Scale and Rotation Invariant Approach to Tracking Human Body Part Regions in Videos

Yihang Bo
Institute of Automation, CAS & Boston College
yihang.bo@gmail.com

Hao Jiang
Computer Science Department, Boston College
hjiang@cs.bc.edu

Abstract

We propose a novel scale and rotation invariant method to track a human subject's body part regions in cluttered videos. The proposed method optimizes the assembly of body part region proposals with the spatial and temporal constraints of a human body plan. This approach is invariant to the object scale and rotation changes. To enable scale and rotation invariance, the human body part graph of the proposed method has to be loopy; efficiently optimizing the body part region assembly is a great challenge. We propose a dynamic programming method to solve the problem. We devise a method that finds N-best whole body configurations from loopy structures in each video frame using dynamic programming. The N-best configurations are then used to construct trellises with which we track human body part regions by finding shortest paths on the trellises. Our experiments on a variety of videos show that the proposed method is efficient, accurate and robust against object appearance variations, scale and rotation changes and background clutter.

1. Introduction

Tracking the movement of human subjects in videos has important potential applications in sports, entertainment and surveillance. To understand detailed human movement, we need to go beyond a simple bounding box representation which is widely used in previous work and try to follow the target's body parts through a long video sequence. Human body part tracking is challenging due to object articulation, rotation and scale changes, appearance variation and strong background clutter.

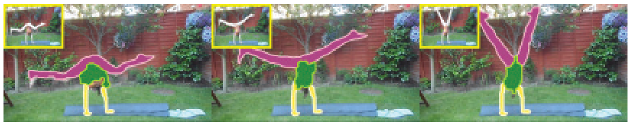

Figure 1. We track human body part regions in videos. The proposed method assembles body part proposal regions and is invariant to the target's scale and rotation. It is able to find the optimal assembly in a single pass optimization. In this figure, arm regions are yellow, legs are magenta and torso is green.

In this paper, we propose a novel method to track the regions of human body parts in cluttered videos. Different from previous methods, the proposed method represents each body part as a region in an image. Instead of trying to fit each body part to a template in different rotations and scales, the proposed method has the advantage that it does not have to quantize the scale and rotation and exhaustively enumerate all the possible cases. By using the relative size of the body parts, we construct a scale and rotation invariant method, which is able to match the target with unknown scale and rotation in an efficient single pass optimization. Our method works in the continuous rotation and scale domain and thus eliminates the quantization errors. In this paper, We also propose an N-best whole body configuration selection method and a dynamic programming solution to optimize the body part region tracking in videos. The experiment results show that our method is more robust than previous pictorial structure methods, especially when the target has large scale or rotation changes. Fig. 1 illustrates the problem we tackle.

Pictorial structure methods are popular methods for human pose estimation and tracking, in which rectangular body part candidates are assembled together to follow a valid body plan. If tree structured body plan is used, efficient dynamic programming method can be used to find the optimal body part assembly. In [1], distance transform method has been proposed to greatly speed up the dynamic programming procedure from being proportional to the square of the number of body parts to a linear complexity. In [2], pictorial structure method is applied to tracking human body parts in cluttered videos. In [4, 3], improved body part detectors based on shape context or histogram of oriented gradients greatly improve the result of the pictorial structure methods. Tree structure formulations have an inherent double counting issue that often causes the miss-detection of one arm or one leg because there are no constraints among the limbs in the tree. Non-tree structure methods have been proposed to tackle the problem. Approximation methods using belief propagation [5], linear programming [6] and dual decomposition [7] have been studied. Pictorial structure methods have also be combined with object foreground segmentation [7, 6] to achieve more reliable results. One of the difficulties of the traditional pictorial methods is that they have to roughly know the target scale or at least the range so that we can quantize the scales and search in each of them. If the target object's scale is unknown, searching through numerous possible cases for the optimal result is slow. In this paper, instead of using rectangular body parts, we use *regions* to represent each body part. Such a scheme enables us to build an efficient method which is invariant to scale and rotation changes.

