The Tenth Asian Conference on Computer Vision (ACCV) 2010

Finding Human Poses in Videos Using Concurrent Matching and Segmentation

Hao Jiang

Boston College, Chestnut Hill, MA 02467, USA

Abstract. We propose a novel method to detect human poses in videos by concurrently optimizing body part matching and object segmentation. With a single exemplar image, the proposed method detects the poses of a specific human subject in long video sequences. Matching and segmentation support each other and therefore the simultaneous optimization enables more reliable results. However, efficient concurrent optimization is a great challenge due to its huge search space. We propose an efficient linear method that solves the problem. In this method, the optimal body part matching conforms to local appearances and a human body plan, and the body part configuration is consistent with the object foreground estimated by simultaneous superpixel labeling. Our experiments on a variety of videos show that the proposed method is efficient and more reliable than previous locally constrained approaches.

1 Introduction

Detecting human poses in videos has important potential applications in video editing, movement analysis, action recognition, and human computer interaction. It is challenging due to body part articulation, self-occlusion and background clutter.

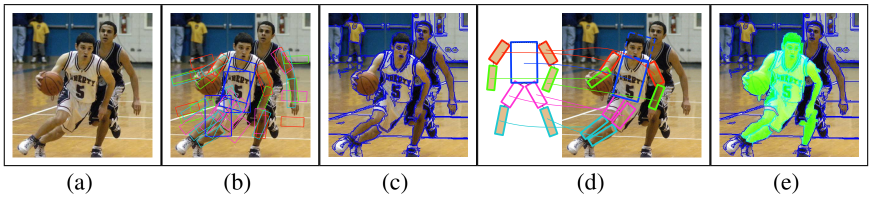


Fig. 1. To detect the human pose in (a), we first detect body part candidates (small set of samples are illustrated in (b)), partition the image into superpixels in (c), and then we concurrently optimize body part matching in (d) and foreground segmentation by superpixel labeling in (e). The cardboard model in (d) is extracted from a single exemplar image; it contains only the information about object foreground colors and body part rectangle shapes.

In this paper, we detect 2D poses of a specific human subject in monocular videos using a cardboard model built from a single exemplar image. Fig.1 illustrates the problem we tackle: we *concurrently optimize body part matching and object segmentation* for robust pose detection. Concurrently optimizing object matching and segmentation enables more robust results since the two closely related tasks support each other. However, the concurrent optimization is a great challenge due to huge number of feasible configurations. To make things worse, for our application it is difficult to obtain good initializations for both tasks. We therefore have to solve a hard combinatorial problem. In this paper, we propose a highly efficient linear relaxation method to optimize matching and segmentation concurrently for robust human pose detection in videos.

Previous research on simultaneous matching and segmentation [25] focuses on active contours and partial differential equation approaches. These methods require initial contours to be close to real targets; deformable models are used and it is hard to extend them to articulated object pose detection, which our method tackles.

PoseCut [18] is the first method for simultaneous human pose estimation and figure-ground separation. It has been successfully applied to 3D human pose tracking in a multiple-camera setting. PoseCut uses a parametric 3D human body model. It explicitly enumerates a small set of poses around an initial guess and uses an efficient dynamic graph-cuts method to compute the optimal foreground estimation for each hypothesis. A gradient descent method is further used to find the optimal pose. PoseCut requires good pose initialization in each video frame. Our proposed method removes this constraint.

Shape prior scheme [19, 26] is a popular way to combine pose estimation and segmentation in restricted pose domains such as walking and running. With the object shape prior, pose estimation is able to achieve reliable results; the estimated poses can be further combined with a segmentation algorithm to obtain accurate object foreground estimation. Unfortunately, for unconstrained human poses, the shape prior is weak. The widely used iterative approach [19, 26] that alternates between shape matching and segmentation does not work for our problem, since neither matching nor segmentation provides a good initialization for the other. We need methods that optimize object matching and segmentation concurrently instead of iteratively.

Due to the overwhelming computational complexity, concurrent optimization of matching and segmentation for unconstrained human pose detection has not yet been achieved. The contribution of this paper is that we propose a novel linear method that efficiently solves the problem. In this method, body part matching finds the optimal pose that follows local appearances, resembles a human body plan, and the covering region is consistent with the object foreground estimated simultaneously by superpixel labeling. The linear optimization can be relaxed and efficiently solved using a branch and bound method. This linear approach is general and can be easily extended to generic object matching and segmentation.

1.1 Related Work

Different methods have been proposed for detecting human poses in videos and images. With multiple view videos [1, 23], 3D poses can be detected. Extracting 3D poses in single view videos [2, 24] is currently limited to movements in specific domains. In this paper, we focus on finding 2D human poses in single view videos. Previous methods for 2D human pose estimation use holistic models or body part graph models.

The holistic approach treats a human body as a whole entity. Poses are estimated either by using pose classification [13] if the object can be segmented from the background, or by matching exemplars in databases [3–5]. Dynamic models have also been combined with exemplar methods [6] to improve the performance in pose tracking. To estimate unconstrained poses, exemplar methods become increasingly complex since we have to deal with huge pose databases.

Human poses can also be estimated by matching a body part graph model to target images. Body part matching scheme is flexible; it is able to model complex poses with a compact representation. The challenge is how to search for the optimal pose in a huge number of feasible configurations. If the body part relation graph is a tree, polynomial algorithms exist. Felzenszwalb et al. propose an efficient dynamic programming method [9] for pose estimation. Ramanan et al. [7, 20] estimate human poses in cluttered images using efficient message passing on trees. Recently, Andriluka et al. [28] extend the tree methods and devise a strong pose detector. Tree structure methods sometimes over-count local evidences because no constraints are directly applied among tree branches.

To solve the over-counting problem, non-tree graph models have also been studied. Pairwise constraints between body parts are introduced to form loopy relation graphs. Searching for poses using non-tree models is NP-hard. A branch and bound method [29] is proposed to solve problems with extra constraints on legs. Approximation methods based on belief propagation [14, 16], mathematical programming [15, 11] and probability sampling [10, 12] have also been proposed. In these methods, overlapping body parts are uniformly penalized, which relieves the over-counting issue, but at the same time, also introduces an undesired penalty to true overlapping body part configurations.

