European Conference on Computer Vision 2012 (ECCV'12)

Finding People Using Scale, Rotation and Articulation Invariant Matching

Hao Jiang

Computer Science Department, Boston College, Chestnut Hill, MA 02467, USA

Abstract. A scale, rotation and articulation invariant method is proposed to match human subjects in images. Different from the widely used pictorial structure scheme, the proposed method directly matches body parts to image regions which are obtained from object independent proposals and successively merged superpixels. Body part region matching is formulated as a graph matching problem. We globally assign a body part candidate to each node on the model graph so that the overall configuration satisfies the spatial layout of a human body plan, part regions have small overlap, and the part coverage follows proper area ratios. The proposed graph model is non-tree and contains high order hyper-edges. We propose an efficient method that finds global optimal solution to the matching problem with a sequence of branch and bound procedures. The experiments show that the proposed method is able to handle arbitrary scale, rotation, articulation and match human subjects in cluttered images.

Keywords: Human pose, scale and rotation invariant matching, global optimization.

1 Introduction

Finding human subjects in cluttered images is a challenging task and it has many important potential applications. In this paper, we match a human subject in images and label the body part regions such as torso, arms and legs. The target object may have different scales and rotations. Most current pictorial structure approaches quantize the scale and rotation and optimize on the discrete cases. As the scaling range increases, searching through a huge number of discrete cases soon becomes impractical. The question is whether it is possible to efficiently match a human target without enumerating the quantized scales and rotations. In this paper, we address this problem and propose an efficient global optimization method that is able to match human subjects in images with unknown scale and rotation.

In contrast to the cardboard model that uses rectangle or polygon body parts, we match region candidates in images so that the combination of these regions forms a valid human body layout. The region candidates are from object independent proposals [19] and successively merged superpixels [18]. The proposed method assembles candidate regions and labels them as arm, leg and torso.

Fig. 1. We label human part regions in images by matching a graph model to the target human subject. The arm regions are red, legs are green and torsos are blue. The proposed matching method is scale, rotation and articulation invariant.

Different from pictorial structure methods [9, 2], the proposed method does not detect bar structures or obtain them from region candidates; instead it directly optimizes the region assembly. By directly working on part region candidates, our method is efficient and when properly constructed it is invariant to scale, rotation and object articulation. Fig. 1 illustrates matching human part regions using the proposed method.

Finding human poses in images has been intensively studied. If object foreground segmentation is available, poses can be estimated using regression and machine learning approaches [5]. In [22], object foreground proposals and latent structured models are used to find human poses. Other top-down methods detect poses by matching exemplars in databases [6–8]. These top-down pose estimation methods work best when poses are in a small domain. If poses are unconstrained, the performance of these methods degrades. Methods that reply on object foreground segmentation are also limited by the quality of figure-ground separation, which itself is a hard problem especially for segmenting human subjects.

Bottom-up pose estimation methods detect body parts and then assemble them into a human-shaped object. Pictorial structure model is widely used, in which arms, legs, torso and head are represented as rectangle or polygon patches. The coupling body parts form a graph model. Different methods have been proposed to optimize the body part assembly. Tree structure models [1–3, 17] allow efficient inference using dynamic programming. Non-tree models that include more constraints among body parts have also been intensively studied [10, 11, 4].

Part based methods also benefit from image segmentation. Object foreground segmentation helps part detection and pose verification [9]. In [4], part assembly is optimized as a max-cover to the object foreground. Even rough foreground estimation is found useful to improve pose estimation [17]. In [13, 12], part candidates (the parallel bars) are extracted from superpixel boundaries and then grouped into a stick figure. Superpixels have also been used in [14] to improve the pictorial structure methods. Joint foreground segmentation and pose estimation

for pedestrians have been studied in [16, 20]. In [21], object segmentation and graph matching are optimized together to achieve reliable unconstrained pose estimation.

The key obstacle for the pictorial structure methods is that it is hard to make the model adapt to the unknown scale of target objects. The body part assembly has to be optimized for each quantized scale and sometimes each rotation. This would be a slow process if we have to enumerate many discrete cases. Fitting rectangle structures to superpixel boundaries is able to make pose estimation scale invariant [13]; however, this procedure may lose detection of body parts. Apart from rectangle body parts, rectangle image patches (poselets) have also been used to match human subjects [23]. Poselet is not scale and rotation invariant. In this paper, we propose a method that directly matches regions. The body part regions are assembled so that the overall configuration fits a human body model. Such a scheme is scale, rotation and articulation invariant when properly constructed.

