Foreground Consistent Human Pose Estimation Using Branch and Bound*

Jens Puwein[1], Luca Ballan[1], Remo Ziegler[2], and Marc Pollefeys[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Vizrt

Abstract. We propose a method for human pose estimation which extends common unary and pairwise terms of graphical models with a global foreground term. Given knowledge of per pixel foreground, a pose should not only be plausible according to the graphical model but also explain the foreground well.

However, while inference on a standard tree-structured graphical model for pose estimation can be computed easily and very efficiently using dynamic programming, this no longer holds when the global foreground term is added to the problem.

We therefore propose a branch and bound based algorithm to retrieve the globally optimal solution to our pose estimation problem. To keep inference tractable and avoid the obvious combinatorial explosion, we propose upper bounds allowing for an intelligent exploration of the solution space.

We evaluated our method on several publicly available datasets, showing the benefits of our method.

1 Introduction

Single image human pose estimation has received a lot of attention over the past few years. The goal is to localize each body part of a human body in a given image. This allows for a higher level of understanding of the image itself and, potentially, it can be used to facilitate other complementary computer vision tasks like image segmentation, 3D reconstruction, activity recognition, and image retrieval.

In this paper, we aim at estimating the 2D locations of all the joints of a human body under uncontrolled imaging conditions. A common approach to this problem is to use tree-structured graphical models to represent the human pose as a set of joints, or as a set of limbs, linked by edges representing bones or kinematic constraints between limbs, respectively [4,25].

While inference in these models can be carried out very efficiently using dynamic programming [25], they lack the possibility of considering global information depending on all the body parts at the same time. An example of such a scenario is when per pixel foreground probabilities of the given image are available. To account for this additional information, the pose should not only be

* Electronic supplementary material - Supplementary material is available in the online version of this chapter at `http://dx.doi.org/10.1007/978-3-319-10602-1_21`. Videos can also be accessed at `http://www.springerimages.com/videos/978-3-319-10601-4`

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 315–330, 2014.

plausible with respect to the graphical model, but also explain the foreground. Therefore a more complex inference model needs to be used, which besides the common unary and pairwise terms also includes a global foreground term favoring solutions explaining the given foreground information.

In order to guarantee the global optimality of our solutions, we propose to optimize the model using a branch and bound based algorithm [10]. However, despite the fact that branch and bound intelligently explores only the promising regions of the solution space, it can be prohibitively slow since this space is actually the Cartesian product of all the domains of each individual unknown, and hence it grows exponentially with the number of unknowns. To keep our inference computationally tractable, we propose a set of upper bounds specifically designed for our pose estimation problem, and a way to decouple the estimation of rarely overlapping limbs while still maintaining the global optimality.

The performance of the proposed method was evaluated on four publicly available datasets (KTH, Parse, Leeds and Buffy [24,25,6,5]), showing the potential improvements achieved by our method.

2 Related Work

One of the most common and efficient ways of estimating a human pose from a single image is to formulate the problem as an inference on a tree-structured graphical model, where nodes express the position, the orientation, and the scale of each body limb [4] or the position of each body joint [25], and where edges between nodes correspond to kinematic constraints between limbs or to bones between joints. While inference on such models can be performed exactly and very efficiently using dynamic programming, it fails to capture some higher level dependencies that can occur between limbs, for instance when these are overlapping in the image space. This problem has been addressed by including occlusion terms [11,18] or repulsive edges [2]. A solution can be obtained through Gibbs sampling [11] or by using a loopy graphical model and loopy belief propagation [18,2]. Loopy models can also be expressed as an ensemble of tree-structured models by enforcing the equality of corresponding nodes in the different trees, as proposed by Sapp *et al.* [17]. In their work, different levels of agreement are proposed. For the full agreement between all submodels, convergence is not guaranteed. Another way of dealing with loopy models is branch and bound, which leads to a globally optimal solution and has been shown to be efficient [19,22,20]. Going beyond local reasoning, Kohli *et al.* [7] simultaneously solve for human pose and segmentation using dynamic graph cuts. Their algorithm, however, is susceptible to local minima and requires a good initialization of the pose. To solve for a model including global terms, sampling techniques are a popular choice. Zhang *et al.* [26] propose a data-driven Markov Chain Monte Carlo framework using a tree-based grammar to explore the space of human poses, trying to globally explain the foreground regions and the edges as well as possible while trying to fulfill body constraints. Similarly, Rauschert and Collins [15] use a data-driven, coarse-to-fine Metropolis Hastings sampling scheme also incorporating the likelihood of all image pixels and the domain knowledge in the proposal function.