Our proposed method is related to region tracking. Video segmentation [8] yields 3D superpixels in videos and these superpixels do not directly correspond to human body parts but just homogeneous color regions. The region partition result also greatly depends on the parameters which are hard to set correctly to segment all the desired body parts. Previous region tracking methods [9] and [10] also do not explicitly give the human body parts.

There have been few previous methods that explicitly track human body part regions. In [13], a region based method is proposed to find the body part regions of pedestrians in images. This method matches the specific shape of the upper and lower body of a pedestrian. It is therefore hard to be generalized to detect complex human poses in cluttered videos. In [11], legs and torso are extracted from cluttered images using region merging and shape parsing on multiple level image segmentation. This method targets at human pose detection in a single image. In [12], a linear method is proposed to find people and their body parts by assemble region proposals [14] in images. This method has no mechanism to enforce that part region detection is consistent through multiple frames in videos. In this paper, our proposed method tracks human body part regions in cluttered videos.

Our proposed method is also related to the N-best method [15], which finds the top N-best human poses on a tree human body plan and uses dynamic programming to optimize the tracking. The difference of our proposed method to the N-best method is that we track regions instead of rectangular body parts. Our method is scale and rotation invariant, while the N-best method needs to know roughly the object's scale. We also extract the N-best whole body configurations in each video frame, but we use a loop graph body plan while [15] uses a tree structure. Our method also explicitly controls the smooth movement of each body part while the method in [15] only enforces that the overall object silhouette evolves smoothly.

The contribution of this paper is a novel method to track human body part regions in videos. We first find body part region candidates and then optimize the body part region assembly so that they satisfy relative position, size, symmetry constraint in each frame and motion continuity constraint across multiple video frames using dynamic programming. The proposed method is able to track human poses in complex videos and it is scale and rotation invariant.

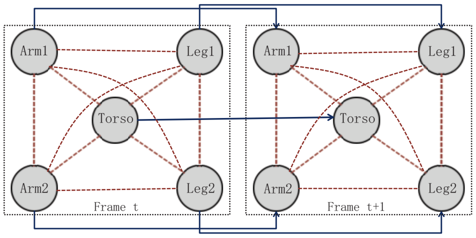

Figure 2. Body part interaction graph of two adjacent frames. The nodes represent human body parts, red edges represent the spatial constraints and blue edges represent temporal constraints among body parts.

2. Method

2.1. Overview

Our model uses a human body plan with five body parts: a torso, two arms and two legs. The coupling between the body parts in each video frame and between successive video frames follows the graph in Fig. 2. Each node in the graph indicates a body part and the edges indicate the interactions among these body parts. We not only include the arm-torso, leg-torso, arm-arm and leg-leg constraints, we also enforce the coupling between arms and legs. The graph edges also connect the corresponding body parts between successive video frames to enforce that each part evolves smoothly through time.

We formulate body part region tracking into an optimization problem: we assign a body part region to each of the graph node so that the assignment cost is minimized. In this paper, the body part region candidates are obtained from the object class independent proposals [14]. These region proposals are randomly merged superpixels whose overall shape has a high objectness score. Different from homogeneous superpixels, the region proposals have a high chance to include correct body parts even if the target subject wears uniform color clothing. The proposed method assembles the chosen region proposals to form a valid human configuration.

For each node assignment, we obtain an estimated whole body region configuration. The cost of a whole body region configuration contains several terms as shown in Eq. (1). Let f_k be the body part region configuration in frame k and n is the number of video frames.

$$E(f) = \min(\underbrace{\sum_{k=1}^{n}(\alpha \cdot P(f_k) + \beta \cdot G(f_k) + \gamma \cdot O(f_k) + \delta \cdot A(f_k))}_{\text{Intra-Frame Energy}}$$
$$+ \underbrace{\sum_{k=2}^{n}(\eta \cdot S(f_k, f_{k-1}) + \phi \cdot L(f_k, f_{k-1}) + \theta \cdot H(f_k, f_{k-1})))}_{\text{Inter-Frame Energy}} \quad (1)$$

