Multi-Level Inference by Relaxed Dual Decomposition for Human Pose Segmentation

Huayan Wang
Department of Computer Science
Stanford University
huayanw@cs.stanford.edu

Daphne Koller
Department of Computer Science
Stanford University
koller@cs.stanford.edu

Abstract

Combining information from the higher level and the lower level has long been recognized as an essential component in holistic image understanding. However, an efficient inference method for multi-level models remains an open problem. Moreover, modeling the complex relations within real world images often gives rise to energy terms that couple many variables in arbitrary ways. They make the inference problem even harder. In this paper, we construct an energy function over the pose of the human body and pixel-wise foreground / background segmentation. The energy function incorporates terms both on the higher level, which models the human poses, and the lower level, which models the pixels. It also contains an intractable term that couples all body parts. We show how to optimize this energy in a principled way by relaxed dual decomposition, which proceeds by maximizing a concave lower bound on the energy function. Empirically, we show that our approach improves the state-of-the-art performance of human pose estimation on the Ramanan benchmark dataset.

1. Introduction

Parsing articulated objects (*e.g.*, the human body) from an image or video sequence has been an area of great interests in computer vision. People try to recover the pose of the human body from various image cues as well as general knowledge on the body structure. The pose is often specified by the position and orientation of each body part (see Fig. 1, top row). Foreground / background segmentation (see Fig. 1, bottom row) is another important task in understanding the content of an image (or video sequence). However, due to the human body's large diversity in appearance and variability in the articulated structure, correctly segmenting it out from an uncontrolled background is extremely difficult without prior knowledge on the configuration of the body parts.

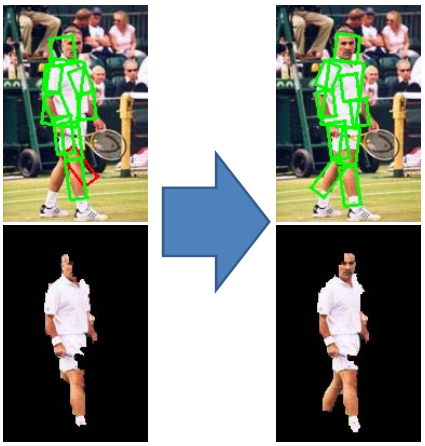

Figure 1. Joint inference on multiple levels (pose level and segmentations level) helps both tasks. **Left:** initial state (solution in the first iteration). **Right:** final state. Note that the estimation on head and left leg are corrected by segmentation cues.

Intuitively, solving each of these two problems should help the other. We would expect the estimation of the human pose to be a strong high-level guide for the foreground segmentation task. And the segmentation cues could also help improve the overall pose estimation (*e.g.*, some configuration of the arm should be more likely if the image cues suggest an arm-like segmentation). Fig. 1 shows an example from our experimental results. Jointly solving these two problems helps get more accurate pose estimation as well as better foreground segmentation.

To solve these two problems together, we construct a unified model over the human poses as well as pixel-wise foreground / background labeling. The model incorporates cues on the high level, such as the joints between body parts, and on the low level, such as pair-wise contrast between pixels. Our model also has energy terms that involve all body parts. For example, one energy term encourages the configuration of all body parts, altogether, to fully "explain" the

pixels that are likely to be foreground. It thus penalizes, for example, the case where the estimations of two legs overlap significantly, leaving another highly limb-like region unexplained by any other body parts (*e.g.*, the arms).

However, performing efficient inference on such multi-level models is an open problem. And the energy term that couples all body parts makes it even harder. In order to tackle the inference problem, we propose a relaxed variation of dual decomposition. Specifically, the inference problems on different levels are solved independently as slave problems but communicate via coordinating messages in the dual decomposition framework. The intractable energy terms are handled by further relaxation on the dual function. Primal solutions are constructed (in various ways) from the dual solutions and selected based on the original energy function. This is, we only need to evaluate the energy function on a handful of solutions constructed from dual solutions, instead of minimizing it directly. This essentially accommodates arbitrary complex energy terms as long as they can be *evaluated* (instead of *minimized*) efficiently.

The main contributions of this paper are in three aspects. Firstly, for pose estimation, the energy term that encourages the estimated pose to "fully" explain the foreground-like regions is a novel contribution that has significant impact on the performance, since it helps avoid cases where the same region is explained by two overlapping body parts, whereas another region of the body is not covered at all. Secondly, we believe that the use of an intractable energy function to rerank hypotheses derived from a simpler method, which has been used with great success in the natural language community [6], could be further exploited in computer vision tasks. Lastly, applying dual decomposition to "multi-level" inference, which is a commonly encountered scenario in computer vision, could have a wide spectrum of applications.