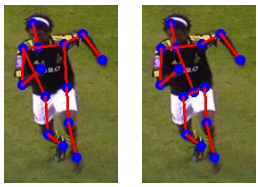
Instead of relying on the maximum a posteriori (MAP) estimate of a tree structured model, Park and Ramanan [12] propose to infer the N-best solutions according to a tree structured model. Assuming that the correct pose is among the N best, further more expensive processing is applied to determine the correct solution. Starting from this idea, Vahid and Sullivan [24] extract the N-best poses and rerank them using an SVM-Rank formulation including a global segmentation term. Along the same paradigm, Ladicky *et al.* [8] introduce a joint pixel-wise and part-wise formulation. First, poses from the set of the N-best poses are added iteratively as long as it decreases the cost. Then, for all the added poses, each pixel is assigned to a person and to a body part. Their approach can deal with multiple people and missing/occluded body parts.

The approach presented in this paper builds on the model introduced by Yang and Ramanan [25]. However, please note that our method is not constrained to this model and any other efficient graphical model can be easily used in its place. A global segmentation term is added to this model, similarly to the work of Vahid and Sullivan [24]. However, differently from their approach, we propose to rely on a branch and bound optimization technique to avoid the premature selection of the N-best solutions, and hence to guarantee the optimal solution. The guarantee of global optimality differentiates our approach from sampling based techniques like [26,15].

3 Our Approach

3.1 Standard Model

We build upon the tree structured graphical model introduced by Yang and Ramanan [25], consisting of 14 joints as shown in Figure 1(a). Differently from them however, we consider an additional joint for the lower end of the spine between the left and the right hip joint, and define the torso as the body part identified by this new joint and the bottom of the head. Together with the edges of the human kinematic chain, these 15 joints, depicted in Figure 1(b), define a tree $G = (V, E)$ with nodes V and edges E. For each joint $i \in V$, let l_i identify its (x, y)-position in the image space, and let $t_i \in \{1, ..., T_i\}$ denote its type. The type essentially captures the relative orientation of a joint with respect to its parent in the tree model. Different orientations lead to different appearances. The appearance of each type t_i is modeled using a HOG descriptor [1], describing the distribution of image gradients in a local region. Let $\phi(I, l_i, t_i)$ denote the descriptor of joint i with type t_i extracted at location l_i in image I. Pairwise costs are given by a deformation model favoring frequently encountered relative positions of connected parts i and j. The corresponding feature vector $\psi(l_i, l_j)$ is given as $[(x_i - x_j), (x_i - x_j)^2, (y_i - y_j), (y_i - y_j)^2]^T$, encoding the differences in x- and y-coordinates, respectively. $\psi(l_i, l_j)$ is weighted differently for each type, providing a link between the appearance of a part and its relative location w.r.t. to its parent. Bias terms $b_i^{t_i}$ and $b_{i,j}^{t_i,t_j}$ capture the probabilities of encountering specific parts and types and pairs of parts and types. After adding the bias



(a) (b)

Fig. 1. (a) The standard model with 14 joints. (b) Our model with 15 joints.

terms and weighting the feature vectors from the appearance model ϕ and the deformation model ψ, we end up with the following scoring function S:

$$S(I,l,t) = \sum_{i \in V} \left(b_i^{t_i} + \phi(I,l_i,t_i)^T w_i^{t_i} \right) + \\ + \sum_{i,j \in E} \left(b_{i,j}^{t_i,t_j} + \psi(l_i,l_j)^T w_{i,j}^{t_i,t_j} \right) \tag{1}$$

Concatenating all the individual weight vectors and bias terms as w and subsuming all feature vectors within $\Phi(I,l,t)$, an equivalent linear model $S(I,l,t) = \Phi(I,l,t)^T w$ is obtained. Parameters w are learned using structured support vector machine [21,23]. For more details, please refer to [25].

3.2 Augmented Model

The scoring function $S(I,l,t)$ captures the local appearance and the deformations of a generic human body. However, in many scenarios, a good guess of the foreground shape of the body can be obtained from an image, for instance, through global color models, background subtraction or image/video matting. It is then desirable to find the pose which, globally, best accounts for the foreground shape, and is also plausible and faithful w.r.t. the scoring function $S(I,l,t)$.

Hence, given a per-pixel foreground estimate $FG(p)$, the pose should not only have a high score $S(I,l,t)$, but also explain foreground regions in $FG(p)$ as much as possible. To this end, we introduce a generative model Ω mapping joint positions l to sets of image points $\Omega(l)$ representing the human body silhouette in the image space in that specific pose. Each body part $(i,j) \in E$ is modeled as a rectangle $R(l_i,l_j)$ of predefined width, as illustrated in Figure 2(e). The silhouette $\Omega(l)$ is therefore defined as the union of all these rectangles, i.e. as $\Omega(l) = \bigcup_{i,j \in E} R(l_i,l_j)$. Using this generative model, the previously described scoring function $S(I,l,t)$ is augmented with a global foreground term $F(l)$ defined as

$$F(l) = \sum_{p \in \Omega(l)} FG(p), \tag{2}$$

where $FG(p) \in [0,1]$ is the per-pixel foreground estimate evaluated on a given pixel p, and it indicates the confidence value that that pixel belongs to the foreground.