The intra-frame spatial terms include the body part region shape matching cost $P(f_k)$, distance between pairs of body parts $G(f_k)$, body part overlap region $O(f_k)$ and body part area ratio term $A(f_k)$. These intra-frame terms are small if each body part region has the right shape, connected part regions have small distance, body parts have small overlap, and the area ratio between body parts conform to a body plan. The inter-frame temporal terms enforce the smooth transition of each body part region from one frame to the next. $S(f_k, f_{k-1})$ enforces the shape similarity of regions in successive frames. $L(f_k, f_{k-1})$ quantifies the weight center position changes and $H(f_k, f_{k-1})$ quantifies the color histogram changes. Coefficients α, β, γ, δ, η, ϕ and θ are constant coefficients to control the weight among different terms.

2.2. Formulation

In the following, we elaborate the details of the proposed formulation.

Body part shape matching cost (P): We first extract region candidates using object class independent proposals [14]. Each region is a potential candidate for a specific human body part. Each body part, such as an arm or a leg, has a sequence of templates. We quantify the similarity of the shape of a region proposal to a template using the Euclidean distance between their corresponding inner distance shape descriptors [16]. The shape descriptor of a region is defined as the distance histogram between each pair of the points inside the region. When computing the histogram, we normalize these distances with the longest distance between two points in the region. We further normalize the inner distance histogram with the number of distance pairs. This shape descriptor is scale and rotation invariant. It is also roughly articulation invariant.

The overall part assignment cost P is defined as:

$$P(f_k) = \sum_{i} c(i, f_k(i))$$

where i is the index of a body part, $f_k(i)$ denotes the chosen region candidate for part i, and $c(i, f_k(i))$ is the cost of assigning region candidate $f_k(i)$ to body part i. c is computed as the shortest distance of region $f_k(i)$'s shape descriptor to those of the templates of part i.

To reduce the number of region candidates, we use RANSAC to partially remove the background clutter. We match the SIFT features in each video frame to a number of previous frames and future frames. Since target object often only occupies a small portion of each video frame, the SIFT features on the human subjects are outliers in the RANSAC feature matching in which we use an affine global motion model. Using only the inlier matching, we warp the surrounding video frames to the current frame and we compute the median of the stack of these images on each image pixel. The result is an estimated background with which a background subtraction can be used to extract a rough object foreground. Note that this partial background removal procedure is optional.

Parts distance (G): Apart from enforcing that each body part region has the right shape, we further ensure that connected body part regions should not be far away but tend to have small distance. Let t be the index of torso and j to be the index of limbs. We compute the minimum boundary distance $d(f_k(j), f_k(t))$ between limb j and torso. The part distance term is:

$$G(f_k) = \sum_{j \in \mathcal{L}} d(f_k(j), f_k(t))$$

where $\mathcal{L}$ is the set of limbs.

Parts overlap (O): We penalize the overlap between body part regions with term O. This penalty term pushes body part regions away so that we can find both arm and leg regions. This is a soft term; overlapped parts are still allowed.

$$O(f_k) = \sum_{\{i,j\} \in \mathcal{N}} \frac{\mathcal{A}(F_k(i) \cap F_k(j))}{\mathcal{A}(F_k(i) \cup F_k(j))}$$

where $F_k(i)$ is the estimated region for part i in video frame k, $\mathcal{N}$ is the body part pair set, which includes the arm-arm, leg-leg, arm-torso, leg-torso and arm-leg pairs, and function $\mathcal{A}$ gives the region area.

Parts size ratio (A): Different body parts, such as an arm and a leg, may have similar shape descriptors. We need more clues to increase the chance of a correct region assignment. We enforce the correct size ratios between pairs of body parts. We model part size ratio by a Gaussian distribution, which is estimated by using a training set.

$$A(f_k) = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} \frac{(r(f_k(i), f_k(j)) - \mu_{i,j})^2}{\sigma_{i,j}^2}$$

Here $r(f_k(i), f_k(j))$ is the area ratio of region candidate $f_k(i)$ for part i to region candidate $f_k(j)$ for part j, $\mu_{i,j}$ and $\sigma_{i,j}^2$ are the mean and variance of the Gaussian distribution. $\mathcal{P}$ is the set of body parts.