2. Related Work

Human pose estimation has been extensively studied in the computer vision literature [16, 8, 1, 9, 18, 20, 2, 17]. Many of these works build on the pictorial structure model [7]. Success of the method largely depend on whether the image cues are effectively used, for which people have been trying iterative learning [16], boosted part detectors [1], and multiple heterogenous part detectors [18]. High level knowledge such as human-object interaction [20] has also been incorporated to boost the performance. More efficient inference schemes have also been explored such as progressive pruning [8] and multi-scale inference [17]. Jiang [9] proposed to estimate the human pose by finding a consistent "max-covering", which also captures the intuition that the body parts should "fully" explain foreground regions. But their model does not incorporate segmentation cues and was not evaluated on public benchmark datasets.

Researchers have long been exploiting the idea of combining shape models and segmentation [13, 4, 10, 5, 2, 15]. However, how to perform joint inference on the two levels effectively and efficiently has been the major issue in all these works. For example, in *Objcut* [13], they first sample from the shape model and then solve the segmentation problem given the shape samples. So the two levels are not solved jointly. In *Posecut* [4, 10], inference was done by repeatedly solving the segmentation problem given different poses of the human body, and picking the pose with the lowest segmentation energy. So there is no direct influence from the segmentation module to the pose estimation module. Lubomir *et al* [2] addressed the pose detection and segmentation using "Poselets". However their approach emphasizes on learning discriminative poselets using 3D-annotated data, whereas segmentation is computed based on poselet detection results and the two problem are not solved jointly. Packer *et al* [15] worked on jointly solve the contour model and segmentation. Inference turns out to be the challenging part and they tackled it by searching with proposal moves from super-pixels.

Ladicky *et al* [14] incorporated object detector responses into CRFs for multi-class image segmentation. Each object candidate is treated as a higher level node, which connects to all relevant pixel nodes. We adopt a similar structure for the body-part candidates and relevant pixels. However, they were dealing with an easier problem where the objects (such as cars, pedestrians) are independent to each other, whereas in our problem the body parts are closely related.

Dual decomposition is a technology developed in the optimization community. Intuitively it decomposes the original problem into multiple slave problems, coordinated by exchanging messages with a shared master. When applied to MRF energy minimization, it is related to LP relaxation [12]. Generally speaking, its advantages are in two aspects. On one hand, the slave problems can be solved in parallel in a distributed manner [19]; on the other hand, the original hard problem can be decomposed into easier slave problems that are tractable by off-the-shelf tools. Our use of dual decomposition falls into the second category.

3. Model Formulation

We define the energy function over a graph (see Fig. 2) that has two types of variables (nodes). On the lower level we have one binary variable for each pixel, indicating whether it belongs to the human body (foreground) or background. We denote them by $\mathbf{x} = \{x_i\}_{i=1}^{N}$. These variables specifies the foreground segmentation. For specifying the human pose, we assume that for each body part we have a discrete set of candidate pose configurations (each comes with three parameters specifying position and orientation). The candidates come from a discretization and pruning step that can be treated as a black box to subsequent steps. In

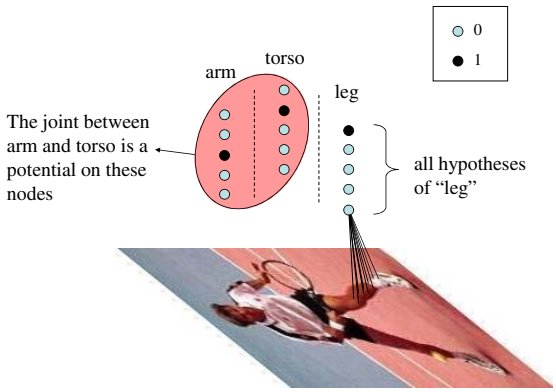

Figure 2. **The primal problem.** Each part-candidate node is linked to all pixels within the extent of that body part. We only show several such links on one part-candidate node for clarity. All nodes are binary. Each joint involves two groups of nodes corresponding to the two body parts. The energy of the joint is well defined only when we have one node turned on (filled) in each group.

practice we used the method of [1] to select the top hypotheses for each body part. One could also use any other method. Our model's higher level then has one binary variable for each candidate configuration of each body part. These variables indicate whether the candidate is chosen (on) or not (off) in the current body configuration. For example, if we consider ten candidate configurations for each of the ten body parts, we would have one hundred such variables. We denote them by $\mathbf{y} = \{y_{jk}\}_{M\times K}$, where j indexes the M body parts. In our setting, $M = 10$ as we have head, torso, and left/right upper/lower arms/legs. And k indexes the K candidates for each body part. In order to get a valid configuration of the human body, we need to impose the constraint that each body part should have one (and only one) candidate turned on. That is, $\sum_{k=1}^{K} y_{jk} = 1$ holds true for all j's, where K is the number of candidates for that body part. For convenience we will denote this feasible set by $\mathcal{C}$, so the constraints can be written as $\mathbf{y} \in \mathcal{C}$.

Recall that each part-candidate y_{jk} comes with a position and orientation. So we can compute a Gaussian mask in the image plane using the "standard" width and length[1] of that body part. The Gaussian mask assigns a weight to each pixel, which we denote by $\{w_{jk}^i\}_{i=1}^N$, where $w_{jk}^i = 1$ if pixel i lies right on the skeleton, and decreases exponentially as we move away from it.