Image segmentation has been used to support body part graph matching in different fashions. Mori [17] uses superpixels to guide pose detection. Ramanan [20] proposes an effective learning method to generate strong body part detectors using soft image segmentation. Johnson et al. [27] use image segmentation to enhance local part detection. Ferrari et al. [21] pre-segment images to obtain rough foregrounds in human upper body pose estimation. In these methods, segmentation is not jointly optimized with the tree structure body part matching. In [8], a pre-segmented object foreground is used to constrain the layout of body parts in the optimization. In the sequential process, the segmentation result greatly affects the performance of pose estimation.

Even though intensively studied, finding human poses in cluttered images and videos is still largely unsolved. In this paper, we focus on detecting a specific subject's poses in single view videos. We propose an efficient linear method that concurrently optimizes body part matching and foreground segmentation. It works for dynamic background videos and unconstrained human poses, and to our knowledge no previous methods are able to achieve the concurrent optimization efficiently in such settings.

2 Concurrent Matching and Segmentation for Pose Estimation

Our task is to detect the poses of a specific human subject in videos. To distinguish the target subject, we extract a cardboard model from a single exemplar image. This procedure is necessary since the target subject may be among a group of people in videos; in this case, a generic pose detector will not work. The cardboard model includes 9 body parts, i.e., a torso and 8 half limbs; each part's appearance is represented by the average RGB color; the foreground color histogram is also stored. We jointly optimize body part matching and foreground estimation for robust pose detection. Formally, we try to find body part matching $\mathcal{X}$ and foreground estimation $\mathcal{Y}$ in a constrained optimization:

$$\min_{\mathcal{X},\mathcal{Y}} \{ B(\mathcal{X}) + S(\mathcal{Y}) + \sum_{(u,v)\in I} |h_{u,v} - g_{u,v}| \} \qquad (1)$$

s.t. $h_{u,v} = 1$ if body parts cover point (u, v), otherwise $h_{u,v} = 0$.
$g_{u,v} = 1$ if point (u, v) is in a foreground superpixel, otherwise $g_{u,v} = 0$.

Each feasible $\mathcal{X}$ determines a body part configuration in the target image and its cost is $B(\mathcal{X})$. $B(\mathcal{X})$ is small if we detect the true pose. Foreground estimation $\mathcal{Y}$ is obtained by superpixel labeling, and its cost is $S(\mathcal{Y})$, which is small if we label the real object foreground. In Eqn.(1), (u, v) represents points in the target image, $h_{u,v}$ is the body part covering map determined by $\mathcal{X}$, and $g_{u,v}$ is the foreground map determined by $\mathcal{Y}$. The term $\sum_{(u,v)\in I} |h_{u,v} - g_{u,v}|$, where I is the target image point set, penalizes the discrepancy between the body part covering and the foreground estimation. By minimizing the objective function in Eqn.(1), we find the optimal body part matching and foreground segmentation that are also consistent with each other.

2.1 Body Part Candidates and Superpixels

Before proceeding to concurrent optimization, we find body part candidates and partition the target image into superpixels. We search for body part candidates using Chamfer matching and color matching. The model shape and colors are extracted from a single exemplar image. Chamfer matching correlates body part bars to the distance transform of the target image edge map. The color differences of each body part with the target image at different locations and orientations are also computed. Chamfer matching costs and color matching costs are linearly combined to form the local body part matching costs. Using non-minimum suppression, we locate body part candidates.

Each half limb candidate is represented by two end points, a rotation angle and a rectangle of specific size. We further group the half limb candidates into full limb candidates and reject apparent wrong pairs: if the distance between the end points of two candidates is greater than a threshold, they cannot be connected together. For the combined limb candidate, its cost is the linear combination of the upper and lower body part matching costs, the distance between the connection joints and the difference between the two sub-limb angles.

We use a graph-cuts method [22] to over-segment images into superpixels. A superpixel contains image pixels that have similar appearance. The superpixels do not consistently partition target objects into body parts. However, points on each limb tend to be in the same superpixel. The overall object coverage, consisting of a bunch of smaller patches, forms a stable foreground region.

The concurrent optimization in Eqn.(1) is a hard combinatorial problem. Due to huge number of feasible configurations for body part assignment and superpixel labeling, exhaustive search is not feasible. Our strategy is to construct a linear formulation and devise an efficient solution.

2.2 The Linear Optimization

We express Eqn.(1) as a linear optimization in the following three steps:

First, We express the body part matching cost $B(\mathcal{X})$ in Eqn.(1) using linear functions and determine how body part covering map $h_{u,v}$ is related to body part assignments. We introduce indicator variables $x_{n,i}$. If body part n selects candidate i, $x_{n,i} = 1$, and otherwise $x_{n,i} = 0$. We also use (n, i) to denote the candidate i of body part n. After merging the upper and lower limbs, we have 5 parts. Note that the pose estimation still gives 9-part matching result.

Let the cost of assigning candidate i to body part n be $c_{n,i}$, which can be computed as discussed in §2.1. The overall body part assignment cost is $\sum_{n\in\mathcal{P}} \sum_{i\in\mathcal{A}(n)} (c_{n,i} \cdot x_{n,i})$, where $\mathcal{P}$ is the set of body parts and $\mathcal{A}(n)$ is the candidate set of part n. Since each body part selects one and only one candidate, we have $\sum_{i\in\mathcal{A}(n)} x_{n,i} = 1, \forall n \in \mathcal{P}$.