Apart from the intra-frame body part layout terms, we further enforce that the body parts move continuously through the video. The temporal consistency terms are defined as follows.

Shape consistency across frames (S): The shape of each body part does not change rapidly between adjacent video frames. The whole silhouette of the estimated whole body region of the target also should change smoothly from frame to frame. We introduce the shape consistency term S to enforce the smooth transition. Here, the shape of a region is quantified by the histogram of its boundary orientations [17]. Note that we do not use the inner distance descriptor here because we do not require the articulation invariant property. The non-invariant boundary orientation histogram is stronger and more suitable than the inner distance feature in the non-invariant application.

Let $\mathbf{s}_{f_k(i)}$ be the shape descriptor of region candidate $f_k(i)$ for part i in frame k. We use $\mathbf{s}_{f_k}$ to indicate the shape descriptor for the whole estimation foreground region in frame k. The shape consistency term is defined as:

$$S(f_k, f_{k-1}) = \sum_{i \in \mathcal{P}} \|\mathbf{s}_{f_k(i)} - \mathbf{s}_{f_{k-1}(i)}\| + \|\mathbf{s}_{f_k} - \mathbf{s}_{f_{k-1}}\|$$

Note that the boundary orientation histogram is not normalized; it thus also contains the region size information. By minimizing S, we enforce the shape and size consistency of the estimated target through multiple video frames.

Location consistency (L): Similar to the shape consistency, we also prefer that body part location does not change abruptly in successive frames. We penalize large displacement of the corresponding body part centroids in successive video frames. Let $l_{f_k(i)}$ be the centroid of part candidate $f_k(i)$ for i in frame k. Part position term is defined as

$$L(f_k, f_{k-1}) = \sum_{i \in \mathcal{P}} \|l_{f_k(i)} - l_{f_{k-1}(i)}\|$$

Color consistency (H): We assume that the appearance of the target does not change abruptly through time. The color consistency term H enforces that body part color to be stable in successive video frames. We use RGB histogram to quantify the color of the human body parts. The color term H is defined as:

$$H(f_k, f_{k-1}) = \sum_{i \in \mathcal{P}} \|\mathbf{h}_{f_k(i)} - \mathbf{h}_{f_{k-1}(i)}\|$$

where $\mathbf{h}_{f_k(i)}$ is the color histogram of the chosen candidate region for part i in frame k.

By combining these items, we obtain a complete energy function. We search for the assembly f of body part region proposals to minimize the energy function. The proposed model is non-tree. Therefore, dynamic programming cannot be directly used. Direct exhaustive search is not an option because there is a huge set of feasible body part configurations. In the following, we propose a method to transform the problem so that we can use dynamic programming to find an approximate solution.

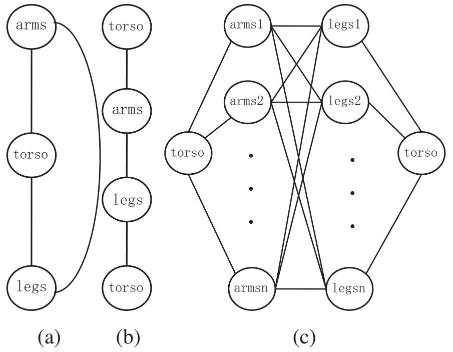

Figure 3. (a) If we treat two arms as an arm node and two legs as a leg node, the body part interaction graph forms a circle. (b) For dynamic programming, we break the torso node into two nodes on both sides of a chain. (c) The trellis for a torso candidate. We find the shortest paths on the trellis to obtain the top N-best whole body configurations.

2.3. Optimization by Dynamic Programming

As shown in the body graph in Fig. 2, the body part regions are coupled to each other not only within each video frame but also between successive video frames. The body graph is thus non-tree. However, if we treat each sub-graph in a single frame as a node, the meta-graph forms a chain, for which dynamic programming can be used to optimize the body part region assignment.