Algorithm 1. Branch and bound inference

push pair $(\bar{E}(\mathcal{H}_0), \mathcal{H}_0)$ into queue and set $\hat{\mathcal{H}} = \mathcal{H}_0$
repeat
  split $\hat{\mathcal{H}} = \hat{\mathcal{H}}_1 \cup \hat{\mathcal{H}}_2$ with $\hat{\mathcal{H}}_1 \cap \hat{\mathcal{H}}_2 = \emptyset$
  push pair $(\bar{E}(\hat{\mathcal{H}}_1), \hat{\mathcal{H}}_1)$ into queue
  push pair $(\bar{E}(\hat{\mathcal{H}}_2), \hat{\mathcal{H}}_2)$ into queue
  pop $\hat{\mathcal{H}}$ with the highest score
until $|\hat{\mathcal{H}}| = 1$

The goal is now to maximize the following scoring function

$$\arg\max_{l,\ t} E(I, l, t) = S(I, l, t) + \lambda F(l), \tag{3}$$

where λ is a constant weighting the global foreground term w.r.t. $S(I, l, t)$.

Pose estimation aims at fitting a model, which typically has a predefined number of parts, into an image. Therefore, placing a part at a wrong location means that the foreground region which actually corresponds to that part is likely to not be explained (if that part does not overlap with another part), and hence lowering the overall score. In the absence of false foreground regions, regions wrongly labeled as background should not bias the model towards wrong solutions. However, false foreground regions might induce errors, since covering such a false region with a body part can increase the score E. This fact can be mitigated by using a conservative foreground mask.

3.3 Optimization

While inference on tree structured graphical models, like the one in Equation 1, can be performed very efficiently using dynamic programming, this no longer holds when a global term considering all the joints at the same time is added. This is the case with term $F(l)$ in Equation 3. Therefore, to optimize the new problem, a different optimization technique is required.

To this aim, we propose to use branch and bound on the set of possible joints configurations l, inspired by the work of Sun *et al.* [20], who applied branch and bound to loopy graphical models, and also inspired by Lampert *et al.* [9], who applied branch and bound to subwindow search. Apart from its generality, one of the advantages of branch and bound is that it guarantees to find the globally optimal solution.

We now describe how branch and bound is employed for our problem. The algorithm starts with the trivial set $\mathcal{H}_0$ defined as the set of all possible joint configurations hypotheses, *i.e.*$\mathcal{H}_0 = \prod_{i=1}^{15}\{1, ..., w_{image}\} \times \{1, ..., h_{image}\}$, the Cartesian product of the possible (x, y)-positions of all joints. Throughout the branch and bound iterations a priority queue is maintained where the considered sets of hypotheses are ordered in terms of a quality bound function $\bar{E}$ which upper bounds the maximum score E that any pose of a given set can possibly achieve. The best candidate $\hat{\mathcal{H}}$ of all the sets within the queue is considered

for further processing. If $\hat{\mathcal{H}}$ consists of a single hypothesis, then the optimum is obtained. Otherwise, the set is split into two disjoint sets of hypotheses $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$. Different branching strategies exist. We use a very simple strategy and split the hypotheses by splitting the largest remaining image coordinate interval of all joints in half. The new bounds $\bar{E}(\hat{\mathcal{H}}_1)$ and $\bar{E}(\hat{\mathcal{H}}_2)$ for those sets are computed, and both candidate sets are added to the priority queue. Since the bounds are tighter (smaller sets of hypotheses), it may be that none of these sets will be on top of the priority queue. The algorithm terminates when a single hypothesis is returned, and this hypothesis is guaranteed to be the global optimum. The advantage of using branch and bound is that it does not explore regions of the solution space which are not promising. The reader is referred to Algorithm 1 for a schematic illustration.

In order to guarantee the convergence of the branch and bound algorithm to the globally optimal solution, the quality bound function $\bar{E}$ needs to satisfy the following two conditions:

1. None of the hypotheses in $\mathcal{H}$ can achieve a higher score than $\bar{E}(\mathcal{H})$. More precisely, for each joint configuration $l \in \mathcal{H}$, and each type configuration t, $\bar{E}(\mathcal{H}) \geq E(I, l, t)$ has to hold.
2. If the set of hypotheses contains a single configuration, the bound has to be exact. More precisely, for each $l \in \mathcal{H}_0$, $\bar{E}(\{l\}) = \max_t E(I, l, t)$ has to hold.

3.4 Quality Bound Function $\bar{E}$

A valid quality bound function $\bar{E}$ can be defined in terms of multiple upper bounds $\bar{E}_i$ by always selecting the smallest value $\bar{E}_i$, i.e. $\bar{E} = \min_i \bar{E}_i$. It is easy to see that condition 1 is satisfied if every upper bound $\bar{E}_i$ satisfies condition 1, while condition 2 is satisfied if at least one of the bounds $\bar{E}_i$ satisfies condition 2. In the following sections, two different upper bounds $\bar{E}_1$ and $\bar{E}_2$ are introduced, each having its own advantages and disadvantages depending on the set of hypotheses being bounded. We combine these two upper bounds as $\bar{E} = \min(\bar{E}_1, \bar{E}_2)$.