Apart from matching local appearances, body parts also need to follow a body plan: the end points of limbs should be close to the appropriate torso end point. Fig.2(a) illustrates the relation among body parts. The degree that a body part configuration follows a valid body plan can be quantified as:

$$\sum_{n \in \mathcal{P}, n \neq t} || \sum_{i \in \mathcal{A}(n)} \mathbf{p}_{n,i} x_{n,i} - \sum_{k \in \mathcal{A}(t)} \mathbf{t}_{n,k} x_{t,k} || \,, \tag{2}$$

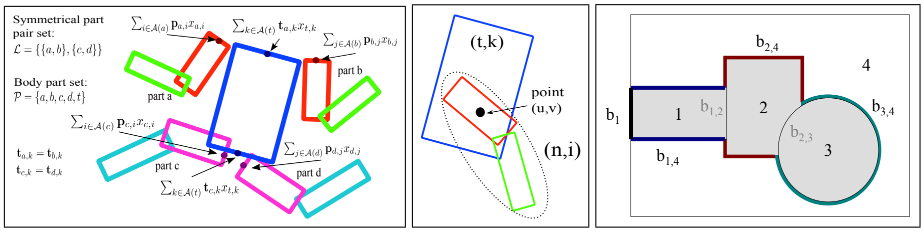
where $\mathbf{p}_{n,i}$ is the upper end point of candidate (n, i); $\mathbf{t}_{n,k}$ is the end point of torso candidate k and the end point is adjacent to part n; t is the torso. The notations are illustrated in Fig.2(a). $||.||$ is the L_1 norm. The L_1 norm terms can be linearized using auxiliary variables: $\min|\xi|$ is equivalent to $\min(\eta)$, s.t. $-\eta \leq \xi \leq \eta, \eta \geq 0$. The complete linear form is in Eqn.(3).

Limbs also tend to be symmetrical in spatial locations relative to the torso. If we draw a line segment between the upper arm or the upper leg joints, the center should be close to one suitable end of the torso. The following term is included to quantify the degree of symmetry:

$$\sum_{\{n,m\} \in \mathcal{L}} || \sum_{i \in \mathcal{A}(n)} \mathbf{p}_{n,i} x_{n,i} + \sum_{j \in \mathcal{A}(m)} \mathbf{p}_{m,j} x_{m,j} - 2 \sum_{k \in \mathcal{A}(t)} \mathbf{t}_{n,k} x_{t,k} || \,,$$

where $\mathcal{L}$ is the set of symmetrical body part pairs. The notations are illustrated in Fig.2(a). We also use the L_1 norm so that this term can be linearized using the auxiliary variable trick. The body part matching cost can then be represented as the linear combination of the local matching cost, the degree that it follows a body plan and the symmetry cost.

For human pose detection, simply optimizing the above body part matching energy is insufficient because it has a strong bias towards single limb detection. To solve the problem, we assemble body parts so that their overall covering is similar to the object foreground, which, as discussed later, is obtained simultaneously by superpixel labeling. To this end, we introduce auxiliary variables $h_{u,v}$ to represent the body part covering map. Here (u, v) is a point in the target image point set I. If point (u, v) is covered by the estimated body configuration, we wish $h_{u,v}$ to be 1, and otherwise, 0.


Symmetrical part pair set:
$\mathcal{L} = \{\{a,b\},\{c,d\}\}$
Body part set:
$\mathcal{P} = \{a,b,c,d,t\}$
$\sum_{i \in \mathcal{A}(a)} \mathbf{p}_{a,i} x_{a,i}$
$\sum_{k \in \mathcal{A}(t)} \mathbf{t}_{a,k} x_{t,k}$
$\sum_{j \in \mathcal{A}(b)} \mathbf{p}_{b,j} x_{b,j}$
part a
part b
$\sum_{i \in \mathcal{A}(c)} \mathbf{p}_{c,i} x_{c,i}$
$\sum_{j \in \mathcal{A}(d)} \mathbf{p}_{d,j} x_{d,j}$
$\mathbf{t}_{a,k} = \mathbf{t}_{b,k}$
$\mathbf{t}_{c,k} = \mathbf{t}_{d,k}$
part c
part d
$\sum_{k \in \mathcal{A}(t)} \mathbf{t}_{c,k} x_{t,k}$
(t,k)
point (u,v)
(n,i)
$b_{2,4}$
4
b_1
1
$b_{1,2}$
2
$b_{3,4}$
$b_{2,3}$
3
$b_{1,4}$

(a) (b) (c)

Fig. 2. (a) Notations for body part matching. (b) Part covering. (c) A toy example of superpixel labeling; the gray region is the foreground.

$h_{u,v}$ is constrained by the body part assignment variables $x_{n,i}$:

$$\sum_{\forall (n,i) \text{ covers } (u,v)} x_{n,i} \geq h_{u,v},\ 0 \leq h_{u,v} \leq 1,\ \forall (u,v) \in I.$$

If (u, v) is not covered by any part candidates, $h_{u,v}$ is set to 0. With such constraints, if no body part covers (u, v), $h_{u,v}$ has to be 0; if at least one body part covers (u, v), $h_{u,v}$ can be as big as 1, but it still can be 0. We therefore need to further make sure that $h_{u,v}$ must be 1 if at least one body part covers the pixel:

$$h_{u,v} \geq x_{n,i},\ \forall (n,i) \text{ covers } (u,v)\,.$$

As an example, in Fig.2(b), there are two part candidates covering point (u, v). The relation between $h_{u,v}$ and x is: $x_{n,i} + x_{t,k} \geq h_{u,v}$, $h_{u,v} \geq x_{n,i}$, $h_{u,v} \geq x_{t,k}$ and $0 \leq h_{u,v} \leq 1$. It is easy to verify that $h_{u,v}$ is indeed the body part covering map.

Next, we represent term $S(\mathcal{Y})$ in Eqn.(1) in linear form and relate it to the foreground map $g_{u,v}$. We introduce binary variable y_i to indicate whether superpixel i is on the foreground or background. If superpixel i is on the foreground, $y_i = 1$, and otherwise $y_i = 0$. To quantify the cost of labeling a superpixel as foreground, we compute the smallest distance from each color in the superpixel to the foreground colors in the template and sum all the color distances to form the superpixel labeling cost. Denoting the cost as the same c as the body part labeling cost but with a single index, the overall cost of the foreground estimation is $\sum_{i \in \mathcal{V}} (c_i \cdot y_i)$, where $\mathcal{V}$ is the set of superpixels in the target image.