Unfortunately, there are too many possible body part configurations of the whole body estimation. If we do not prune the whole body part configurations in each single frame, the dynamic programming on the chain meta-graph would be too big to be solved since its complexity is proportional to the square of the number of the whole body configurations in each video frame. We devise an efficient method to find the N-best whole body configurations in each video frame, where N is a relatively small number that DP is able to handle. This is similar to the spirit of the approach in [15]. But, since our sub-graph in each single frame is loopy, the N-best method in [15] cannot be used here. We propose a new method to extract the N-best configurations for a loopy structure.

Let's take a closer look at the body plan graph. If we treat two arms as a single node, and two legs as a single node, the meta-graph is in fact a circle as shown in Fig. 3 (a). We can further convert the circle into a chain in which the torso node appears twice on both ends of the chain as shown in Fig. 3 (b). We can now apply dynamic programming to find N-best configurations on the chain with a constraint that the two torso nodes have to select the same region candidate. The torso consistency condition is not too much of a problem. We simply fix the torso candidate for both ends of the chain and then run a standard dynamic programming. For each fixed torso candidate, we keep the sorted N-best whole

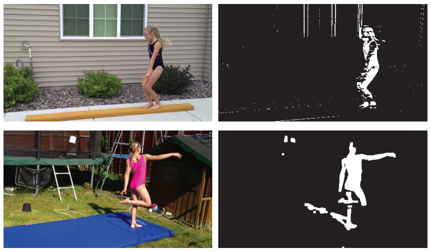

Figure 4. Examples of background removal results. Left column shows the original video frames and the right column shows the foreground images.

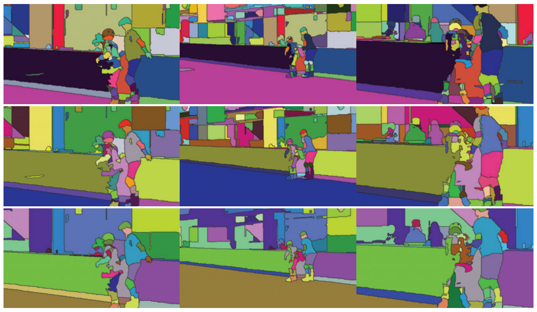

Figure 5. Examples of video segmentation results with three different thresholds from row one to row three.

body region configurations. We further merge the sorted N-best regions for all the fixed torso candidates to obtain the final N-best whole body configurations. Assuming that there are K torso candidate regions, with a divide and conquer algorithm that recursively merges half of the solution sets, the merging can be completed in $O(NK)$ time. It is easy to verify that this method indeed finds the N-best whole body configurations.

With the N-best whole body configurations, we generate a trellis network whose nodes correspond to each whole body configuration in each video frame. We connect the nodes between successive video frames with edges. Each node has a node cost that is determined by the intra-frame cost terms P, G, O, A and each edge has edge cost that is determined by the inter-frame cost terms S, L, H. The optimization in Eq.(1) is thus equivalent to finding a path whose total node cost and edge cost is minimum. The shortest path problem can be efficiently solved by dynamic programming.

3. Experimental Results

We apply our method to a set of challenging test video sequences, which involve complex human poses and actions. The first four videos are Youtube videos and the last one is from the HumanEVA dataset [18]. In the following, we show both the qualitative results and ground truth comparison results.

3.1. Qualitative Results

In the experiment, we extract human body part candidates from the object class independent proposal regions [14]. We partially remove the background clutter using a RANSAC method as discussed in section 2.2. This method does not completely remove the clutter. It also does not remove the shadow of the human target, which will be treated as outliers in the matching. Fig. 4 shows two examples of background removal results. The partial background removal method eliminates a lot of background clutter, but we still do not have a clean target foreground and it is challenging to detect the human body part regions. We further apply the proposed method to extract the human body part regions. Fig. 6 shows our tracking results on the five test videos. The frames in Fig. 6 are regularly sampled from the results. Our method reliably tracks human body part regions in these challenging videos.