Grouping regions into a human shape is not a new concept. The jigsaw puzzle problem has been studied in [15], where the over-segmented superpixels are grouped together to fit a human model. Since superpixels are not able to group regions with different colors or textures, body parts with non-uniform appearance are often split into multiple superpixels. A parse tree method is proposed to merge superpixels in [15]. The parse tree may become huge and hard to process. As a compromise, a sequential procedure is applied: legs are first detected and then the torso is predicted from the leg detection using polygon matching. In [24], accurate body part region labeling has been achieved for pedestrians. This method replies on the shape priors of pedestrians and pedestrian detectors; it is thus hard to extend to matching people with arbitrary poses. Grouping a set of regions into a human shape and labeling the part regions is still an open problem.

The contribution of this paper is that we propose a global optimization method to match human body part regions. Our method groups superpixels [18] and region proposals [19] so that their spatial correlations and region ratios fit a human model. Our method is able to handle arbitrary human poses. It is scale and rotation invariant and can be globally optimized using a fast branch and bound approach.

2 Method

We treat human part matching as an assignment problem. We assign a candidate region to each body part so that the configuration follows the model constraints. Here the body part candidates are segments from successive superpixel merging and object independent region proposals. The body part model is shown in Fig. 2. The corresponding graph model has five nodes that represent torso, arms and legs. The hyper-edges linking the nodes indicate the torso-arms, torso-legs and arms-legs constraints. These constraints enforce the spatial layout, overlapping area, symmetry, size ratio, and overall region coverage. Given a set of

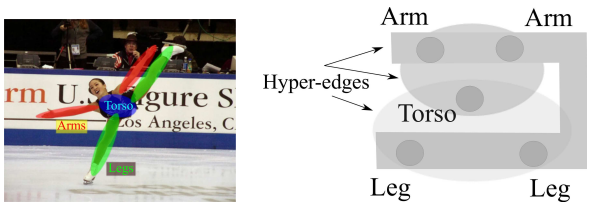


Fig. 2. Left: The 5-part body model. Right: The interaction of body parts in a graph. The graph includes five nodes and three hyper-edges among them.

candidate regions, we optimize the body part assignment on the graph model: each graph node selects a candidate region so that the following energy function is minimized.

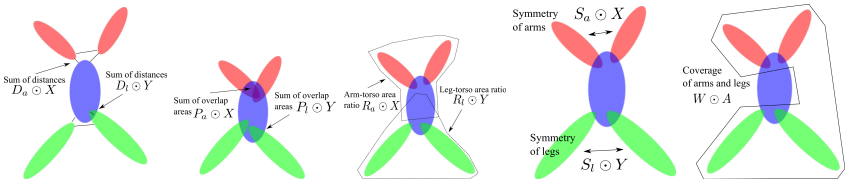
$$\min_{L,s}\{\mathcal{U}(L)+\alpha\mathcal{D}(L)+\beta\mathcal{P}(L)+\eta\mathcal{R}(L)+\gamma\mathcal{S}(L)+\mu\mathcal{W}(L,s)\} \qquad (1)$$

$$\text{s.t. } L \text{ is a valid part assignment, and } s \text{ is the scale estimation.}$$

where $\mathcal{U}(.)$ is the unary assignment cost, which is small if part candidate regions have similar shape to the corresponding part templates. $\mathcal{D}(.), \mathcal{P}(.), \mathcal{R}(.)$ and $\mathcal{S}(.)$ are tri-part terms, which are small if the labeling of arms-torso combination and the legs-torso combination satisfies specific constraints. $\mathcal{D}(.)$ quantifies the distance between specific body parts. $\mathcal{P}(.)$ penalizes selecting region candidates that are overlapping. $\mathcal{R}(.)$ enforces the relative sizes among body parts. $\mathcal{S}(L)$ encourages selecting regions with symmetrical appearance for arms or legs. $\mathcal{W}(.)$ is used to control the interaction among arms and legs, and encourages the overall coverage of arms and legs to fit a target size, i.e., the arms and legs do not overlap much and the overall area approaches a predicted value. During the body part region labeling, we estimate the scale s simultaneously. The coefficients of $\alpha, \beta, \eta, \gamma$ and μ control the weight among different terms. In this paper, we set $\eta = 0.1$, $\alpha = \gamma = 0.01$ and $\beta = \mu = 0.001$. The energy function is invariant to the scale, rotation and object articulation. Due to the loopy structure and high order terms, finding optimal body part region assignment is a challenging problem. In the following, we propose an efficient global optimal solution to this problem.