Simply minimizing the superpixel assignment cost would result in a small foreground estimation. We need to constrain the size of the foreground segmentation to remove the bias. Assuming that the area of superpixel i is r_i, we constrain the object foreground to have an approximate area s_f, which is the exemplar foreground area. We therefore need to minimize $|\sum_{i \in \mathcal{V}} (r_i \cdot y_i) - s_f|$. The absolute value of the area difference can be linearized using auxiliary variables.

Besides, we hope that an object foreground contains a group of connected superpixels. Since connected regions tend to have small perimeter, we minimize the overall boundary length of the foreground superpixels to implicitly enforce this constraint. Let $b_{i,j}$ be the length of the common boundary between the neighboring superpixels i and j, and b_i be the length of the common boundary between superpixel i and the image bounding box. The perimeter of the foreground region is: $\sum_{\{i,j\} \in \mathcal{N}_s} (b_{i,j} \cdot |y_i - y_j|) + \sum_{i \in \mathcal{D}} (b_i \cdot y_i)$, where $\mathcal{N}_s$ is the set of neighboring superpixel pairs; $\mathcal{D}$ is the set of superpixels adjacent to the image bounding box. Fig.2(c) illustrates a toy example in which $\mathcal{N}_s = \{\{1,2\},\{2,3\},\{1,4\},\{2,4\},\{3,4\}\}$ and $\mathcal{D} = \{1\}$ and the above equation computes the foreground perimeter. The above connectivity term can also be linearized using auxiliary variable tricks. The superpixel labeling energy $S(\mathcal{Y})$ is therefore the linear combination of the three terms: the superpixel color matching term, the size term and the connectivity term.

To facilitate the comparison of foreground estimation with body part covering, we introduce auxiliary variables $g_{u,v}$ to represent the foreground map at the image pixel level. If (u, v) in the target image is covered by a superpixel, $g_{u,v}$ has the same value as the superpixel label: $g_{u,v} = y_i, \forall (u,v) \in R_i\,, i \in \mathcal{V}$, where R_i is the point set of superpixel i.

Finally, we are ready to express the complete optimization. we have formulated $B(\mathcal{X})$, $S(\mathcal{Y})$, $h_{u,v}$ and $g_{u,v}$ in Eqn.(1) using linear functions and linear constraints. With the above settings, $\sum_{(u,v)\in I} |h_{u,v} - g_{u,v}|$, where I is the set of points in the target image, equals the difference between the body part covering region and the foreground region estimated in the superpixel labeling. When minimizing the total energy $B(\mathcal{X}) + S(\mathcal{Y}) + \sum_{(u,v)\in I} |h_{u,v} - g_{u,v}|$, we find the optimal body part matching and foreground estimation that are consistent with each other. This consistency criterion is soft, and therefore it allows partial mismatches between the body part rectangles and the foreground superpixels. The concurrent optimization is also a principled way to solve the over-counting issue without introducing an undesired penalty for truly overlapping body parts, since body parts are now encouraged to fit the foreground instead of being simply pushed away from each other.

2.3 Relaxation and Branch and Bound Solution

Pose estimation can therefore be formulated as the following linear optimization:

$$\min\{\sum_{(u,v)\in I} z_{u,v} + \alpha_1 \sum_{n\in\mathcal{P}} \sum_{i\in\mathcal{A}(n)} (c_{n,i}\cdot x_{n,i}) + \alpha_2 \sum_{n\in\mathcal{P}, n\neq t} \sum_{l=1}^{2} p_n^{(l)} + \tag{3}$$

$$\alpha_3 \sum_{\{n,m\}\in\mathcal{L}} \sum_{l=1}^{2} q_{n,m}^{(l)} + \beta_1 \sum_{i\in\mathcal{V}} (c_i\cdot y_i) + \beta_2 [\sum_{\{i,j\}\in\mathcal{N}_s} (b_{i,j}\cdot y_{i,j}) + \sum_{i\in\mathcal{D}} (b_i\cdot y_i)] + \beta_3 w\}$$

$$\text{s.t.} \sum_{i\in\mathcal{A}(n)} x_{n,i} = 1, \forall n\in\mathcal{P}. \qquad \sum_{\forall(n,i)\ \text{covers}\ (u,v)} x_{n,i} \geq h_{u,v}, 0 \leq h_{u,v} \leq 1\,, \forall (u,v)\in I.$$

$$h_{u,v} \geq x_{n,i}, \forall (n,i)\ \text{covers}\ (u,v)\ , \forall (u,v)\in I.$$

$$g_{u,v} = y_i, \forall (u,v)\in R_i, \forall i\in\mathcal{V}. \quad -z_{u,v} \leq g_{u,v} - h_{u,v} \leq z_{u,v}, \forall (u,v)\in I.$$

$$-p_n^{(l)} \leq \sum_i p_{n,i}^{(l)} x_{n,i} - \sum_k t_{n,k}^{(l)} x_{t,k} \leq p_n^{(l)},\ l = 1..2, n\in\mathcal{P},\ n\neq t.$$

$$-q_{n,m}^{(l)} \leq \sum_i p_{n,i}^{(l)} x_{n,i} + \sum_j p_{m,j}^{(l)} x_{m,j} - 2\sum_k t_{n,k}^{(l)} x_{t,k} \leq q_{n,m}^{(l)},$$

$l = 1..2, \{n,m\}\in\mathcal{L}$ which includes two arms and two legs, t is the torso.

$$-y_{i,j} \leq y_i - y_j \leq y_{i,j}, \forall \{i,j\}\in\mathcal{N}_s. \quad -w \leq \sum_{i\in\mathcal{V}} (r_i\cdot y_i) - s_f \leq w.$$

All variables $\geq 0, x, y$ are binaries.