Fig. 5 shows the video segmentation [8] results for one of the test sequences with different parameter settings. As shown in this example, video segmentation has a hard time to determine which parameter to use to detect different body part regions. In a long video sequence, video segmentation also tends to break a long narrow 3D superpixel and thus loses tracking of a region. In contrast, the proposed method explicitly estimates the body part regions and can be used to reliably track human body part regions in long video sequences.

3.2. Quantitative Results

We compare our method with the N-best method [15] using the ground truth body part region labeling of the five test videos. For fair comparison, we use the same background clutter removed images as the input to the N-best method. N-best method [15] is not scale invariant; we give the method the advantage of knowing the correct object scale. Our proposed method is scale and rotation invariant; it does not use the scale information.

Since the N-best method outputs a stick figure detection, we first convert the stick figure result into the body part region format for further comparison. The body part regions for the N-best method are obtained by thickening the lines between the end points of each body part with a proper width. For each body part region detection, we define the detection score as $\mathcal{A}(P \cap G)/\mathcal{A}(P \cup G)$, where P is the region of the body part detection and G is the region of the ground truth region of the corresponding body part, $\mathcal{A}$ gives the area of a region.

Table 1 shows the region detection score comparison of the proposed method with the N-best method. Our proposed method shows great improvement for almost all the test cases. The overall average part detection score of the proposed method is always higher than the average part detection score of the N-best method.

