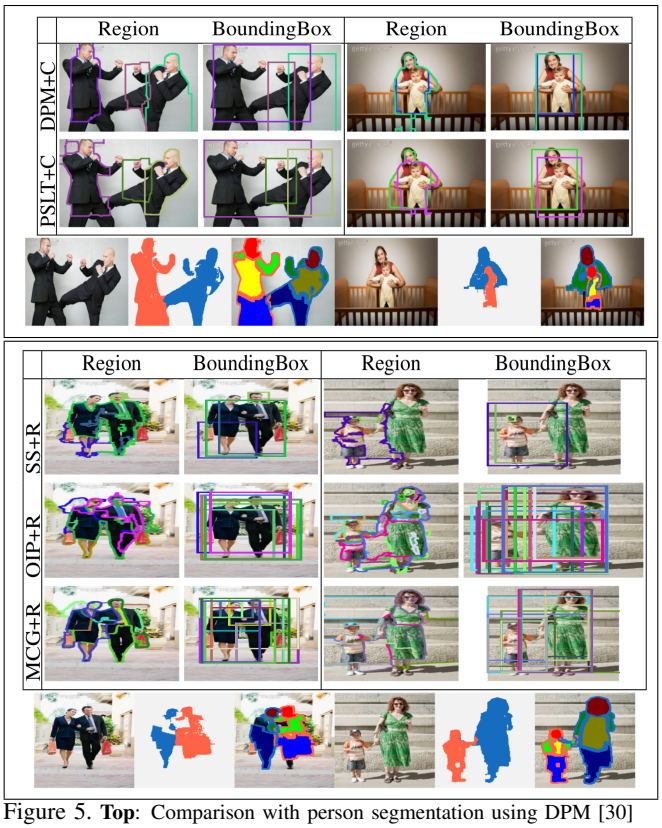


Figure 5. **Top**: Comparison with person segmentation using DPM [30] and Poselet (PSLT) [23] detectors combined with GrabCut (C) [29], and **Bottom**: object proposal methods (selective search (SS) [17], object independent proposal (OIP) [16], and MCG [18]) combined with an RCNN (R) person detector [15] (Bottom). For each set, our results are shown in the last row.

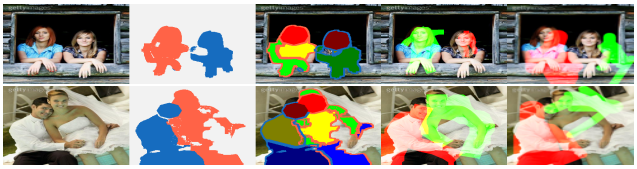


Figure 6. Comparison with methods that use human pose detectors [21, 43]. Our method's results are in column 2 (instance segmentation) and 3 (part segmentation). Column 4 shows results of [21] and column 5 shows results of [43]. Here we show the pose masks before CRF refinement.

line that greedily finds the grouping of each person sequentially (**Greedy**). For the latter, after the lowest cost group is found, the regions in that group are removed and we proceed to the next one until all the head regions are exhausted. Note that naive exhaustive search is extremely slow due to the huge search space.

Table 1 shows the results. Our full method's strong results relative to both these baselines reveals the role of our region assembly optimization. Our efficient global optimization is necessary.

Comparison with bounding box detector methods: One widely used method (e.g., [9, 10, 4, 23]) to extract human instances is to first detect people in a set of bounding boxes, and then obtain pixel-level segmentation. To test such a baseline, we use a deformable part model (**DPM**) person detector [30] and poselet (**Poselet**) person detector [23], and refine the segmentation with GrabCut [29]. We adjust the threshold of the person detectors to the lower side so that they can detect more people instances. We also adjust the parameters of GrabCut to achieve the best performance.

As shown in Fig. 5 (top), when the people have complex poses, interactions, and occlusions, the bounding boxes from person detectors are not accurate. It is a non-trivial task for a pixel-level segmentation method to correct such errors without manual interaction. Indeed, our method gives consistently superior results to the detector based approach (see Table 1 DPM and Poselet).

Comparison with CNN object detectors using region proposals: Another method for human instance segmentation is first generating many region proposals and then using a classifier to extract true human instances, e.g. [22]. RCNN [15] can also be modified to achieve such a function. To compare this idea to our method, we test three kinds of region generation methods: selective search [17], MCG [18], and object-independent proposals [16]. Each rectangle image patch that encloses a region proposal is then sent to RCNN to determine the probability of the image patch containing a human instance. For fair comparison, apart from the original dataset for training, we also include the LSP images [20] in the refinement, which improves the baseline's human classification result.

Fig. 5 (bottom) shows sample results. In images with tangled people instances, region proposals often have a hard time to obtain full human segmentations, because the human structures are not directly used in these region proposal methods. Table 1 shows the quantitative comparison. Our method gives better results.

Comparison with methods using human pose detectors: Next, we compare our approach to two stick figure pose detectors (postprocessed to provide segmentations). The first uses a flexible stick figure person detector [26]; the second is based on CNNs for part detection [43].