2.1 Finding Body Part Candidates

Before optimizing the body part configuration, we first find candidate regions for each body part. Different from the approach in [15], we do not merge small regions during the optimization, instead we select parts from a large set of candidate regions to form a human body assembly. The proposed method assumes that "correct" body part segments are in the candidate set. It is not necessary that separate arms or legs are detected; we allow merging of arms or legs into a single region. At first sight, this setting seems limited. However, we can almost



(a) Distance (b) Overlap (c) Area Ratio (d) Symmetry (e) Coverage

Fig. 3. The constraints on body parts and their notations.

always obtain roughly correct body part segments from object independent region proposals and progressively merged superpixels. Object independent region proposals [19] provide thousands of region candidates in an image by segmentation with randomly selected seed points and region pruning by object priors. This method works well to identify part regions on a human subject even when they are composed of sub-regions with different appearance. To further improve the chance of obtaining parts such as arms or legs, we also include candidate regions generated by progressively merging over-segmented superpixels. The merging process starts from fine superpixels [18] and then successively merges two neighboring superpixels with the most similar color histogram and the weakest boundary. With the object independent region proposals and successively merged superpixels, there is a high chance that the true body part segments are included in the candidate sets. Note that we do not require accurate part candidates; our method is robust when handling region merging and inaccurate candidates.

Given the region candidates, we solve a combinatorial search problem to assemble regions so that the overall configuration resembles a human subject. Naive exhaustive search is not feasible. We propose an efficient global optimization method.

2.2 The Formulation

We formulate the optimization in this section. The basic idea is to construct the optimization so that it can be linearized for fast solution.

We introduce some notations. We define an arm assignment tensor X and a leg assignment tensor Y. The arm tensor $X = [x_{i,j,k}]$ whose element $x_{i,j,k}$ indicates the assignment of region candidate i to arm one, region candidate j to arm two and candidate k to torso. And, similarly we define the leg tensor $Y = [y_{i,j,k}]$ to indicate the assignment of parts i, j, and k to leg one, leg two and torso respectively. The elements of X and Y are indicator variables whose values are either 0 or 1. In X or Y there is a single 1 element and every other element is 0. We also define a torso assignment vector $Z = [z_i]$, where $z_i = 1$ if torso selects candidate i, and otherwise $z_i = 0$.

The Unary Term: Each region candidate has a cost when assigned to a body part. We measure the shape similarity of each candidate region to the template. We use the inner distance [25] histogram to quantify the shape of a segment. The

shape descriptor is the histogram of the distance between each pair of points in a region. It can be efficiently computed using dynamic programming in $O(n^3)$ time, where n is the number of points in the region. The histogram has 20 bins in the range from 0 to the longest pairwise distance. We further normalize the histogram by the number of point pairs. The normalized inner distance histogram is scale and rotation invariant and roughly articulation invariant.

For each part p, e.g., arm, leg or torso, we have a set of exemplars $\{e_1, e_2, ...e_{k_p}\}$ in which e_i is the inner distance histogram of the ith template shape. The cost of the assignment of a candidate whose shape descriptor is h is defined as $\min_i ||h - e_i||$ where $||.||$ is the Euclidean distance. We build assignment cost tensor $U = [u_{i,j,k}]$ and $V = [v_{i,j,k}]$, where $u_{i,j,k} = a(i) + a(j)$ and $a(.)$ is the arm assignment cost for a candidate, $v_{i,j,k} = l(i) + l(j)$ and $l(.)$ is the leg assignment cost. The torso assignment cost vector is denoted as $T = [t_i]$, where t_i is the assignment cost of torso candidate i. In this paper, we keep the top 100 candidates for arm, leg and torso based on their local matching costs. The overall unary part assignment cost is

$$\mathcal{U} = U \odot X + V \odot Y + Z \odot T, \tag{2}$$

where $\odot$ is the operator to sum the product of corresponding tensor elements.
Distance Term: A valid body configuration requires that the chosen arm candidates and leg candidates should be adjacent to the selected torso candidate. Arms or legs also tend to be close to each other. The distance term is

$$\mathcal{D} = D_a \odot X + D_l \odot Y, \tag{3}$$

where D_a and D_l are distance tensors for arms and legs. $D_a = [d_{i,j,k}]$ where $d_{i,j,k} = d_{i,j} + d_{i,k} + d_{j,k}$, and we define $d_{i,j}$ as the distance between the closest points on the boundaries of arm candidate regions i and j, $d_{i,k}$ and $d_{j,k}$ are distances from arm candidates i and j to torso candidate k. Tensor D_l is similarly defined for legs. The shortest distances between region contours can be efficiently computed using the distance transform. The notations for the distance term are illustrated in Fig. 3(a).
Overlap Term: Simply minimizing the boundary distances among part regions does not guarantee a correct body part layout, since overlapping regions also have small boundary distances. We minimize the overlap between arms, legs, and torso:

$$\mathcal{P} = P_a \odot X + P_l \odot Y, \tag{4}$$

in which $P_a = [p_{i,j,k}]$ is an arm overlap tensor whose element $p_{i,j,k} = p_{i,j} + p_{i,k} + p_{j,k}$; $p_{i,j}$ is the overlapping area between arm candidate regions i and j, $p_{i,k}$ and $p_{j,k}$ are the overlapping areas of arm candidate regions i and j with torso region k. The leg overlap tensor P_l is similarly defined to penalize the overlap between legs, and between legs and torsos. The notations are illustrated in Fig. 3(b).
Size Ratio Term: A valid matching also should maintain correct size ratio between parts. The size ratio is also important for distinguishing arms and legs.