Upper Bound $\bar{E}_1$. Let us first consider an alternative global foreground term $\tilde{F}$ defined as

$$\tilde{F} = \sum_{i,j \in E} Seg(l_i, l_j), \tag{4}$$

where each pairwise score $Seg(l_i, l_j)$ is defined as $\sum_{p \in R(l_i, l_j)} FG(p)$. Adding this new term $\tilde{F}$ to the scoring function $S(I, l, t)$ leads to the following scoring function

$$E_{pairwise}(I, l, t) = S(I, l, t) + \lambda \tilde{F}(l). \tag{5}$$

Differently from the original foreground term, the new $\tilde{F}$ maintains the tree structure of $S(I, l, t)$. Therefore inference on $E_{pairwise}(I, l, t)$ can be performed efficiently using dynamic programming.

However, $\tilde{F}$ counts foreground pixels multiple times if body parts overlap, and therefore we conclude that

$$E_{pairwise}(I, l, t) \geq E(I, l, t) \tag{6}$$

for every pose (l, t). In Equation 6, equality holds if and only if none of the rectangles defining the foreground silhouette overlap.

Hence, an upper bound for $E(I, l, t)$ can be defined as

$$\bar{E}_1(\mathcal{H}) = \max_{l \in \mathcal{H},\ t} E_{pairwise}(I, l, t). \tag{7}$$

Due to the underlying tree structure, $\bar{E}_1(\mathcal{H})$ can be computed very efficiently by constraining the dynamic programming to the configurations in $\mathcal{H}$. In order to quickly evaluate $Seg(l_i, l_j)$ one can resort to integral images computed for rotated versions of the original foreground map $FG(p)$. In this way, each body part rectangle becomes an axis-aligned rectangle in the respective rotated foreground map. Integrals can then be evaluated using lookups in the corresponding integral images. In our implementation, integral image angles were quantized to steps of one degree.

Upper Bound $\bar{E}_2$. Since the new foreground term $\tilde{F}$ introduced in the previous section may count image foreground evidence multiple times, there is no guarantee that condition 2 holds in general for the upper bound $\bar{E}_1$. We therefore introduce a second upper bound $\bar{E}_2$ as follows.

Given the current set of hypotheses $\mathcal{H}$, a conservative estimate of the body silhouette is given by

$$\bar{\Omega}(\mathcal{H}) = \bigcup_{l \in \mathcal{H}} \Omega(l), \tag{8}$$

which equals to the union of all the silhouettes corresponding to each individual pose hypothesis. Hence, an upper bound for the original $F(l)$ is

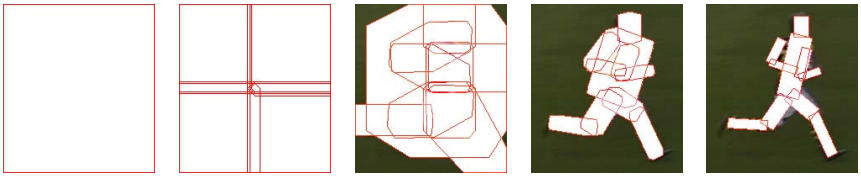
$$\bar{F}(\mathcal{H}) = \sum_{p \in \bar{\Omega}(\mathcal{H})} FG(p). \tag{9}$$

Finally, we define the upper bound $\bar{E}_2$ as

$$\bar{E}_2(\mathcal{H}) = \bar{S}(I, \mathcal{H}) + \lambda \bar{F}(\mathcal{H}), \tag{10}$$

where $\bar{S}(I, \mathcal{H})$ is the maximum value achievable by $S(I, l, t)$ in $\mathcal{H}$, i.e., $\bar{S}(I, \mathcal{H}) = \max_{l \in \mathcal{H}, t} S(I, l, t)$. Due to the tree structured nature of $S(I, l, t)$, $\bar{S}(I, \mathcal{H})$ can be computed efficiently using dynamic programming on the set of hypotheses $\mathcal{H}$.

Note that the new upper bound $\bar{E}_2$ fulfills both condition 1 and 2 of the branch and bound algorithm.



(a) (b) (c) (d) (e)

Fig. 2. Branch and bound: the algorithm iteratively narrows down the search space. White pixels indicate the set $\bar{\Omega}(\mathcal{H})$ of Equation 8. In the initial set of hypotheses, every joint can lie anywhere in the image (a). Once branch and bound terminates, the set of hypotheses corresponds to a single pose (e).