We compute an image specific appearance model that is used in multiple energy terms in our model. Specifically a foreground model is computed by fitting a Gaussian mixture model (in the HSV color space) to all pixel i weighted by $\sum_{j,k} w_{jk}^i$. And a background model is computed similarly but using the weights $\sum_{j,k}(1 - w_{jk}^i)$. Then we compute, for each pixel, the posterior probability of being foreground, denoted by $\{p_i\}_{i=1}^N$. Note that w_{jk}^i and p_i are pre-computed and fixed during the entire inference procedure.

Given the above settings, we aim at minimizing an energy function consisting of these terms:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{y}} \mathbf{E}^P = \min_{\mathbf{x},\mathbf{y}} [&(\gamma_1 \mathbf{E}_{pixel-single}(\mathbf{x}) + \gamma_2 \mathbf{E}_{pixel-pair}(\mathbf{x}) + \\ &\gamma_3 \mathbf{E}_{part-pixel}(\mathbf{x},\mathbf{y}) + \gamma_4 \mathbf{E}_{part-single}(\mathbf{y}) + \\ &\gamma_5 \mathbf{E}_{joints}(\mathbf{y}) + \mathbf{E}_{residual}(\mathbf{x},\mathbf{y})] \\ &\text{s.t.} \quad \mathbf{y} \in \mathcal{C}.\end{aligned} \tag{1}$$

The first term includes the singletons on pixels:

$$\mathbf{E}_{pixel-single} = \sum_i \mathbf{1}(x_i = 1) \log \frac{1 - p_i}{p_i} \tag{2}$$

And we have the standard contrast sensitive pairwise potential [3] between all pairs of adjacent pixels:

$$\mathbf{E}_{pixel-pair} = \sum_{i,j} \mathbf{1}(x_i \neq x_j) \exp(-\theta_\beta \|I_i - I_j\|^2), \tag{3}$$

where θ_β is estimated by $\frac{1}{2\langle \|I_i - I_j\|^2 \rangle_{i,j}}$, where $\langle \cdot \rangle$ means averaging.

$\mathbf{E}_{part-pixel}$ includes pairwise potentials connecting each part-candidate node with relevant pixels:

$$\mathbf{E}_{part-pixel} = \sum_{i,j,k} \mathbf{1}(x_i = 0, y_{jk} = 1) w_{jk}^i \tag{4}$$

To avoid the unnecessary cost of connecting each part-candidate to too many pixels we threshold w_{jk}^i by a small positive number. Note that this energy term is submodular as it only penalize the $(1, 0)$ case, where a part-candidate is turned on but the pixel underneath is labeled as background.

$\mathbf{E}_{part-single}$ includes singleton potentials on the part-candidate nodes:

$$\mathbf{E}_{part-single} = \sum_{j,k} \mathbf{1}(y_{jk} = 1) \sum_i w_{jk}^i (1 - p_i) \tag{5}$$

$\mathbf{E}_{joints}$ models the joints between body parts. Note that each joint in our setting involves two groups of nodes representing all candidates for the two body parts. So this energy term is well defined only if the constraint $\mathbf{y} \in \mathcal{C}$ is satisfied. The joint is modeled by a three dimensional Gaussian over the relative position and angle between the two body parts. We learn the joint parameters and compute the joint energy using the same method as in [1].

[1] For the width and length of each body part, we use the same numbers as in [1].

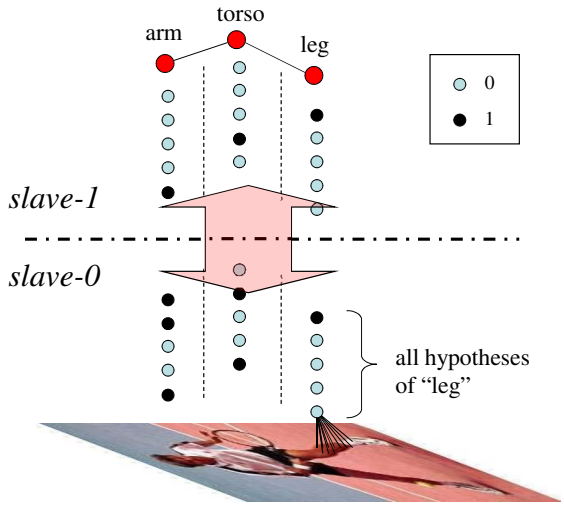

Figure 3. **Dual decomposition.** The larger body-part nodes (red) on the top can take on K possible states. Each state corresponds to one binary node underneath. In *slave-0*, the feasibility constraints are relaxed, such that each group could have an arbitrary number of nodes turned on (filled).