We enforce that the arm-torso ratio, leg-torso ratio, arm-arm ratio and leg-leg ratio conform to the priors. The ratio term is

$$\mathcal{R} = |R_{at} \odot X - r_{at}| + |R_{lt} \odot Y - r_{lt}| + |R_{aa} \odot X - r_{aa}| + |R_{ll} \odot Y - r_{ll}|, \quad (5)$$

where r_{at}, r_{aa}, r_{lt} and r_{ll} are the arm-torso, arm-arm, leg-torso and leg-leg region ratio priors, and $r_{aa} = r_{ll} = 1$. $R_{at} = [r^{(at)}_{i,j,k}]$ is the arm-torso ratio tensor and $r^{(at)}_{i,j,k} = (b_i + b_j)/b_k$ where b_i and b_j are the areas of arm candidate i and j, and b_k is the area of torso candidate k. The arm-arm ratio tensor $R_{aa} = [r^{(aa)}_{i,j,k}]$, where $r^{(aa)}_{i,j,k} = b_i/b_j$. The leg-torso ratio tensor R_{lt} and leg-leg ratio tensor R_{ll} are similarly defined. The notations are illustrated in Fig. 3(c). We use the L1 norm here so that we can linearize the ratio term by introducing auxiliary variables.

Symmetry Term: The arms and legs are symmetrical parts that usually have similar appearance. We minimize their histogram difference:

$$\mathcal{S} = S_a \odot X + S_l \odot Y, \quad (6)$$

where S_a and S_l are the symmetry tensors for arms and legs. We have $S_a = [s_{i,j,k}]$, $s_{i,j,k} = ||H_i - H_j||$, where H_i and H_j are the normalized color histograms of arm candidate regions i and j. S_l is similarly defined for the legs. When minimizing the symmetry term, we prefer to select arms and legs with similar appearance as shown in Fig. 3(d).

The Overall Coverage of Arms and Legs: The above terms do not explicitly constrain the layout of arms and legs. Without further constraints, the legs and arms may choose closely overlapping region candidates. Here we control their overall region coverage so that they should occupy a preferred region size. To this end, we find a set of "finer" segments so that all the region candidates can be represented as the union of these small units. In this paper, we use over-segmented superpixels as the unit regions. Let w_n be a variable to indicate whether unit region n is part of the object region and let $W = [w_n], n = 1..N$, where N is the number of unit regions. Let a be the total area of the template arm and leg regions and A be the vector of the areas of the unit regions, we minimize

$$\mathcal{W} = |sW \odot A - a| \quad (7)$$

Subject to:

$$w_n \leq 1, \;\; w_n \leq F_n \odot X + G_n \odot Y, \; n = 1..N$$
$$w_n \geq x_{i,j,k}, \; \forall f^{(n)}_{i,j,k} = 1, \; n = 1..N$$
$$w_n \geq y_{i,j,k}, \; \forall g^{(n)}_{i,j,k} = 1, \; n = 1..N$$

where F_n and G_n are 0-1 arm and leg mask tensors for unit region n. We define $F_n = [f^{(n)}_{i,j,k}]$ where $f^{(n)}_{i,j,k} = 1$ if arm candidate region i or region j covers unit region n; G_n is defined similarly. In such a setting, if an arm or a leg region covers unit region n, $w_n = 1$ and otherwise $w_n = 0$. Therefore, $W \odot A$ equals the total area of the region covered by the arms and legs. The coverage is scaled by s for scale invariance. Notations are illustrated in Fig. 3(e).

2.3 Linearization and Branch and Bound

Combining all the terms, we have a complete minimization problem. However, this optimization is still hard to solve due to huge number of variables and constraints. We decompose the optimization into slave linear programs corresponding to each torso candidate. Each of the sub-problems becomes much simpler and can be quickly solved. For a 3D tensor M whose last dimension is k, we denote $M^{(k)}$ as the kth slice of tensor M. For instance, $X^{(k)}$ and $Y^{(k)}$ indicate the arm and leg assignment given the torso selection k. We use such a notation for all the matrices including U, V, D, P, S, R, F and G. We also estimate the scale s by computing the ratio between the model torso area and the area of current torso candidate k; the scale estimation is denoted as $\hat{s}_k$. The linear optimization corresponding to torso region k is written as follows:

$$\begin{aligned}&\min\{(U^{(k)} + \alpha D_a^{(k)} + \beta P_a^{(k)} + \gamma S_a^{(k)}) \odot X^{(k)} + (V^{(k)} + \alpha D_l^{(k)} + \\ &\beta P_l^{(k)} + \gamma S_l^{(k)}) \odot Y^{(k)} + t_k + \eta(q_{aa} + q_{ll} + q_{at} + q_{lt}) + \mu(w^+ + w^-)\}\end{aligned} \tag{8}$$