The variables $x_{n,i}, y_i, h_{u,v}$ and $g_{u,v}$ follow the previous definitions. The auxiliary variables $z_{u,v}$, $p_n^{(l)}$, $q_{n,m}^{(l)}$, $y_{i,j}$ and w are included to help turn the L_1 norm terms into linear functions. Coefficients $p_{n,i}^{(l)}, l = 1..2$, are the elements of $\mathbf{p}_{n,i}$, and $t_{n,k}^{(l)}, l = 1..2$, are the elements of $\mathbf{t}_{n,k}$; $\mathbf{p}$ and $\mathbf{t}$ are defined in Eqn.(2). In the objective function, the terms with α coefficients correspond to the body part assignment cost $B(\mathcal{X})$ in Eqn.(1); β coefficient terms correspond to the superpixel labeling cost $S(\mathcal{Y})$; and the z term is the covering consistency $\sum_{(u,v)\in I} |h_{u,v} - g_{u,v}|$ in Eqn.(1). The α and β coefficients are selected manually by trial and error; they are fixed in all the experiments. Typical values are $\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 0.1$ and $\beta_1 = \beta_2 = \beta_3 = 0.01$.

In this formulation, the variables x and y for body parts and superpixels are binary. The map variables g and h are continuous. Directly solving the mixed integer program is not feasible. We relax it for an approximate solution. A relaxation of both x and y into continuous variables yields weak results, in which the superpixel indicator variables y often obtain equal value and body part assignment does not benefit from the decisions on y. We therefore only relax x to continuous variables in [0,1] and keep y as binary variables. The relaxed problem can be efficiently solved by a branch and bound method.

An initial random superpixel labeling is used to estimate an upper bound of the optimization. The branch and bound method picks up a superpixel and generates two branches: one labels the superpixel 1, and the other labels it 0. The lower bound of the optimization for each branch is computed using the linear program by relaxing all the other variables in Eqn.(3). If the lower bound is greater than the current upper bound, the branch is cut; otherwise it is expanded by including two branches for another superpixel. The upper bound is updated whenever an integer solution for each y is obtained in a branch. This procedure repeats until every superpixel obtains binary solution in each surviving branch.

Our method quickly converges. In the relaxation solution, very few x variables are nonzero. Keeping only the body part candidates that correspond to these non-zero assignment variables, we solve the full integer program. Since there are few variables, the exhaustive search converges quickly. We can further lower the complexity by reducing the number of g and h variables. We define them on coarser image blocks instead of image pixels. We use 2500 g and h variables respectively in the optimization. With about 100 torso candidates, 10 thousand candidates for each full limb and a few hundred superpixels, the average running time for the concurrent optimization is about 25 seconds on a 2.8GHz machine.

3 Experimental Results

We evaluate the proposed method on a variety of video sequences. The test data include recorded videos and the videos from the web of total 4413 frames and 755-frame videos from the HumanEva dataset [30]. The four recorded sequences contain complex poses and strong background clutter. We select the sequences of three different subjects in different actions from the HumanEva dataset. These sequences are from camera one, whose view has the strongest background clutter. For each test sequence, we use the proposed concurrent optimization method to match a cardboard model, estimated from a single exemplar image in the sequence, to the target images to estimate human poses. For fair comparison, we use the same “walking” pose exemplar in each sequence for all the testing methods.

To verify the usefulness of the concurrent optimization approach, we compare it with some of the variations. We first test whether using superpixel labeling alone would yield satisfactory foreground estimation. If this were the case, we could use a sequential optimization instead of the more complex concurrent optimization. Fig.3 shows that the superpixel labeling alone cannot yield reliable foreground segmentation. Without a global shape constraint, it gives lots of false positives and false negatives. The concurrent optimization is necessary, and it helps to obtain a roughly correct foreground

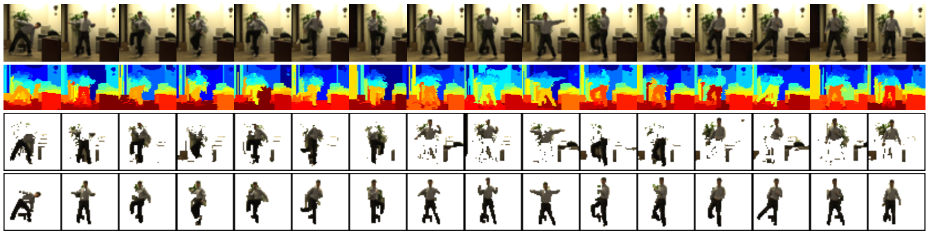


Fig. 3. Foreground estimation comparison. Row 1: sample images from sequence lab-man-I. Row 2: superpixel partitions of images. Row 3: foreground estimation using superpixel labeling alone. Row 4: foreground estimation using the concurrent optimization.

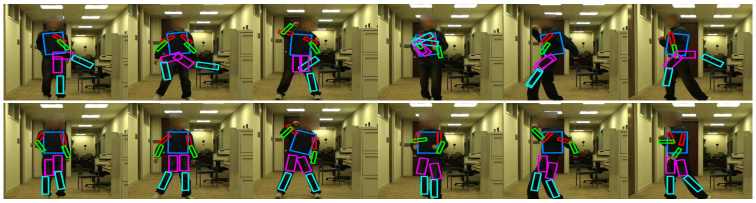


Fig. 4. Comparison with the dynamic programming method. The 1st row: the DP sample results for the lab-man-II sequence. The 2nd row: pose estimation using the proposed method.

estimation as shown in Fig.3. With a "taller" torso rectangle, the head of the subject is also labeled as foreground in the concurrent optimization.