$\mathbf{E}_{residual}$ contains two parts $\mathbf{E}_{residual} = \mathbf{E}_{r1} + \gamma_6 \mathbf{E}_{r2}$ that both involve all part-candidate variables. Specifically,

$$\mathbf{E}_{r1} = \frac{\sum_i \mathbf{1}(\max_{j,k}(y_{jk} w^i_{jk}) < \delta_1) \log \frac{p_i}{1-p_i}}{\sum_i \mathbf{1}(\max_{j,k}(y_{jk} w^i_{jk}) < \delta_1)} \tag{6}$$

where the indicator function indicates that the pixel i is not covered by any of the body parts. So this energy term captures the intuition that areas that appear highly likely to belong to the human body would better be "explained" by some body part. And

$$\mathbf{E}_{r2} = \sum_i \mathbf{1}(\max_{j,k}(y_{jk} w^i_{jk}) < \delta_2) x_i \tag{7}$$

penalizes pixels that are far from the human body being labeled as foreground. In practice, we use a much smaller value for δ_2 than for δ_1. Note that these two terms intrinsically couple all body parts (as they account for mutual overlapping) and cannot be decomposed among (pairs of) them.

4. Dual Decomposition

Now we have the primal problem being the constrained energy minimizing problem (1). In this section we show how to optimize it with dual decomposition.

We first give some intuition behind the whole procedure. Our model naturally consists of two levels (see Figure 3). The higher level resembles the pictorial structure model, which has a simple tree structure among the body parts. The lower level only contains submodular potentials, which can be solved efficiently by graph cut [3, 11]. The part-candidate nodes are shared between these two levels, and naturally convey messages to coordinate them.

Formally, we make a copy of each of the part-candidate variables, and add the constraint that the two copies of the same variable should take the same value:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}^0,\mathbf{y}^1} [&\gamma_1 \mathbf{E}_{pixel-single}(\mathbf{x}) + \gamma_2 \mathbf{E}_{pixel-pair}(\mathbf{x}) + \\ &\gamma_3 \mathbf{E}_{part-pixel}(\mathbf{x}, \mathbf{y}^0) + \gamma_4 \mathbf{E}_{part-single}(\mathbf{y}^1) + \\ &\gamma_5 \mathbf{E}_{joints}(\mathbf{y}^1) + \mathbf{E}_{residual}(\mathbf{x}, \mathbf{y}^1)] \\ \text{s.t.} \quad &\mathbf{y}^0 = \mathbf{y}^1, \quad \mathbf{y}^1 \in \mathcal{C} \end{aligned} \tag{8}$$

The minimization problem (8) is equivalent to the primal problem (1) defined in the last section. For simplicity of subsequent equations we define

$$\begin{aligned} \mathbf{E}^0(\mathbf{x}, \mathbf{y}^0) = &\gamma_1 \mathbf{E}_{pixel-single}(\mathbf{x}) + \gamma_2 \mathbf{E}_{pixel-pair}(\mathbf{x}) \\ &+ \gamma_3 \mathbf{E}_{part-pixel}(\mathbf{x}, \mathbf{y}^0) \end{aligned} \tag{9}$$

$$\mathbf{E}^1(\mathbf{y}^1) = \gamma_4 \mathbf{E}_{part-single}(\mathbf{y}^1) + \gamma_5 \mathbf{E}_{joints}(\mathbf{y}^1) \tag{10}$$

Then (8) becomes

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}^0,\mathbf{y}^1} &(\mathbf{E}^0(\mathbf{x}, \mathbf{y}^0) + \mathbf{E}^1(\mathbf{y}^1) + \mathbf{E}_{residual}(\mathbf{x}, \mathbf{y}^1)) \\ \text{s.t.} \quad &\mathbf{y}^0 = \mathbf{y}^1, \quad \mathbf{y}^1 \in \mathcal{C} \end{aligned} \tag{11}$$

The Lagrange dual is formed by relaxing the coupling constraints on $\mathbf{y}^{0,1}$ by Lagrange multipliers,

$$\begin{aligned} g(\lambda) = \min_{\mathbf{x},\mathbf{y}^0,\mathbf{y}^1} &(\mathbf{E}^0(\mathbf{x}, \mathbf{y}^0) + \mathbf{E}^1(\mathbf{y}^1) + \mathbf{E}_{residual}(\mathbf{x}, \mathbf{y}^1) + \\ &\lambda \cdot (\mathbf{y}^0 - \mathbf{y}^1)) \\ \text{s.t.} \quad &\mathbf{y}^1 \in \mathcal{C} \end{aligned} \tag{12}$$

Recall that $\mathbf{y}^0$ and $\mathbf{y}^1$ are vectors, whose dimensionality equals the number of body parts times the number of candidate configurations for each body part. And λ is a vector of the same dimensionality.

For any value of λ, the dual function $g(\lambda)$ is a lower bound on the primal energy function. And the dual problem can be maximized by the sub-gradient method, which needs to first solve the minimization problem inside $g(\lambda)$.