Subject to:

$$\begin{aligned}&|X^{(k)}| = 1,\ |Y^{(k)}| = 1\\ &-q_{aa} \le R_{aa}^{(k)} \odot X^{(k)} - 1 \le q_{aa},\ -q_{ll} \le R_{ll}^{(k)} \odot Y^{(k)} - 1 \le q_{ll}\\ &-q_{at} \le R_{at}^{(k)} \odot X^{(k)} - r_{at} \le q_{at},\ -q_{lt} \le R_{lt}^{(k)} \odot Y^{(k)} - r_{lt} \le q_{lt}\\ &\hat{s}_k W \odot A - a = w^+ - w^-\\ &w_n \le 1,\ \ w_n \le F_n^{(k)} \odot X^{(k)} + G_n^{(k)} \odot Y^{(k)},\ n = 1..N\\ &w_n \ge x_{i,j,k},\ \forall f_{i,j,k}^{(n)} = 1,\ n = 1..N\\ &w_n \ge y_{i,j,k},\ \forall g_{i,j,k}^{(n)} = 1,\ n = 1..N\\ &\text{All the variables are non-negative, } X \text{ and } Y \text{ are binary.}\end{aligned}$$

Here $|X^{(k)}|$ and $|Y^{(k)}|$ denote the summation of all the elements in a matrix. t_k is the unary cost of torso candidate k. The nonnegative auxiliary variables q_{aa}, q_{ll}, q_{at}, q_{lt} equal the absolute value terms $|R_{aa}^{(k)} \odot X^{(k)} - 1|$, $|R_{ll}^{(k)} \odot Y^{(k)} - 1|$, $|R_{at}^{(k)} \odot X^{(k)} - r_{at}|$, $|R_{lt}^{(k)} \odot Y^{(k)} - r_{lt}|$ and $w^+ + w^-$ equals $|\hat{s}_k W \odot A - a|$, when the objective function is minimized. There are K slave mixed integer linear programs, each of which has K^2 arm and leg pairwise variables and N unit superpixel variables. In this paper $K = 100$ and and N is around 1000. We notice that when the torso selection is fixed, the only coupling between the arms and legs is the region overlapping constraints, which implies that each slave program can be solved quite efficiently.

We use branch and bound method to obtain the integer solution to each mixed integer slave program. Each slave program has the format $\min cu : Du = d$, where u includes the binary X and Y variables, and continuous w, q variables. We compute the lower bound by solving the relaxed linear program in which the binary constraints on X and Y variables are discarded. Any feasible integer

solution provides an upper bound, which can be initialized using the local best part matching.

New search tree branches are generated on the node with the smallest lower bound. We introduce integer cuts on the most fractional variable (the variable closest to 0.5). For the node with the lowest lower bound, a new cut $u_i = 0$ or $u_i = 1$ where u_i is either an X variable or a Y variable is included in the linear program. We do not have to solve each linear program from scratch, since there is only one more new constraint included in each branch and cut iteration. By introducing slack variables, $u_i = 0$ or equivalent $u_i \leq 0$ becomes $u_i + v_{i,0} = 0$, and $u_i = 1$ or equivalent $u_i \geq 1$ becomes $u_i - v_{i,1} = 1$ where $v_{i,0} \geq 0, v_{i,1} \geq 0$. u_i is a basic variable and its right hand side is a fractional number in the final simplex tabular. For the $u_i = 0$ branch, we subtract the original u_i row from $u_i + v_{i,0} = 0$, and for the $u_i = 1$ branch, we subtract $u_i - v_{i,1} = 1$ from the u_i row. In either case, we turn $v_{i,0}$ or $v_{i,1}$ into a basic variable that is not feasible because it has negative value on the right hand side. The dual-simplex method is then applied in pivoting and usually it takes very few steps to regain the optimal solution. We discard the branch whose linear program solution is infeasible or is greater than the current upper bound. Most of the branches are pruned quickly.

We keep track of the upper bound B_u and lower bound B_l of the solution. B_l is the lowest lower bound of all the active search tree nodes. Branch and bound can be terminated prematurely when the tolerance gap $\delta = \frac{2(B_u - B_l)}{(B_u + B_l)}$ is reached, and the objective is upper bounded by $(\delta+2)/(2-\delta)$ times the global minimum. In this paper, we terminate the iteration when $\delta \leq 10^{-3}$. After solving each slave program, the optimal solution of the original problem is the minimum of all the slave programs.