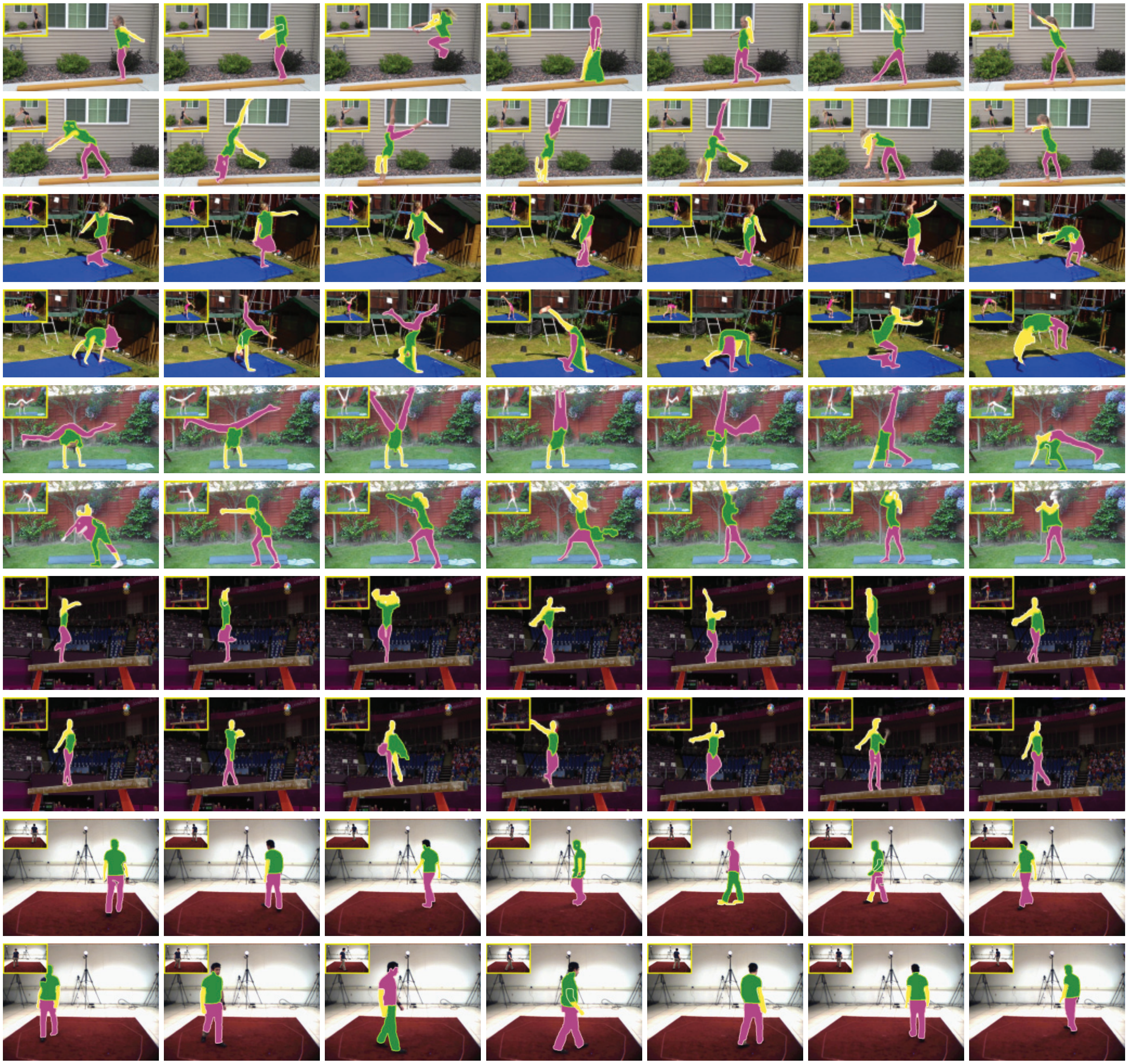

Figure 6. Sample results of the proposed methods on five test videos. These test videos are challenging. They have different human body movements such as jumping, flipping, handstand, flat turn and walking. Our method reliably detects and tracks the human body parts in these test videos. Torsos are green, arms are yellow, and legs are magenta. Original videos frames are embedded at the top left corner of the images.

Fig. 7 illustrates the detection ratio curves of the proposed method and the N-best method. Each detection curve shows the proportion of the "correct" detected body parts with respect to a threshold. A detection is deemed correct if its detection score is greater than the threshold. Therefore, the detection rate is 0 if the threshold is 1 and 1 if the threshold is 0. Fig. 7 shows that the proposed method gives better performance than the N-best method in terms of detection rate.

4. Conclusion

We propose a novel human part regions tracking method. The proposed method does not require initialization and is

	N-best Arms	Ours Arms	N-best Legs	Ours Legs	N-best Torso	Ours Torso	N-best All	Ours All	N-best Mean	Ours Mean
Video1	13.96	**25.90**	**45.30**	37.37	24.99	**40.31**	45.70	**62.45**	32.49	**41.51**
Video2	12.15	**32.49**	24.71	**43.87**	42.61	**56.41**	38.47	**62.43**	29.49	**48.80**
Video3	12.62	**25.00**	42.69	**42.99**	**45.41**	44.03	48.75	**67.98**	37.37	**45.00**
Video4	22.54	**25.93**	44.76	**54.29**	51.20	**53.81**	50.21	**67.77**	42.18	**50.45**
Video5	22.29	**56.10**	**65.32**	64.17	49.75	**63.18**	62.96	**84.58**	50.08	**67.01**
Mean	16.71	**33.08**	44.56	**48.54**	42.79	**51.55**	49.22	**69.04**	38.32	**50.55**

Table 1. Comparison of the average scores of the N-best [15] and the proposed method. The values show average detection scores scaled by 100. Each estimated body part region is compared against the ground truth labeling and the score is defined as the ratio of their region intersection to that of the region union.