We decompose the dual function to separate $\mathbf{y}^0$ and $\mathbf{y}^1$, which gives rise to a relaxed dual function:

$$\begin{aligned} g(\lambda) \geq \tilde{g}(\lambda) = &\min_{\mathbf{x},\mathbf{y}^0} (\mathbf{E}^0(\mathbf{x}, \mathbf{y}^0) + \lambda \cdot \mathbf{y}^0) + \\ &\min_{\mathbf{y}^1 \in \mathcal{C}} (\mathbf{E}^1(\mathbf{y}^1) - \lambda \cdot \mathbf{y}^1) + \\ &\min_{\mathbf{x},\mathbf{y} \in \mathcal{C}} \mathbf{E}_{residual}(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{13}$$

Note that the feasibility constraint is only imposed on $\mathbf{y}^1$ but not on $\mathbf{y}^0$, which makes the first minimization problem

in (13) tractable since it only contains submodular potentials thus can be solved by graph cut. We call this problem *slave-0*. Note that this does not make any difference in the primal problem as we also have the constraint that $\mathbf{y}^0 = \mathbf{y}^1$. Intuitively, in the dual problem, the coordinating messages eventually want $\mathbf{y}^0$ to be equal to $\mathbf{y}^1$, so in this regard the feasibility constraint does affect $\mathbf{y}^0$. The second minimization problem, which we call *slave-1*, is equivalent to the standard pictorial structure model, where we have one variable for each body part, taking on K possible states. This K-label problem has a simple tree structure and can be solved by max-product message passing. Then the solution to the K-label problem can be readily translated into the binary variables $\mathbf{y}^1$, which naturally satisfy the feasibility constraint $\mathbf{y}^1 \in \mathcal{C}$. The third term in (13) is a constant that does not depend on λ, which is eliminated in computing the sub-gradient, so we do not even bother solving this (intractable) minimization problem.

Given the current value of λ, let $\bar{\mathbf{x}}(\lambda)$, $\bar{\mathbf{y}}^0(\lambda)$ and $\bar{\mathbf{y}}^1(\lambda)$ be the optimal solutions to *slave-0* and *slave-1*. The sub-gradient of the relaxed dual function at λ is given by

$$\nabla \tilde{g}(\lambda) = \bar{\mathbf{y}}^0(\lambda) - \bar{\mathbf{y}}^1(\lambda) \tag{14}$$

To sum up, the sub-gradient method for maximizing the relaxed Lagrange dual $\tilde{g}(\lambda)$ proceeds by repeating the two steps:

1. Solve the two slave problems given the current value of λ.

2. Compute sub-gradient as in (14) and update λ by $\lambda \leftarrow \lambda + \alpha_t \nabla \tilde{g}(\lambda)$,

where α_t is the step size indexed by iteration t, which can be set adaptively as elaborated in [12].

To interpret the sub-gradients as coordinating messages, we consider a single part-candidate variable y_{jk}. It has two copies in $\mathbf{y}^0$ and $\mathbf{y}^1$ respectively. Suppose in some iteration the two copies agree, i.e., $y^0_{jk} = y^1_{jk}$, then we have $\{\nabla \tilde{g}(\lambda)\}_{jk} = 0$ according to (14), and the corresponding element in λ would not change in the next iteration. However, if $y^0_{jk} = 1$ but $y^1_{jk} = 0$, we have $\{\nabla \tilde{g}(\lambda)\}_{jk} = 1$, which would increase the penalty of turning on y^0_{jk} and decrease the penalty of turning on y^1_{jk}. So the sub-gradient method tries to pull the two copies of each part-candidate variable towards each other.

Up until now, we haven't really used the term $\mathbf{E}_{residual}$ as it is eliminated in the sub-gradient. Recall that our initial goal is to solve the primal problem. So we have to construct feasible primal solutions from the dual solutions (as discussed in the next section), and pick the one with minimum primal energy. In this procedure we only have to evaluate the primal energy function on a handful of primal solutions suggested by the dual solutions. Thus we can accommodate any energy term as long as it can be *evaluated* (instead of *minimized*) efficiently.

However, several desirable properties of the original dual function $g(\lambda)$ are lost in the relaxation. For example, $g(\lambda)$ is a *tight* lower bound, which equals the optimal primal energy we could possibly achieve, so the dual gap (the different between current primal objective and dual objective value) is a very informative quantity that can be used, for example, in setting the step size α_t. But this no longer holds true for the relaxed dual function $\tilde{g}(\lambda)$. Moreover, even though we can always pick the best primal solution by evaluating the original energy function, the quality of the primal solutions constructed can be affected by the fact that we are maximizing the relaxed dual function. These are the price we pay for accommodating intractable energy terms.

5. Constructing Primal Solutions

The dual solutions from the slave problems are either infeasible or incomplete for the primal. In this section, we show how to construct feasible (and better) primal solutions from the solutions of each of the two slave problems.