3 Experiment

An Example: Fig. 4 shows the example of matching a human subject using the proposed method. In this example, we generate about 1000 candidate regions. The local matching costs for the torso, leg and arm are shown in Fig. 4(b), (c) and (d), where brighter color indicates that a region is more likely to be a specific body part. The unary part cost is computed by matching the normalized inner distance histograms of the region candidates to those of the template shapes. Local matching is noisy and as shown in Fig. 4(e) a simple greedy method that selects the best match for each part does not give satisfactory result. The proposed method constructs a mixed integer program corresponding to each torso candidate. Here we keep the top 100 candidates for the torso, arm and leg. Our optimization yields much better result. The top 5 matching results are shown in Fig. 4 (f)-(j). The optimal matching accurately localizes the body parts in this example. The proposed method is also efficient; the optimization takes less than 10 seconds on a 2.8GHZ machine.

Proposal Regions and Object Foreground Segmentation: The region candidates from the object independent proposals and successively merged superpixels are not always able to give the overall human subject foreground. The

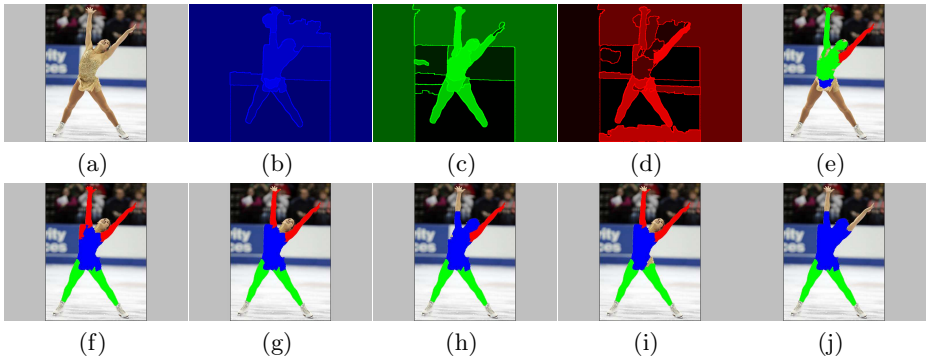


Fig. 4. A matching example using the proposed method. (a) Input image. (b), (c) and (d) show the local matching costs of the candidate regions to the torso, leg and arm templates (the brighter a segment, the more likely it is a corresponding body part). (f)-(j) show the top 5 matching results using the proposed method. (red, green and blue indicate arm, leg and torso regions respectively).

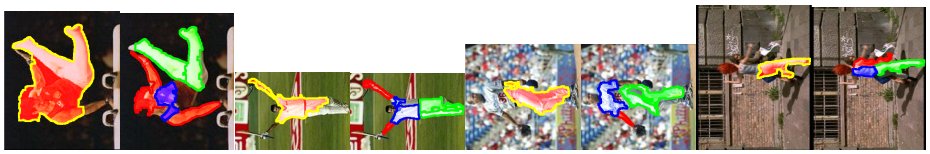


Fig. 5. Object foreground is not always in the region candidates. The odd number images show the closest region candidates to the object foreground. The proposed method uses smaller part candidates and is able to match the target reliably, as shown in the even number images.

sample test images in Fig. 5 are from the 305-image human pose dataset [2]. To make the matching problem more general, we resize the height of each image to 480 pixels so that the human subjects have different scales, and we rotate each image by 90 degrees. The best overall body segmentation from region candidates can be quite far from the ground truth as shown in the odd number images in Fig. 5. The proposed method is able to localize the target by using smaller part regions which are much easier to detect as shown in the even number images in Fig. 5.

Comparison with Competing Methods on Pose Dataset: We further compare the proposed method with competing methods. We first compare the proposed method with a greedy method that assigns the lowest cost candidate to the corresponding body part. The comparison is on the 305-image human pose dataset [2]. The images are scaled so that the height is 480 pixels. The scale factor is not determined due to a variety of image sizes in the dataset. Without loosing generality, we rotate all the images by 90 degrees and we assume that all the testing methods do not know the rotation angle. Due to the noisy local matching costs and lack of constraints among body parts, the simple greedy