Combining $\bar{E}_1$ and $\bar{E}_2$. $\bar{E}_1$ and $\bar{E}_2$ are combined to form the upper bound $\bar{E} = \min(\bar{E}_1, \bar{E}_2)$. $\bar{E}$ fulfills conditions 1 and 2 of the branch and bound algorithm because both $\bar{E}_1$ and $\bar{E}_2$ fulfill condition 1 and $\bar{E}_2$ fulfills condition 2. While $\bar{E}_2$ alone would be sufficient in theory, $\bar{E}_1$ should be included in practice to decrease the computational complexity. Branch and bound terminates once the currently chosen set of hypotheses $\hat{\mathcal{H}}$ is of size one. Note that at this point, the upper bound equals the lower bound since they are both equal to the cost of the single remaining pose. The faster the upper bound decreases, the faster branch and bound terminates.

During the first branch and bound iterations, the chosen sets of hypotheses are large and lead to high values of $\bar{F}$. In these cases, $\bar{E}_1$ provides a much tighter bound than $\bar{E}_2$, and this holds until the double counting in $\bar{E}_1$ leads to $\bar{E}_1 > \bar{E}_2$. Therefore, at the beginning $\bar{E}_1$ quickly guides the branch and bound to a reasonable set of poses, and then $\bar{E}_2$ is active instead. An example of a branch and bound evolution is provided in Figure 2.

3.5 Efficient Inference

Sequential Branch and Bound. An alternative way to optimize $E(I, l, t)$ which does not guarantee global optimality is to apply the branch and bound algorithm sequentially on the tree structure. More precisely, the torso and the head are first inferred jointly (see Figure 3(c)). Subsequently, head and torso are kept fixed, and the legs are inferred (see Figure 3(d)). Finally, head, torso and legs are kept fixed and the configuration of the arms is inferred (see Figure 3(e)). The whole process is summarized in Figure 3. The results obtained in our experiments (and reported in Section 4) suggest that the detection of the torso is the most reliable by far, and correct with a high probability. This justifies estimating the torso and the head first, followed by legs and arms. Legs and arms only seldom interfere with each other, suggesting that, given the torso and the head, their configuration may be inferred correctly using sequential branch and bound.

Decoupling of States. The observations made in the previous section lead us to consider an additional expedient to speed up the branch and bound algorithm

(a) (b) (c) (d) (e)

Fig. 3. Sequential branch and bound: (a) Input image. (b) Foreground probabilities $FG(p)$. In the sequential algorithm, torso and head are estimated first (c), followed by legs (d), and arms (e).

in a way that the global optimality property is preserved. As explained before, the complexity of our problem is caused by the global foreground term $F(l)$ which links all the body parts together, and makes them dependent on each other because of possible overlaps in the image space.

In a realistic scenario, however, not all the limbs overlap, and some of them do it very rarely. This is the case for head, arms and legs. Therefore, in most of the cases, it makes perfect sense to consider the head and the arms independently from the legs. We can exploit this natural characteristic of our solution space to speed up our global optimization. Basically, given a torso location, a very tight upper bound can be obtained by running branch and bound on the configurations of head and arms, and legs independently.

Therefore we first compute a lower bound for the scoring function $E(I, l, t)$ using the sequential approach described in the previous section. Then, for each torso location an upper bound is computed by maximizing $E_{pairwise}(I, l, t)$. Torso locations leading to upper bounds below the current lower bound can be safely discarded. For each remaining location of the torso a set of hypotheses is created and added to the priority queue. The already fixed torso in each set of hypotheses decouples upper and lower body to a large extent.

4 Results

The proposed method was tested on four publicly available datasets, namely: the KTH dataset [24], the Parse dataset [25], the Leeds dataset [6] and the Buffy dataset [5]. The KTH dataset consists of 771 images, where the first 180 images were used for training and the remaining 591 images were used for testing. The images show football players in different poses commonly observable in TV broadcasts. The Parse dataset instead consists of 305 images, where the first 100 images were used for training and the remaining 205 images were used for testing. The images show a wide variety of poses in unconstrained outdoor settings, similar to the Leeds dataset, which consists of 1000 training images and 1000 test images. The Buffy dataset is limited to the upper body and shows scenes from different episodes of the TV show 'Buffy'; 472 images were used for training and 276 for testing.

Table 1. Comparison of the percentage of correctly estimated body parts (strict PCP) on the **KTH** dataset

Method	Head	Torso	U. Arms	L. Arms	U. Legs	L. Legs	**Total**
[25], 26 parts	91.2	99.7	87.2	60.7	85.0	73.3	80.3
[25], 29 parts	91.7	**99.8**	85.2	62.8	85.8	73.9	80.7
[24]	91.7	99.7	**87.8**	**63.4**	**91.5**	80.0	**83.7**
Sequential BB (26 parts)	90.9	99.7	87.1	62.2	**91.5**	**81.1**	83.4
Sequential BB (29 parts)	91.5	**99.8**	84.8	59.7	90.1	79.3	81.9
Global BB (29 parts)	**92.2**	**99.8**	84.2	61.7	91.4	80.2	82.7

Table 2. Comparison of the percentage of correctly estimated body parts (strict PCP) on the **Parse** dataset