Human instance segmentation scores are not reported in [26]; the body part IoU scores are based on different body part region definitions from ours and the code [26] is not publicly available. Thus we compare with the upper bound performance of the **N-best** poses [21] that [26] uses for human segmentation. We follow [26] to prune the N-best poses to remove very close estimates while maintaining the variety; a few thousand candidate poses are extracted. These poses are then refined to person masks following [26]. Instead of selecting the best candidates using the energy as in [26], we directly find candidates that maximize the IoU ratio score using ground truth. We also specify the order of the matching when computing the forward score so that occlusion can be counted away. The score is thus an upper bound of the baseline.

		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

Table 1. Average person instance (Left table) and part (right table) IoU ratio comparison (%) for the UCI and MPII dataset. In the left table, notations include Connected: Connected component method. R-I: RCNN+OIP, R-II: RCNN+MCG, R-III: RCNN+SelectiveSearch, CNN-D: CNN pose detector [43]. F: forward score. B: backward score. In part IoU table: Connected component is denoted as C, Greedy method as G, Nbest as NB and CNN-D as CD.

The CNN pose detector [43] baseline (**CNN-D**) is designed to detect a single person stick figure. To make it generalize to our multi-person images, we use DPM [30] to detect candidate bounding boxes and then apply the CNN pose detector [43] to find poses in each bounding box. We refine the stick figure detection to obtain instance segmentation following [26]. As seen in Table 1, our method outperforms both pose-based methods (NBest and CNN-D) on UCI and MPII overall.

Fig. 6 shows samples of raw masks whose refinement best fits the ground truth regions. Our method is more robust when handling occlusion and complex people interactions than traditional stick figure pose detectors. Apart from the instance segmentation scores, our method also gives better part segmentation scores than the pose detector methods (see Table 1).

Comparison to state-of-the-art in person individuation: We compare to [27], a method specifically aimed at human individuation that represents the state of the art, with our method on all the images from the Buffy dataset episode 4, 5, 6. Our average forward and backward scores are 68.22% and 69.66%, which are higher than the average score of 62.4% reported in [27]. Note that [27] is trained on the Buffy dataset but ours is not. We also compare our people parsing method with the hypercolumn method [46] on the articulated object categories in PASCAL VOC. Our person body part APr at 0.5 is 0.312, which is higher than the hypercolumn approach which has a 0.285 APr at 0.5.

The detector CRF approach [26, 27] also has higher complexity than our method, especially when we do not know the people's orientation. Finding a large set of pose candidates in 10 orientations alone takes 3 minutes with a 3GHz machine. Our method takes less than 10 seconds on region assembly on each image. Our method is rotation invariant. Our method may fail (Fig. 8) if gross errors happen on semantic maps. With better pixel level semantic segmentation, the human instance detection and segmentation result can be further improved.

Application for proxemics recognition: Finally, we demonstrate the utility of our human region parsing for *proxemics* recognition. Proxemics is the study of the spatial separation individuals naturally maintain in social situations. The UCI dataset was created to study proxemics, and is labeled for 6 classes: hand-hand (HH), hand-shoulder (HS), shoulder-shoulder (SS), hand-torso (HT), hand-elbow (HE) and elbow-shoulder (ES).

We use features that include the min and max distances between each pair of upper body part regions of a person pair normalized by the average scale of the two subjects, the normalized horizontal and vertical distance of heads and the scale difference. The data for training and testing are uniformly split at random, following the setup in [31]. To learn the 6 proxemics classes on top of these features, we use a random forest classifier with 100 trees and unlimited tree depth. We repeat the experiment 10 times and report the average accuracy. We do not use ground truth head locations.

	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**
[31]	37	29	50	**61**	38	34	42	38
[39]	31	20	40	20	11	12	22	23
[41]	41.2	35.4	**62.2**	NA	**43.9**	**55.0**	NA	47.54

Figure 7. Sample proxemics recognition. Row one: Our result (D) matches the ground truth (G). Row two: Failure cases. The table shows the average precision (%) in proxemics recognition. Mean(a) is the average of all classes. Mean(b) excludes class HT.

Figure 8. Sample failure cases.

Fig. 7 shows sample classifications and AP scores. Our average AP score is higher than all the competing methods [31, 39, 41]. Our weakness vs. [31] on HT is likely because not only baby hugging but also other hand-on-torso images are classified as HT. Compared to the prior pose detectors, our method is more resistant to large occlusions, non-pedestrian poses, and complex interactions.

4. Conclusion

We propose a novel method to segment human instances and label their body parts using region assembly. The proposed method is able to handle complex human interactions, occlusion, difficult poses, and is rotation and scale invariant. Our branch and bound method is fast and gives reliable results. Our method's results compare favorably to a wide array of alternative methods, and we improve the state of art on proxemics recognition.

Acknowledgements: This research is supported in part by U.S. NSF 1018641 and a gift from Nvidia (HJ) and ONR PECASE N00014-15-1-2291 (KG).

References

[1] P. Dollár, C. Wojek, B. Schiele and P. Perona. Pedestrian Detection: An Evaluation of the State of the Art. TPAMI 2012.