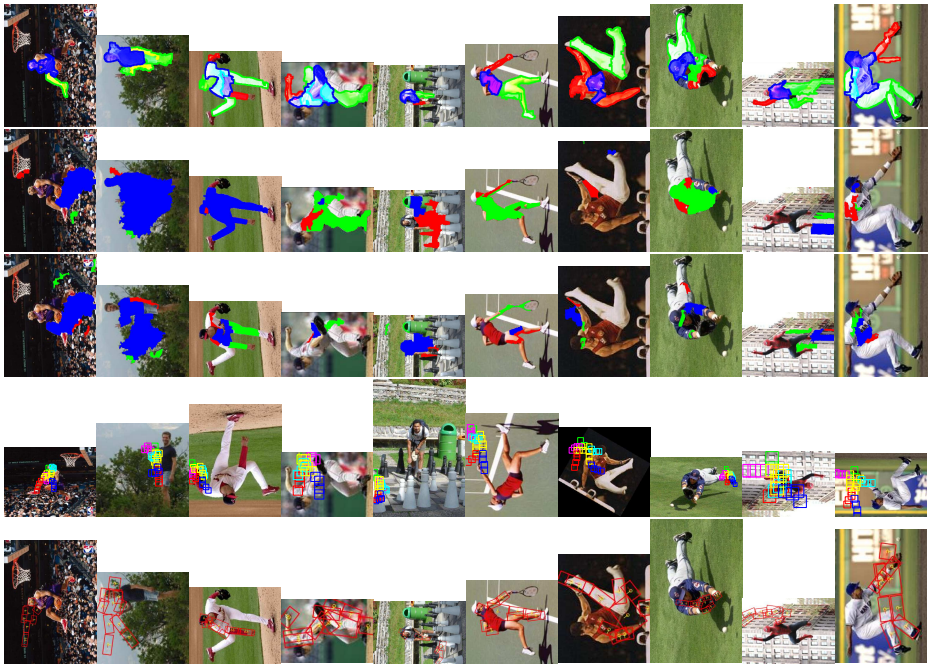
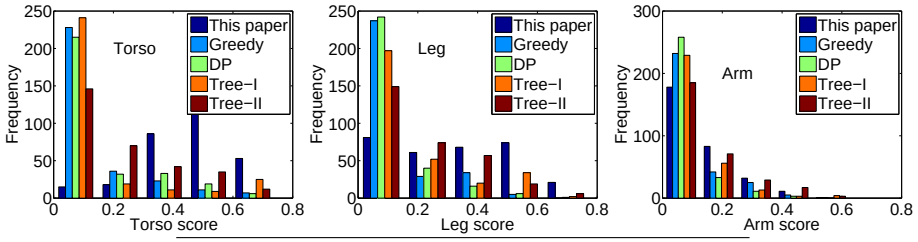


Fig. 6. Sample matching results of the proposed method (row 1), greedy method (row 2), Hough Transform based deformable matching (row 3), a recent people detector [2] (row 4) and 10-part pictorial structure method with strong part detector [3] (row 5).

approach gives poor results. Fig. 6 row 1 shows sample results of the proposed method, and Fig. 6 row 2 shows the matching results of the greedy method. The proposed method yields much better results. The quantitative comparison on all the images in the dataset is shown in Fig. 7. We define the matching score for a part as $|T \cap G|/|T \cup G|$, where T is the target part region, G is the ground truth region of the corresponding part, and $|.|$ computes the area of a region. In this paper, the ground truth regions are obtained from the ground truth joint labeling and by fitting a bar with suitable width to each body part segment. We compute the matching scores for the torso, arms and legs. The matching score is in $[0, 1]$ and the higher the matching score the better the matching; a perfect matching has the score of 1. The proposed method has much higher matching scores than the greedy method.

We compare the proposed method with a Hough Transform based method. In this method, we use a star structure model constrained by the global scale and rotation. The whole model is thus non-tree. The energy function is the linear combination of the unary matching cost, the pairwise matching cost, and the global scale and rotation consistency cost. The pairwise cost enforces the vector from the center of one part to the center of its neighbor part to conform to the model under some unknown rotation and scaling, and it also enforces that the



	This paper	Greedy	DP	Tree-I [2]	Tree-II [3]
Torso	0.4612	0.0979	0.1240	0.1058	0.2025
Leg	0.3205	0.0865	0.0788	0.1444	0.1954
Arm	0.1159	0.0728	0.0497	0.0764	0.1208

Fig. 7. Comparison with competing methods on the 305-image pose dataset [2]. Row 1 shows matching score distributions for torso, leg and arm. Row 2 gives average matching scores of different methods. Higher scores indicate better results.

area ratio of the part pairs follows the model. By quantizing scales and rotations, the optimization of the deformable matching turns into a sequence of dynamic programming on each scale and rotation. This matching method is essentially the extended Hough Transform in which the torso position is voted from all the part candidates. The final result is the matching with the lowest energy. We choose a stretch-out pose as the model spatial layout. As shown in Fig. 6 the dynamic programming (DP) approach gives results worse than the proposed method. The average matching scores and the matching score distributions shown in Fig. 7 confirm the advantage of the proposed method. The DP matching method is not able to handle large object articulations and therefore yields poor results for this dataset.