Method	Head	Torso	U. Arms	L. Arms	U. Legs	L. Legs	**Total**
[25], 26 parts	84.9	89.8	61.5	39.8	75.4	68.0	66.4
[25], 29 parts	84.9	87.8	59.0	36.8	77.6	70.5	66.0
[8]	75.1	83.9	56.8	33.9	71.0	63.9	61.0
[14]	**86.3**	**93.2**	**63.4**	**48.8**	77.1	68.0	**69.4**
Sequential BB	83.4	86.3	60.5	38.8	79.8	72.7	67.3
Global BB	**86.3**	92.7	59.8	40.0	**81.0**	**73.4**	68.7

Table 3. Comparison of the percentage of correctly estimated body parts (loose PCP) on the **Buffy** dataset

Method	Head	Torso	U. Arms	L. Arms	**Total**
[25], 21 parts	97.5	97.8	93.1	66.0	85.6
[8]	**100.0**	**100.0**	**97.5**	**75.4**	**90.9**
Global BB	**100.0**	**100.0**	95.6	71.5	89.0

Table 4. Comparison of the percentage of correctly estimated body parts (strict PCP) on the **Leeds** dataset

Method	Head	Torso	U. Arms	L. Arms	U. Legs	L. Legs	**Total**
[25], 29 parts, 12 types	80.1	84.8	54.0	38.0	71.5	66.5	62.5
[3]	80.1	86.5	56.5	37.4	74.9	69.4	64.3
[14]	**85.6**	**88.7**	**61.5**	**44.9**	**78.8**	**73.4**	**69.2**
Global BB, 29 parts, 12 types	80.0	86.6	53.8	38.8	75.4	70.0	64.3

Each test image is first pre-processed in order to estimate the per-pixel foreground confidence map $FG(p)$. To this aim, the standard tree-structured model of Yang and Ramanan [25] is used to retrieve an estimate of the pose. Subsequently, a mask around this estimate is created by dilating the convex hull of the estimated joints positions. In the end, grabcut [16] is initialized by this mask and used to obtain the foreground map $FG(p)$. Note that in many scenarios where information about foreground and/or background is given a priori, a better

segmentation can be obtained. For the Buffy dataset, segmentations provided by [8] were used.

To increase the expressiveness of the original model $S(I, l, t)$ in Equation 1, 14 auxiliary joints are added to the graph G. This is coherent to what is also done in [25], and the main purpose is to provide appearance models also to the central part of each body limb. We do not branch on these additional unknowns. On the contrary, branch and bound is still run only on the (2x15)-dimensional solution space described before, and the positions of these auxiliary joints are estimated during the maximization procedures in $\bar{E}_1$ and $\bar{E}_2$.

To quantitatively evaluate the results obtained using our approach, the percentage of parts being correctly detected (PCP) was used [5]. Note that with the exception of the Buffy dataset, we use the strict PCP measure, not the loose PCP. In the strict version, if the maximum difference between the locations of two connected joints and the corresponding ground truth locations is less than 50% of the length of the corresponding body part, the location of that part is considered to be correctly estimated. In the loose version used for Buffy, not the maximum, but the average distance is considered. More details on this measure can be found in [13]. Notice that [24] used a different evaluation criterion for KTH.

Table 1, Table 2, Table 3 and Table 4 report the results obtained on all datasets by our branch and bound approach and the approaches proposed in [24], [8] and [25], where available. Note that these methods, including ours, use the model introduced by Yang and Ramanan as the underlying model. Additionally, we compare to state-of-the-arts results achieved by Eichner and Ferrari, and Pishchulin *et al.* [3,14]. In the KTH dataset, the tree structured model of [25] leads to a total score of 80.3% when 26 joints are used. The sequential branch and bound proposed in Section 3.5 outperforms this score by 3.1%, achieving 83.4%, similar to what is achieved using the re-ranking approach of [24]. Using the model consisting of 29 parts, our global branch and bound approach here scores 82.7%, outperforming the sequential version (81.9%). For this dataset, examples of successfully estimated poses are shown in Figure 4.

In the Parse dataset, the model of Yang and Ramanan achieves a score of 66.0% for the model with 29 joints. Our global branch and bound approach instead is able to achieve a score of 68.7%. Figure 5 shows some examples of correctly estimated poses compared with the ones obtained using [25] (29 parts). Figure 6, instead, shows failure cases on both methods. Although our approach is not able to detect the correct pose, its results are closer to the actual solution than the ones obtained using [25]. Figure 7 shows some more failure cases which might be caused by a wrong foreground map.

The loose PCP measure is commonly used for the Buffy dataset. We achieve 89.0% and perform 3.4% better than our baseline implementation of [25]. Note the significant increase in the detection of the lower arms, also shown qualitatively in Figure 8. The method of Ladicky *et al.* [8] performs very well on the Buffy dataset, but seems to have some shortcomings on the Parse dataset.

Fig. 4. KTH Dataset: The top row shows the results of the inference in the standard tree model [25]. The bottom row displays the results obtained using the proposed branch and bound algorithm.