[2] K. Yamaguchi, H. Kiapour, L.E. Ortiz, T.L. Berg. Retrieving Similar Styles to Parse Clothing. TPAMI, vol.37, no.5, 2015.

[3] M.D. Rodriguez and M. Shah. Detecting and Segmenting Humans in Crowded Scenes. ACM MM 2007.

[4] Y. Bo and C.C. Fowlkes. Shape-based Pedestrian Parsing. CVPR 2011.

[5] J. Dong, Q. Chen, X. Shen, J. Yang, S. Yan. Towards Unified Human Parsing and Pose Estimation. CVPR 2014.

[6] T. Lim, S. Hong, B. Han, J.H. Han. Joint Segmentation and Pose Tracking of Human in Natural Videos. ICCV 2013.

[7] H. Wang, D. Koller. Multi-Level Inference by Relaxed Dual Decomposition for Human Pose Segmentation. CVPR 2011.

[8] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts. IJCV, 79:285298, 2008.

[9] J.C. Niebles, B. Han, and L. Fei-Fei. Efficient Extraction of Human Motion Volumes by Tracking. CVPR 2010.

[10] J.C. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting Moving People from Internet Videos. ECCV 2008.

[11] T. Zhao, R. Nevatia. Bayesian Human Segmentation in Crowded Situations. CVPR 2003.

[12] G. Mori, X. Ren, A.A. Efros and J. Malik. Recovering Human Body Configurations: Combining Segmentation and Recognition. CVPR 2004.

[13] H. Jiang. Finding People Using Scale, Rotation and Articulation Invariant Matching. ECCV 2012.

[14] P. Srinivasan, J. Shi. Bottom-up Recognition and Parsing of the Human Body. CVPR 2007.

[15] R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014.

[16] I. Endres, D. Hoiem. Category-Independent Object Proposals With Diverse Ranking. PAMI February 2014.

[17] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, A.W.M. Smeulders. Selective Search for Object Recognition. IJCV, v.104, no.2, page 154-171, 2013.

[18] P. Arbelaez, J. Pont-Tuset, J.T. Barron, F. Marques, J. Malik. Multiscale Combinatorial Grouping. CVPR 2014.

[19] J. Long, E. Shelhamer, T. Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

[20] S. Johnson, M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. BMVC 2010.

[21] D. Park, D. Ramanan. N-Best Maximal Decoders for Part Models. ICCV 2011.

[22] B. Hariharan, P. Arbelaez, R. Girshick and J. Malik. Simultaneous Detection and Segmentation. ECCV 2014.

[23] T. Brox, L. Bourdev, S. Maji, J. Malik. Object Segmentation by Alignment of Poselet Activations to Image Contours. CVPR 2011.

[24] L. Bourdev, S. Maji, T. Brox, J. Malik. Detecting People Using Mutually Consistent Poselet Activations. ECCV 2010.

[25] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic Contours from Inverse Detectors. ICCV 2011.

[26] L. Ladicky, P. Torr, A. Zisserman. Human Pose Estimation using a Joint Pixel-wise and Part-wise Formulation. CVPR 2013.

[27] V. Vineet, J. Warrell, L. Ladicky, P. Torr. Human Instance Segmentation from Video using Detector-based Conditional Random Fields. BMVC 2011.

[28] L.A. Wolsey, G.L. Nemhauser. Integer and Combinatorial Optimization. Wiley, 1999.

[29] C. Rother, V. Kolmogorov, A. Blake. "GrabCut" Interactive Foreground Extraction using Iterated Graph Cuts. Siggraph 2004.

[30] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. IEEE TPAMI, Vol.32, No.9, Sep. 2010.

[31] Y. Yang, S. Baker, A. Kannan, D. Ramanan. Recognizing Proxemics in Personal Photos. CVPR 2012.

[32] M. Andriluka, L. Pishchulin, P. Gehler and S. Bernt. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. CVPR 2014.

[33] X. Fan, K. Zheng, Y. Lin, S. Wang. Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation. CVPR 2015.

[34] V. Chvátal. Linear Programming. W.H. Freeman, 1983.

[35] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. CVPR 2011.

[36] P.F. Felzenszwalb, D.P. Huttenlocher. Pictorial Structures for Object Recognition. IJCV 61(1), Jan. 2005.

[37] Y. Yang, D. Ramanan. Articulated Pose Estimation with Flexible Mixtures-of-Parts. CVPR 2011.

[38] M. Andriluka, S. Roth, B. Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. CVPR 2009.

[39] A. Sadeghi and A. Farhadi. Recognition Using Visual Phrases. CVPR 2011.

[40] A. Toshev, C. Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. CVPR 2014.

[41] X. Chu, W. Ouyang, W. Yang, X. Wang. Multi-task Recurrent Neural Network for Immediacy Prediction. ICCV 2015.

[42] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille. Joint Object and Part Segmentation using Deep Learned Potentials. ICCV 2015.

[43] X. Chen, A. Yuille. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. NIPS 2014.

[44] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, B. Schiele, DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. CVPR 2016.

[45] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model. ECCV 2016.

[46] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik. Hypercolumns for object segmentation and fine-grained localization. CVPR 2015.