Slave-1: Let $\bar{\mathbf{y}}^1$ be the solution to *slave-1*. Here we have that $\mathbf{y}^1 \in \mathcal{C}$. So all we need to do is to find the value for $\mathbf{x}$, *i.e.*, the pixel labels, and put them together to form a valid primal solution. A naive approach is to take $\bar{\mathbf{x}}$, which is the solution from *slave-0*. However, note that when $\mathbf{y}$ is fixed, directly minimizing the primal energy function can be done efficiently by graph-cut (as $\mathbf{E}_{part-pixel}$ and $\mathbf{E}_{r2}$ merge into pixel singletons and all the other terms are constants). Since graph-cut finds the global optimal solution (given $\mathbf{y}$), this choice of $\mathbf{x}$ is guaranteed to be better than using the solution of *slave-0*.

Slave-0: Let $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}^0$ be the solution to *slave-0*. Now $\bar{\mathbf{y}}^0$ does not necessarily satisfy the feasibility constraint. A naive way of imposing the feasibility constraint would be to pick one hypothesis from each group with the highest probability of being turned on conditioned on $\bar{\mathbf{x}}$. However, this usually leads to very unlikely overall body configurations as the mutual consistency is ignored. Note that when $\mathbf{x}$ is fixed, $\mathbf{E}_{part-pixel}$ is a term of singletons on $\mathbf{y}$. We add them to $\mathbf{E}_{part-single}$ and run max-product message passing using this updated singletons and original $\mathbf{E}_{joints}$ to get a consistent body configuration. Then, based on the same procedure as in last paragraph, we fix $\mathbf{y}$ and re-do graph-cut to get $\mathbf{x}$.

Empirically these two methods of constructing primal solutions contribute about equally to the final solutions.

6. Experiments

In all experiments we use the method in [1] to select the top ten (*w.r.t* the marginal distribution) hypotheses for each

body part, and build our model on that. The top-ten candidates already give us a lot of room to improve the performance: for example, on the Ramanan dataset [16], picking the best one (*w.r.t* ground truth) among the top ten (*w.r.t* part marginals from [1]) for each body part gives us an over accuracy of 77.02%[2], compared to the actual performance of 55.2% of [1] and the state-of-the-art performance of 60.88% in [18].

We use the same evaluation metric for human pose estimation as most earlier works: a body part is considered as correct if the distances from both of its ends to the ground truth positions are less than half of the body-part length.

The Ramanan dataset [16] consists of 305 images (100 training and 205 test). Each image comes with the ground-truth labeling of the ten body parts. We run our methods on this dataset with the weights among the energy terms validated on the training set. Our method achieves an overall accuracy of 61.51% on the test set, comparing to the baseline performance of 55.2% in [1], which we used to select candidates. The best performance reported on this dataset was 60.88% in [18]. Detailed quantitative comparison are shown in Table 1.

In Fig. 4 we show our results on pose estimation and segmentation, and compare to the baseline result of [1]. Note that in many cases our improvement is not reflected quantitatively due to the limitation of the evaluation criterion. For example, see Fig. 5 (and also Fig. 6: 6th image right leg, 5th image right arm).

As we could not get the implementation of [18] to regenerate their results, we only compare to some results shown in their paper. The correctness of the body parts in their results is not color coded, so we also show the number of correct parts (as they reported) above each image.

It is noticeable that one major error mode in [1] and [18] corrected by our method is that the limbs are put in a overlapping position leaving other limb-like region "unexplained". As we can see from Table 1, we significantly outperform their methods in estimating leg and arm pose. This is due to the energy term $\mathbf{E}_{r1}$.

7. Conclusion and Future work

In this paper, we proposed a novel multi-level model that jointly solves human pose estimation and foreground segmentation. Inference is done by dual decomposition with a relaxed dual function that accommodates intractable energy terms. A promising further direction is to extend our framework to the temporal domain to handle video sequences. Specifically, given a video sequence, we could replicate the structure in Fig. 2 for each frame, so we have $\{\mathbf{x}^{(t)}\}_{t=1}^{T}$, $\{\mathbf{y}^{(t)}\}_{t=1}^{T}$, and all the energy terms in (1) for each frame. And we add lateral (temporal) potentials both on the pixel level and higher level. As we want "smoothing" over the temporal domain, these lateral potentials are all associative (submodular), so that we can have one *slave-0* over the entire sequence, that performs graph cut on the volume of pixels and part-candidates of all frames. We expect to explore this trajectory in future work.

[2]Note that the value 77.02% is a little bit overestimated as picking the best candidate for each body part independently does not necessarily give rise to a consistent overall configuration.

Acknowledgment

This work was supported by the Mind's eye grant W911NF-10-2-0059, National Science Foundation under Grant No. RI-0917151, and the Office of Naval Research under the MURI program (N000140710747).