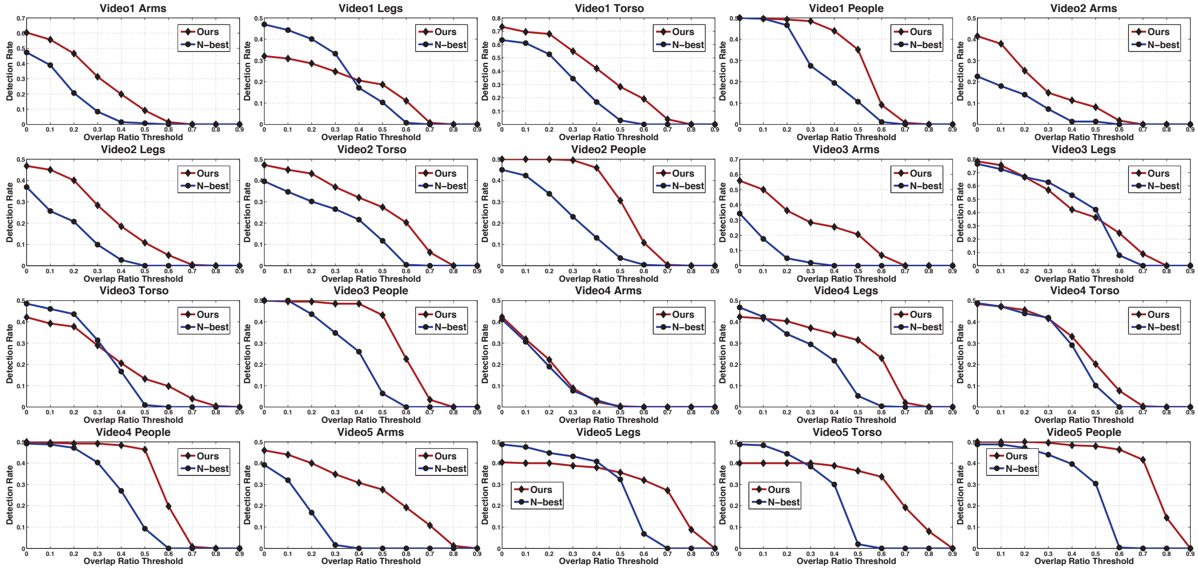

Figure 7. Comparison of the detection rate comparisons of the N-best [15] and the proposed method.

scale and rotation invariant. Our experiments show that the proposed method gives more reliable results when tracking people with different clothing and unconstrained movement than competing methods. Our method is also efficient. We believe that it is a useful tool for many applications such as human movement understanding, surveillance and human computer interaction.

Acknowledgment: This project is supported by the United States NSF funding 1018641. Yihang Bo is supported in part by China National Natural Science Foundation funding 61202327.

References

[1] P. Felzenszwalb, D. Huttenlocher, “Efficient graph-based image segmentation” International Journal of Computer Vision, Vol. 59, No. 2, September 2004.

[2] D. Ramanan, D. A. Forsyth, A. Zisserman, “Tracking People by Learning their Appearance” IEEE Trans. PAMI. Jan 2007.

[3] Y. Yang, D. Ramanan, “Articulated Pose Estimation using Flexible Mixtures of Parts”, CVPR 2011.

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

[5] L. Sigal and M. J. Black, “Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation”, CVPR, 2006.

[6] H. Jiang, “Human Pose Estimation Using Consistent Max-Covering”, IEEE Trans. PAMI, 2011.

[7] B. Sapp, D. Weiss and B. Taskar, “Parsing Human Motion with Stretchable Models”, CVPR 2011.

[8] M. Grundmann and V. Kwatra and M. Han and I. Essa, “Efficient Hierarchical Graph Based Video Segmentation”, CVPR 2010.

[9] Y. J. Lee, J. Kim, and K. Grauman, “Key-Segments for Video Object Segmentation”, ICCV 2011.

[10] D. Tsai and M. Flagg and J.M.Rehg, “Motion Coherent Tracking with Multi-label MRF optimization”, BMVC 2010.

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

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

[13] Y. Bo and C.C. Fowlkes, “Shape-based pedestrian parsing”, CVPR 2011.

[14] I. Endres and D. Hoiem, “Category independent object proposals”, ECCV 2010.

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

[16] H. Ling and D. W. Jacobs, “Shape classification using the inner-distance”, IEEE Trans. PAMI, vol.29, no.2, 2007.

[17] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, CVPR 2005.

[18] L. Sigal and M. J. Black, “HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion”, Technical Report CS-06-08, Brown University, 2006.