We proceed to compare the proposed method with a variation that optimizes only the body part matching. If the symmetrical part constraint is also discarded, we have a tree structure graph model. Pose estimation with a tree structure body plan can be exactly solved using dynamic programming (DP). Fig.4 shows sample comparison results for the lab-man-II sequence. Without a global constraint, the dynamic programming method often loses detection of arms and legs, and it is easily distracted by the background clutter. The quantitative comparison is shown in Fig.6 and Fig.7. Fig.6 compares the normalized histograms of per-frame errors. Without ground truth, we use visual inspection to verify the results. The criterion is that a correct body part detection should be closely aligned with the corresponding body part or hallucinate on the occluded one. Since there are 9 body parts, the per-frame error number is from 0 to 9. A good performance is characterized by an error histogram that is high in the low error range and low in the high error range. The proposed method yields much better result than the simple DP approach. As shown in Fig.7, the average per-frame errors of the proposed method are less than half of the errors of DP.

It is indeed useful to use simultaneous segmentation to globally constrain the pose optimization. The question is whether other global constraints would work as well. We set out to test whether a simple max-covering global constraint would be sufficient. We label all the superpixels as 1 to introduce a max-covering constraint: the body parts should cover a region as big as possible. This formulation penalizes the overlapping body parts equally and prefers a stretched pose. The sample comparison results are shown in Fig.5. The max-covering method has a difficult time to decide whether to accept a body part candidate or to reject it as clutter because it does not use the clues from image segmentation. As shown in Fig.5, the errors of max-covering include both false

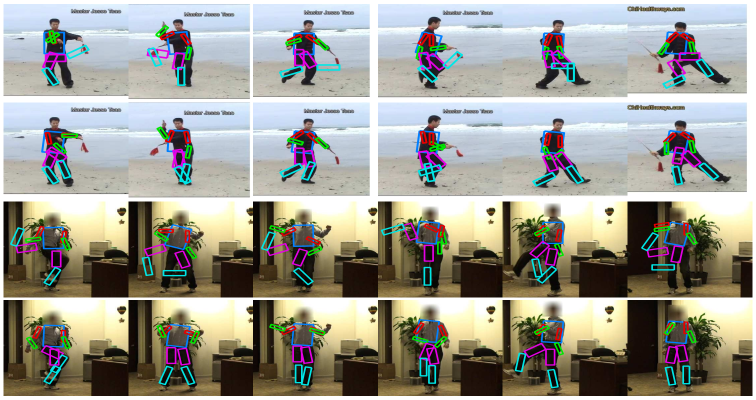
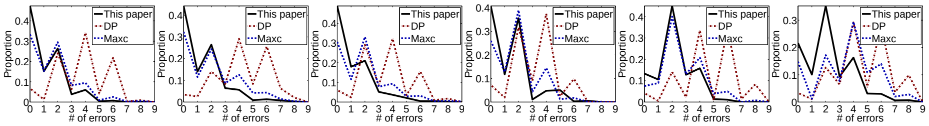


Fig. 5. Comparison with max-covering. The odd rows show the results of max-covering on the taichi and lab-man-I sequence. The even rows show how the proposed method improves the results.



(a) lab-girl (b) lab-man-I (c) lab-man-II (d) lab-man-III (e) taichi (f) dance

Fig. 6. The normalized per-frame error histograms of the proposed method (black solid) with the dynamic programming (DP) (red dash-dotted) and max-covering (Maxc) (blue dotted).

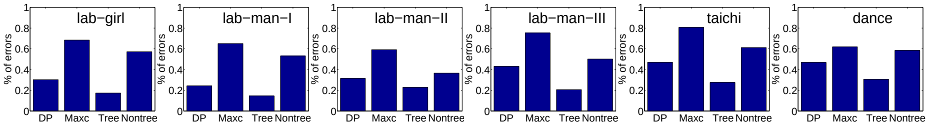


Fig. 7. The ratio of the average per-frame errors of the proposed method to other methods on the test sequences.

positives and false negatives. Simply adjusting the parameters will reduce one class of errors but increase the other. The proposed method does not have such problems and it achieves much better results on the test image sequences. The quantitative comparison in Fig.6 and Fig.7 shows that the proposed method has many fewer errors than max-covering.

The proposed method is indeed better than its variations. But does it have an advantage over other approaches? We first compare the proposed method with the tree inference method [7], a state-of-the-art method for pose detection in videos using a single exemplar. We run the code with [7] on the test videos. The body part detectors are trained from the same walking pose exemplars as those in other testing methods. The sample comparison results are shown in Fig.8. The tree inference method sometimes loses the detections of arms or legs. The proposed method solves the problem by using the global foreground shape constraint. It is also more resistant to clutter. As shown in Fig.8, the proposed method is more reliable in distinguishing two dancers' legs even though they have similar color. The quantitative comparison is shown in Fig.7 and Fig.10.

We further compare the proposed method with a non-tree method [15]. The sample comparison results are shown in Fig.9. The non-tree method uses pairwise prohibition

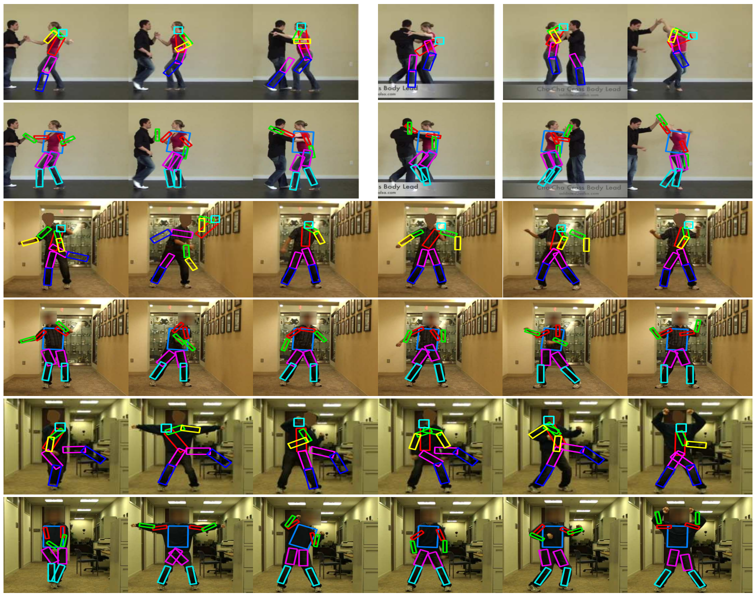


Fig. 8. Comparison with the tree inference method [7]. The odd rows show the results of tree method on the dance, lab-man-III and lab-man-II sequences. The even rows show the results of the proposed method.