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009. 2434, 2435, 2437, 2438, 2439, 2440

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, sep 2009. 2434

[3] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages I: 105–112, 2001. 2435, 2436

[4] M. Bray, P. Kohli, and P. H. S. Torr. Posecut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *ECCV*, pages II: 642–655, 2006. 2434

[5] Y. Chen, L. Zhu, C. Lin, A. L. Yuille, and H. Zhang. Rapid inference on a novel AND/OR graph for object detection, segmentation and parsing. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. MIT Press, 2007. 2434

[6] M. Collins. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70, 2005. 2434

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, Jan. 2005. 2434

[8] V. Ferrari, M. M. Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8, 2008. 2434

[9] H. Jiang. Human pose estimation using consistent max-covering. In *ICCV*, 2009. 2434

[10] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision*, 79(3):285–298, 2008. 2434

[11] Kolmogorov and Zabih. What energy functions can be minimized via graph cuts. *IEEETPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 2004. 2436

[12] N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2010. 2434, 2437

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head	Total
Andriluka [1]	81.4	67.3	59.0	63.9	46.3	47.3	47.8	31.2	32.1	75.6	55.2
Singh [18]	**91.2**	69.3	73.7	61.0	**68.8**	**50.2**	49.8	34.6	33.7	**76.6**	60.88
Our method	88.3	**71.2**	**77.6**	**73.7**	56.1	48.8	**51.7**	**36.1**	**36.6**	75.1	**61.51**

Table 1. Accuracy in percentage for each body part and overall. Note that our method significantly outperforms other methods in estimating the limbs.

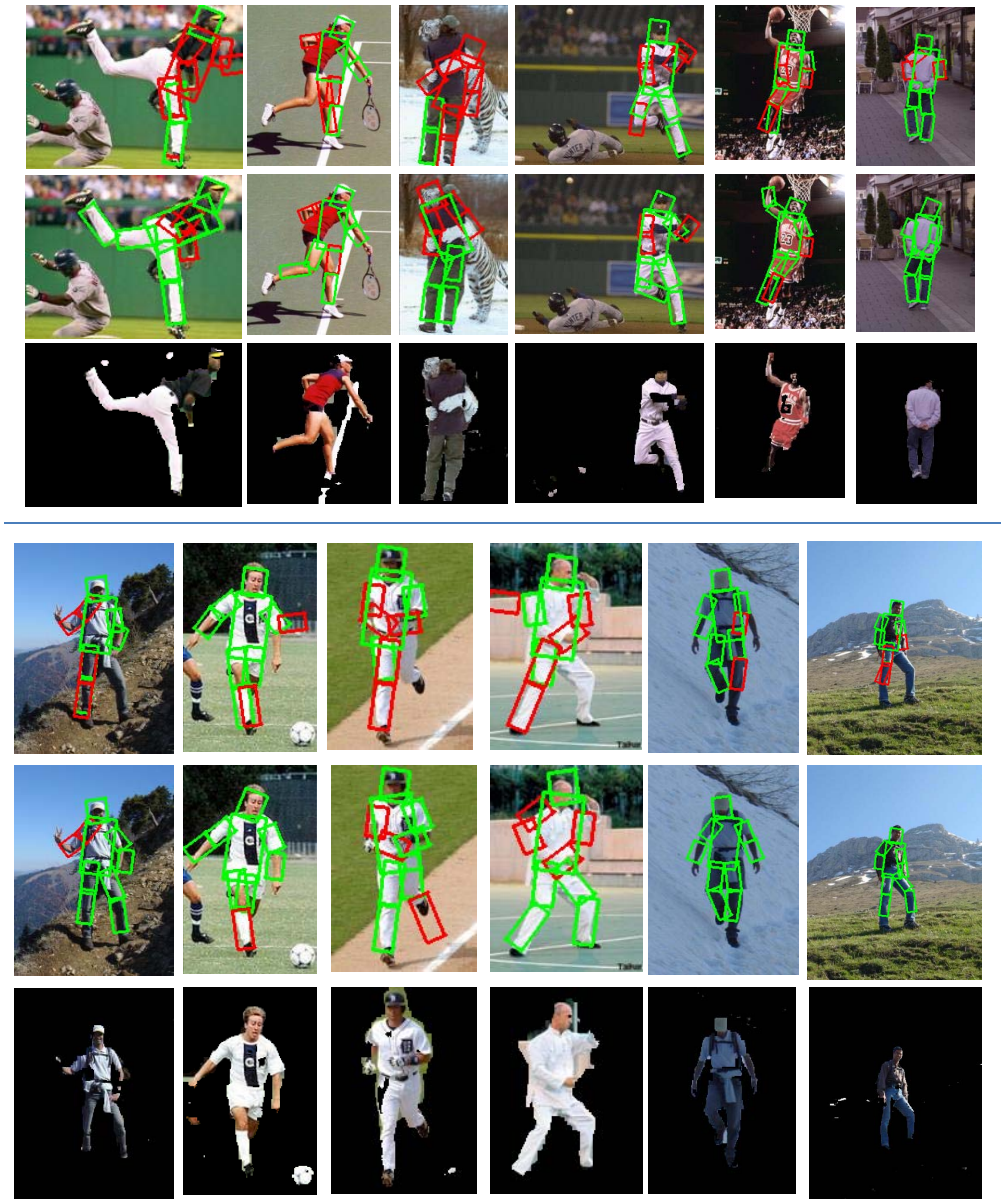

Figure 4. Qualitative comparison to [1]. For each image, **top:** result of [1] (we reproduced using their implementation), **middle:** our result of pose estimation, **bottom:** our result of foreground segmentation with background rendered black. Green box indicates the body part is correct according to the evaluation criterion; red indicates incorrect.