We compare the proposed method with a recent human detector [2] and a pictorial structure method using strong part detectors [3]. The method in [2] is not rotation invariant. We thus rotate each input image from 0 to 360 degrees with 24 steps, and we select the result with the best matching score. Fig. 6 row 4 shows sample matching results of the people detector. The proposed method greatly improves the result. Generating the foreground part segmentation by connecting joint detections of the pictorial structure method [2] and thickening the lines, we can use the region ratio metric to quantitatively measure the matching performance. The ratio of line thickening uses the same scheme as the one in ground truth region generation, i.e., a perfect matching would give a score of 1 for each part. Fig. 7 compares the matching scores between the proposed method and [2]. The proposed method has much better performance. Another pictorial structure method [3] that uses strong local part detectors is further compared with the proposed method. This method operates on discrete scales from 1 to 5 with 10 steps and 24 rotation angles. The pictorial structure method takes about 20 minutes to process each image, while the proposed method takes about 10 seconds in the optimization (the candidate region generation takes

Fig. 8. Sample results of the proposed methods on the human pose dataset [2]. Arm regions are red, legs are green and torsos are blue. The test images are scaled from 1 to 5 and rotated by 90 degrees. The results are rotated back to the normal position and rescaled.

about 60 seconds per image). The comparison is shown in Fig. 6 and Fig. 7. The proposed method has much higher detection scores for the torso and legs than [3] and the arm detection score is comparable with [3]. More sample results of the proposed methods are shown in Fig. 8.

4 Conclusion

We propose an efficient method to localize human subject in images by matching body part region proposals. The proposed linearization scheme and branch and bound approach are able to give global optimal result efficiently. The proposed method is scale, rotation and articulation invariant. It has a clear advantage over competing methods when the target human subject has unknown scale and rotation. The proposed method will be useful for many different applications including human detection, tracking and action recognition.

Acknowledgments. This research is supported by US NSF grant 1018641.

References

1. Ramanan, D.: Learning to Parse Images of Articulated Objects. NIPS 2006.
2. Yang, Y., Ramanan, D.: Articulated Pose Estimation Using Flexible Mixtures of Parts. CVPR 2011.
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. CVPR 2009.
4. Jiang, H.: Human Pose Estimation Using Consistent Max-Covering. ICCV 2009.
5. Rosales, R., Sclaroff, S.: Inferring Body Pose without Tracking Body Parts. CVPR 2000.
6. Mori, G., Malik, J.: Estimating Human Body Configurations Using Shape Context Matching. ECCV 2002.
7. Gavrila, D.M.: A Bayesian, Exemplar-based Approach to Hierarchical Shape Matching. TPAMI, 29(8), 2007.
8. Shakhnarovich, G., Viola, P., Darrell, T.: Fast Pose Estimation with Parameter Sensitive Hashing. ICCV 2003.
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. IJCV 61(1), Jan. 2005.
10. Sigal, L., Black, M.J.: Measure Locally, Reason Globally: Occlusion Sensitive Articulated Pose Estimation. CVPR 2006.
11. Tian, T.P., Sclaroff, S.: Fast Globally Optimal 2D Human Detection with Loopy Graph Models. CVPR 2010.
12. Mori, G., Guiding Model Search Using Segmentation. ICCV 2005.
13. Ren, X.F., Berg, A.C., Malik, J.: Recovering Human Body Configurations Using Pairwise Constraints between Parts. ICCV 2005, 1:824-831.
14. Sapp, B., Jordan, C. and Taskar, B.: Adaptive Pose Priors for Pictorial Structures. CVPR 2011.
15. Cour T. and Shi J., Recognizing Objects by Piecing Together The Segmentation Puzzle. CVPR 2007.
16. Kohli, P., Rihan, J., Bray, M., Torr, P.H.S.: Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts. IJCV, vol.79, no.3 pp.285-298, 2008.
17. Ferrari, V., Manuel, M., Zisserman, A.: Pose Search: Retrieving People Using Their Pose. CVPR 2008.
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-based Image Segmentation. IJCV, vol.59, no.2, 2004.
19. Endres, I. and Hoiem, D.: Category Independent Object Proposals. ECCV 2010.
20. Chen, C., Fan, G.: Hybrid Body Representation for Integrated Pose Recognition, Localization and Segmentation. CVPR 2008.
21. Wang H. and Koller D.: Multi-Level Inference by Relaxed Dual Decomposition for Human Pose Segmentation. CVPR 2011.
22. Ionescu, C., Li, F., Sminchisescu, C.: Latent Structured Models for Human Pose Estimation. ICCV 2011.
23. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting People Using Mutually Consistent Poselet Activations, ECCV 2010.
24. Bo, Y., Fowlkes, C.: Shape-based Pedestrian Parsing, CVPR 2011.
25. Ling, H. and Jacobs, D. W.: Shape Classification Using the Inner-Distance, IEEE Trans. on Pattern Anal. and Mach. Intell., 29(2):286-299, 2007.