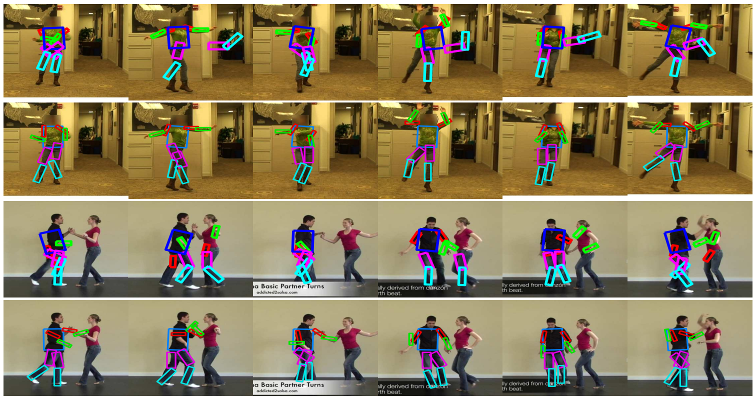
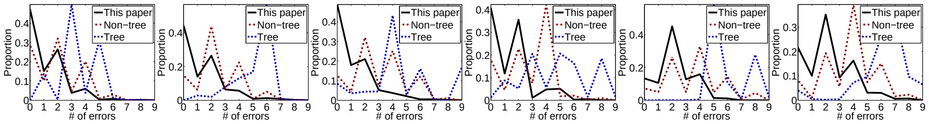


Fig. 9. Comparison with a non-tree method [15]. The odd rows show the results of the non-tree method on the lab-girl and dance sequences. The even rows show the results of the proposed method.

terms to constrain the symmetrical body parts. It uses the same set of body part candidates and costs as the proposed method in the comparison. Compared with the non-tree method, the proposed method works better for complex poses and is more robust in strong clutter. The quantitative results in Fig.7 and Fig.10 confirm the advantage of the proposed method.

Since we optimize both the body part assignment and the foreground superpixel labeling, the byproduct is a rough object foreground estimation. Rows 1-4 in Fig.11 show some sample results. More pose estimation results randomly sampled from the



(a) lab-girl (b) lab-man-I (c) lab-man-II (d) lab-man-III (e) taichi (f) dance

Fig. 10. The normalized per-frame error histograms of the proposed method (black solid) with the tree inference method [7] (blue dotted) and the non-tree method [15] (red dash-dotted).

videos are shown in Fig.11. The proposed method robustly detects poses in the videos. In Fig.11, we also see some part detection errors, especially in the challenging taichi and dance sequences. In our experiments, pose estimation errors are caused mainly by the weak local body part detectors. Using stronger part detectors will further improve the performance.

We also test the proposed method on the ground truth data. Three sequences are selected from the HumanEva [30] dataset. The boxing and walking sequences are downsampled in time while the jogging sequence includes all the frames. All the images are pre-scaled so that the target objects have roughly the same size. These sequences have strong background clutter. The comparison of the proposed method, the non-tree method and the tree methods is shown in Fig.12. Besides the tree method in [7] denoted as tree-I, we also compare with a recent tree method in [28], denoted as tree-II, which

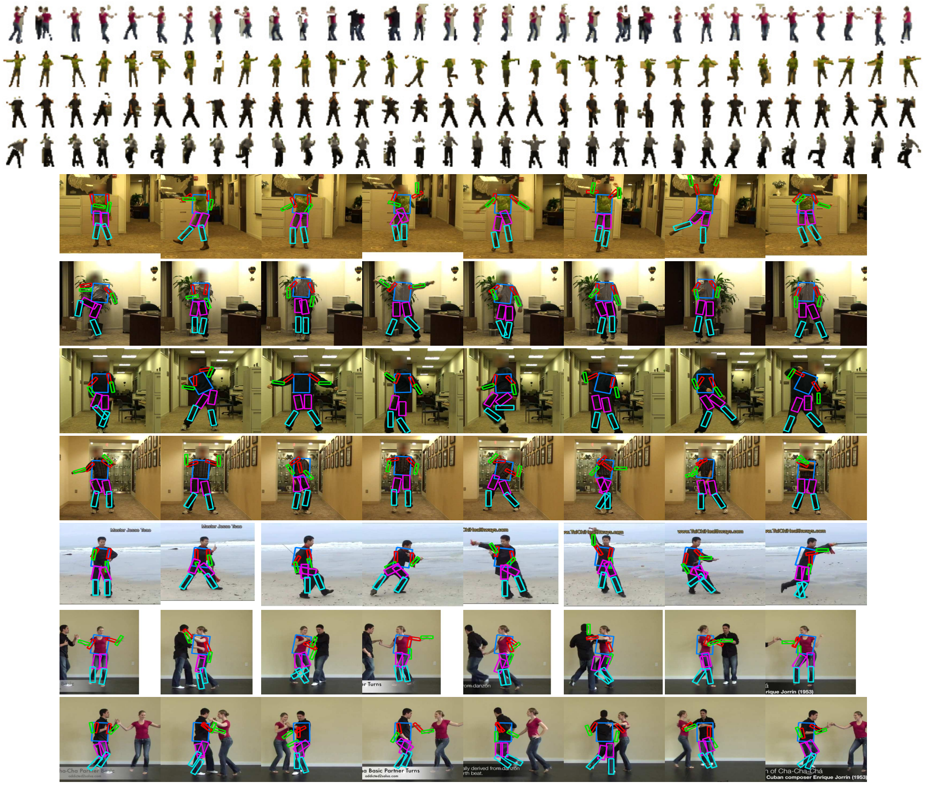
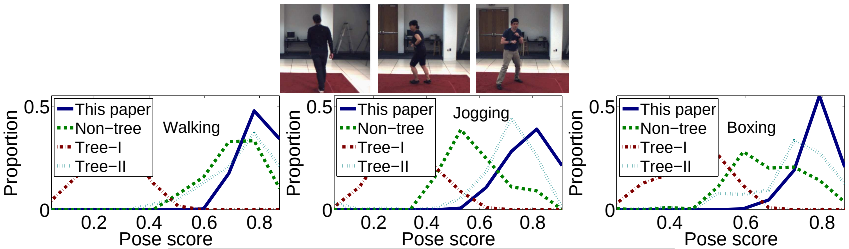