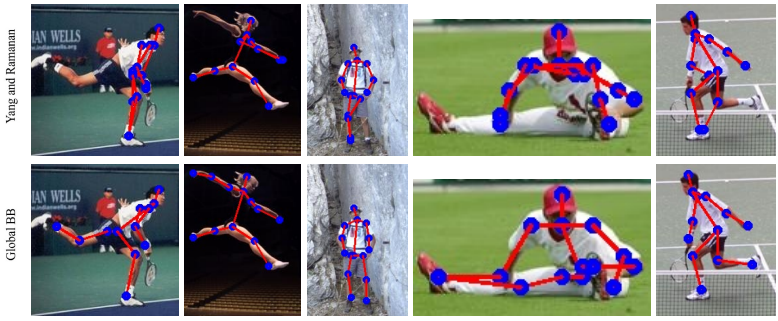


Fig. 5. Parse Dataset: Comparison between the results obtained using the approach of Yang and Ramanan [25] (top row), and the results obtained using our approach (bottom row).

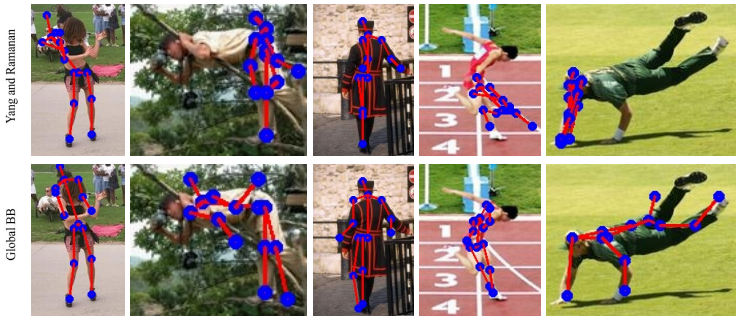


Fig. 6. Parse Dataset: Failure cases for both [25] and our global branch and bound approach

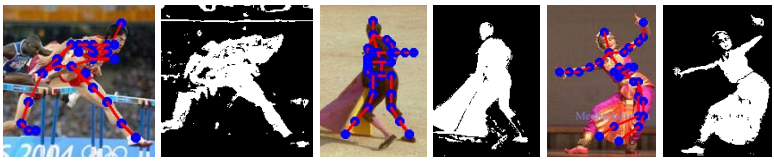


Fig. 7. Parse Dataset: Failure cases due to segmentation errors

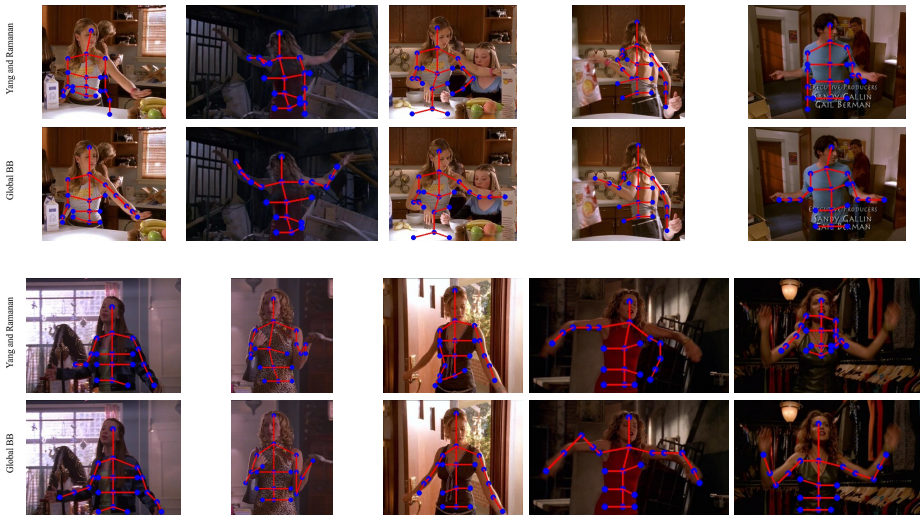


Fig. 8. Buffy Dataset: Comparison between the results obtained using the approach of Yang and Ramanan [25] (top row), and the results obtained using our approach (bottom row).

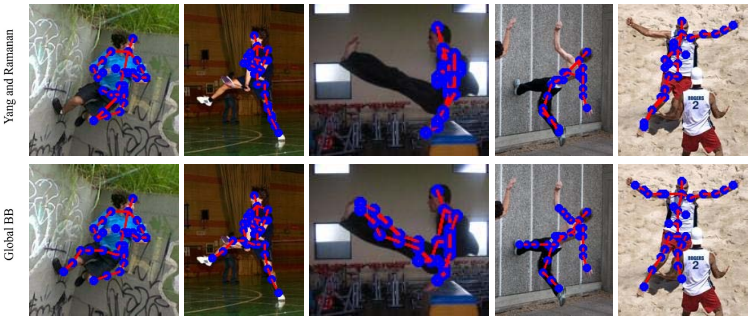


Fig. 9. Leeds Dataset: Comparison between the results obtained using the approach of Yang and Ramanan [25] (top row), and the results obtained using our approach (bottom row).