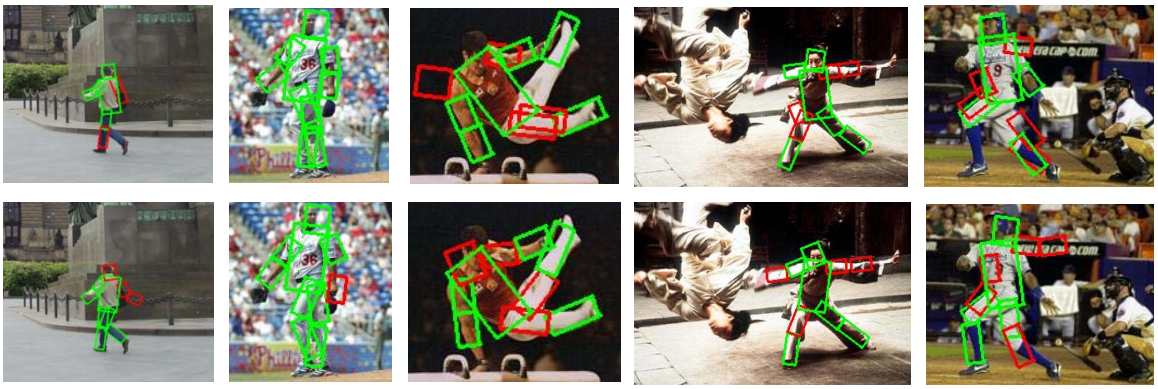

Figure 5. Limitation of the evaluation metric. **Top row:** the results of [1]. **Bottom row:** our results. Some improvements are not reflected by the current evaluation criterion. Green box indicates the body part is correct according to the evaluation criterion; red indicates incorrect.

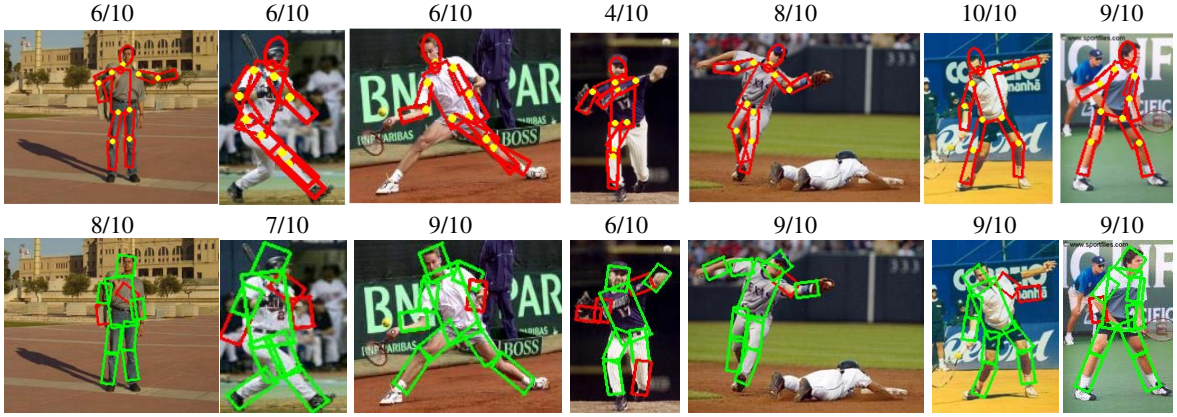

Figure 6. Qualitative comparison to [18]. **Top row:** the results of [18] (copied from their paper). **Bottom row:** our results. As their result boxes are not color coded, we show the number of correct boxes (as reported in [18]) above each image.

[13] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Objcut: Efficient segmentation using top-down and bottom-up cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):530–545, January 2009. 2434

[14] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and CRFs. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 424–437. Springer, 2010. 2434

[15] B. Packer, S. Gould, and D. Koller. A unified contour-pixel model for segmentation. In *ECCV*, 2010. 2434

[16] D. Ramanan. Learning to parse images of articulated bodies. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 1129–1136. MIT Press, 2006. 2434, 2438

[17] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (2)*, volume 6312 of *Lecture Notes in Computer Science*, pages 406–420. Springer, 2010. 2434

[18] V. K. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (3)*, volume 6313 of *Lecture Notes in Computer Science*, pages 314–327. Springer, 2010. 2434, 2438, 2439, 2440

[19] P. Strandmark and F. Kahl. Parallel and distributed graph cuts by dual decomposition. In *CVPR*, pages 2085–2092. IEEE, 2010. 2434

[20] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, June 2010. 2434