Fig. 11. Pose estimation sample results using the proposed method on the test sequences. Row 1-4: object foreground estimation samples. Row 5-11: random samples from lab-girl (548 frames), lab-man-I (779 frames), lab-man-II (1001 frames), lab-man-III (730 frames), taichi (359 frames), dance(woman) (498 frames) and dance(man) (498 frames).



	This paper	Non-tree[15]	Tree-I[7]	Tree-II[28]
Walk	0.793	0.709	0.263	0.733
Jog	0.784	0.582	0.326	0.685
Box	0.788	0.660	0.460	0.725

Fig. 12. Test on HumanEva walking (222 frames), jogging (362 frames) and boxing (171 frames) sequences. Row 1: sample frames. Row 2: pose score histograms of the proposed method, the non-tree [15], tree-I [7] and tree-II [28] methods. Row 3: average pose scores.

uses more robust body part detectors. Tree-II is a state-of-the-art generic pose detector. For fair comparison, we modify the code for [28] so that color is also used in local part detection and we tune its parameters to achieve the best performance. We quantify the performance of pose estimation by the overlapping area of body parts and the corresponding ground truth. We compute the overlapping area for arms and legs and do not count the easiest torso. The total overlapping area is normalized by the sum of all the ground truth limb areas to form the pose score. Fig.12 compares the histograms of per-frame pose scores and the average pose scores of different methods. The proposed method has the highest pose scores in all the tests. Visual inspection shows consistent result. Our method greatly improves the pose detection results. The performance improvement is not a surprise. Our model enforces a global shape constraint through simultaneous segmentation and therefore all the body parts are related through hypergraph edges. The high order constraint is essential for pose estimation in strong clutter.

4 Conclusion

We propose a novel concurrent optimization method to detect human poses in cluttered videos. With a single exemplar image, the proposed method robustly finds human poses in long video sequences. Concurrently optimizing the body part matching and object segmentation is a great challenge due to its huge search space. We efficiently solve the hard combinatorial problem by novel linear relaxation and branch and bound method. Our experiments on a variety of videos show that the proposed method has a clear advantage over locally constrained methods. The linear approach is also general and it can be extended to generic object matching and segmentation.

Acknowledgement. The project is supported by NSF grant 1018641.

References

[1] Bregler, C., Malik, J., Pullen K.: Twist based acquisition and tracking of animal and human kinematics. IJCV, 56(3), 179-194, 2004.

[2] Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. Inter. J. of Robotics Research, 22(6):371-391, 2003.

[3] Mori, G., Malik, J.: Estimating human body configurations using shape context matching. ECCV 2002.

[4] Gavrila, D.M.: A Bayesian, exemplar-based approach to hierarchical shape matching. TPAMI, 29(8), 2007.

[5] Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter sensitive hashing. ICCV 2003.

[6] Toyama, K., Blake, A.: Probabilistic tracking with exemplars in a metric space. IJCV, 48(1):9-19, 2002.

[7] Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: tracking people by finding stylized poses. CVPR 2005.

[8] Jiang, H.: Human pose estimation using consistent max-covering. ICCV 2009.

[9] Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV 61(1), Jan. 2005.

[10] Ioffe, S., Forsyth, D.A.: Probabilistic methods for finding people. IJCV 43(1):45-68, 2001.

[11] Ren, X.F., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. ICCV 2005, 1:824-831.

[12] M.W. Lee and I. Cohen. "Human upper body pose estimation in static images", ECCV 2004.

[13] Rosales, R., Sclaroff, S.: Inferring body pose without tracking body parts. CVPR 2000.

[14] Sigal, L., Black, M.J.: Measure locally, reason globally: occlusion sensitive articulated pose estimation. CVPR 2006.

[15] Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. CVPR 2008.

[16] Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. ECCV 2008.

[17] Mori, G., Guiding model search using segmentation. ICCV 2005.

[18] 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.

[19] Pawan Kumar, M., Torr, P.H.S., Zisserman, A.: OBJCUT. CVPR 2005.

[20] Ramanan, D.: Learning to parse images of articulated objects. NIPS 2006

[21] Ferrari, V., Manuel, M., Zisserman, A.: Pose search: retrieving people using their pose. CVPR 2008.

[22] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV, vol 59, no. 2, 2004.

[23] Gupta, A., Mittal, A., Davis, L.S.: Constraint integration for efficient multiview pose estimation with self-occlusions. IEEE TPAMI, 30(3), 2008, pp. 493-506.

[24] Urtasun, R., Fleet, D., Fua, P.: Temporal motion models for monocular and multiview 3D human body tracking. CVIU, vol.104, no.2, pp. 157-177, 2006.

[25] Yezzi, A., Zollei, L., Kapur, T.: A variational framework for joint segmentation and registration. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis 2001.

[26] Chen, C., Fan, G.: Hybrid body representation for integrated pose recognition, localization and segmentation. CVPR 2008.

[27] Johnson, S., Everingham, M.: Combining discriminative appearance and segmentation cues for articulated human pose estimation. IEEE International Workshop on Machine Learning for Vision-based Motion Analysis, 2009.

[28] Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: people detection and articulated pose estimation. CVPR 2009.

[29] Tian, T.P., Sclaroff, S.: Fast globally optimal 2D human detection with loopy graph models. CVPR 2010.

[30] HumanEva Dataset. `http://vision.cs.brown.edu/humaneva`.