For all datasets observer centric labeling was used for training and testing. This means that the right arm and the right leg of a back facing person are labeled as left arm and left leg and vice versa.

The worst-case runtime complexity for our approach is exponential in the number of states. This happens when branch and bound degenerates to exhaustive search and the algorithm tries all the possible combinations of part positions. In practice, however, branch and bound terminates much earlier because only promising sets of hypotheses are divided further, ignoring and never exploring many large sets of hypotheses with low upper bounds. Experiments were run on an Intel Core i7, 2.8GHz, with 12GB of RAM. The runtime of the global branch and bound ranged between 2 and 10 minutes for most images in the full body datasets. Note that without the methods proposed in Section 3.5, the algorithm can take up to several hours or even days.

5 Conclusion

In this paper, we propose a method for single image human pose estimation which extends the common unary and pairwise terms of graphical models with a global foreground term. In order to guarantee the global optimality of our solutions, we propose to optimize the model using a branch and bound based algorithm. To keep inference tractable and avoid the obvious combinatorial explosion, we propose a set of upper bounds specifically designed for our pose estimation problem, and a way to decouple the estimate of rarely overlapping limbs while still maintaining the global optimality.

We evaluated the performance of the proposed method on four publicly available datasets, showing the benefits of adding a global foreground term. Branch and bound guarantees the best solution according to the specified model. Additionally, we show quantitative results of a sequential version of the proposed branch and bound algorithm.

In conclusion, the global foreground term improves the results when a reasonable segmentation or confidence map for the foreground $F(l)$ is available. Our automatic estimation of $F(l)$ works reasonably well in the tested datasets. However, when it fails, it influences the outcome of the pose estimation algorithm. Figure 7 shows some failure cases due to segmentation errors. Nevertheless, in many scenarios a good foreground model can be easily estimated and therefore we expect this algorithm to work well in such situations.

In future work, we plan to address pose estimation given multiple images of the same person either from multiple views or several neighboring frames of a video sequence. Encouraging consistency between several such input images suggests new challenges in terms of efficient inference and is an encouraging direction for more robustness.

Acknowledgments. This project is supported by a grant of CTI Switzerland, the 4DVideo ERC Starting Grant Nr. 210806 and the SNF Recording Studio Grant.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR (2005)
2. Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V.: 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. In: IJCV (2012)
3. Eichner, M., Ferrari, V.: Appearance sharing for collective human pose estimation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 138–151. Springer, Heidelberg (2013)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
5. Ferrari, V., Marin-Jiminez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Proc. CVPR (2008)
6. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proc. BMVC (2010)
7. Kohli, P., Rihan, J., Bray, M., Torr, P.H.S.: Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. IJCV (2008)
8. Ladicky, L., Torr, P.H.S., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: Proc. CVPR (2013)
9. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Efficient subwindow search: a branch and bound framework for object localization. TPAMI (2009)
10. Land, A.H., Doig, A.G.: An automatic method of solving discrete programming problems. Econometrica (1960)
11. Mori, G.: Guiding model search using segmentation. In: Proc. ICCV (2005)
12. Park, D., Ramanan, D.: N-best maximal decoders for part models. In: Proc. ICCV (2011)
13. Pishchulin, L., Jain, A., Andriluka, M., Thormaehlen, T., Schiele, B.: Articulated people detection and pose estimation: Reshaping the future. In: Proc. CVPR (2012)
14. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: Proc. ICCV (2013)
15. Rauschert, I., Collins, R.T.: A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 704–717. Springer, Heidelberg (2012)
16. Rother, C., Kolmogorov, V., Blake, A.: 'grabcut' - interactive foreground extraction using iterated graph cuts. In: Proc. of ACM SIGGRAPH (2004)
17. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR (2011)
18. Sigal, L., Black, M.J.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: Proc. CVPR (2006)
19. Singh, V.K., Nevatia, R., Huang, C.: Efficient inference with multiple heterogeneous part detectors for human pose estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 314–327. Springer, Heidelberg (2010)
20. Sun, M., Telaprolu, M., Lee, H., Savarese, S.: An efficient branch-and-bound algorithm for optimal human pose estimation. In: Proc. CVPR (2012)
21. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In: Proc. NIPS (2003)

22. Tian, T., Sclaroff, S.: Fast globally optimal 2d human detection with loopy graph models. In: Proc. CVPR (2010)
23. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proc. ICML (2004)
24. Vahid, K., Sullivan, J.: Using richer models for articulated pose estimation of footballers. In: Proc. BMVC (2012)
25. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. TPAMI (2013)
26. Zhang, X., Li, C., Tong, X., Hu, W., Maybank, S., Zhang, Y.: Efficient human pose estimation via parsing a tree structure based human model. In: Proc. ICCV (2009)