Human Pose Estimation in Videos

Dong Zhang, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida
Orlando, Florida, USA
dzhang@cs.ucf.edu, shah@crcv.ucf.edu

Abstract

*In this paper, we present a method to estimate a sequence of human poses in unconstrained videos. In contrast to the commonly employed **graph** optimization framework, which is NP-hard and needs approximate solutions, we formulate this problem into a unified two stage **tree-based** optimization problem for which an efficient and exact solution exists. Although the proposed method finds an exact solution, it does not sacrifice the ability to model the spatial and temporal constraints between body parts in the video frames; indeed it even models the symmetric parts better than the existing methods. The proposed method is based on two main ideas: 'Abstraction' and 'Association' to enforce the intra- and inter-frame body part constraints respectively without inducing extra computational complexity to the polynomial time solution. Using the idea of 'Abstraction', a new concept of 'abstract body part' is introduced to model not only the tree based body part structure similar to existing methods, but also extra constraints between symmetric parts. Using the idea of 'Association', the optimal tracklets are generated for each abstract body part, in order to enforce the spatiotemporal constraints between body parts in adjacent frames. Finally, a sequence of the best poses is inferred from the abstract body part tracklets through the tree-based optimization. We evaluated the proposed method on three publicly available video based human pose estimation datasets, and obtained dramatically improved performance compared to the state-of-the-art methods.*

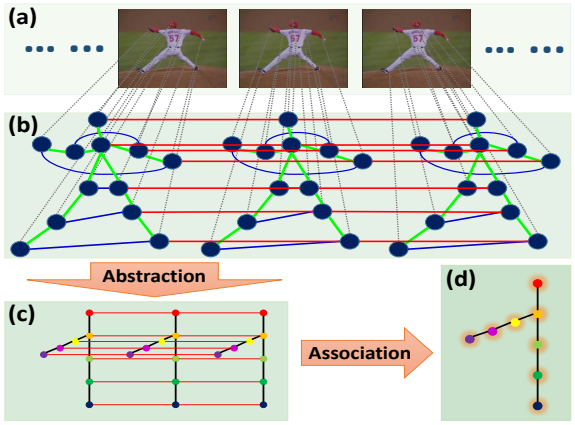

Figure 1. An abstract high-level illustration of the proposed method aiming at removing simple cycles from the commonly employed graph optimization framework for video based human pose estimation problem. All of the above graphs are relational graphs for the problems. In (b), each body part in each frame is represented by a node. Green and blue edges represent relationships between different body parts in the same frame (green ones are commonly used edges in the literature, and **blue ones are important edges for symmetric parts**); red edges represent the consistency constraints for the same body part in adjacent frames. Note that this is only an illustration and not all edges are shown. In the 'Abstraction' stage, symmetric parts are combined together, and the simple cycles within each single frame are removed (shown in (c)); and in the 'Association' stage, the simple cycles between adjacent frames are removed (shown in (d)).

1. Introduction

Human pose estimation is crucial for many computer vision applications, including human computer interaction, activity recognition and video surveillance. It is a very challenging problem due to the large appearance variance, non-rigidity of the human body, different viewpoints, cluttered background, self occlusion etc. Recently, a significant progress has been made in solving the human pose estimation problem in unconstrained single images ([39, 23, 36]); however, human pose estimation in videos ([19, 22, 5]) is a relatively new and challenging problem, which needs significant improvement. Obviously, single image based pose estimation method can be applied to each video frame to get an initial pose estimation, and a further refinement through frames can be applied to make the pose estimation consistent and more accurate. However, due to the innate complexity of video data, the problem formulations of most video based human pose estimation

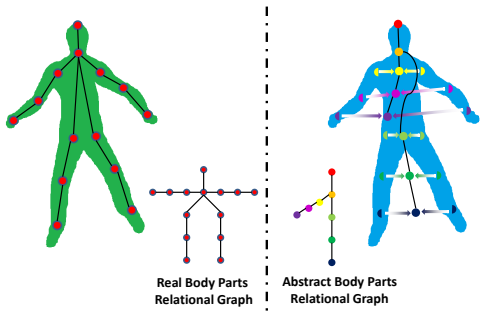

Figure 2. Real body parts vs. abstract body parts. The left side shows a commonly used body part definitions in the literature, and we call these body parts (nodes) 'real body parts', and the graph 'real body part relational graph'. The right side shows the proposed new definition of body parts, basically we combine a pair of symmetric body parts to be one body part, we call these body parts (nodes) 'abstract body parts', since the parts are some abstract concepts of parts, not real body parts, and the graph as 'abstract body part graph'.

methods are very complex (usually NP-hard), therefore, approximate solutions have been proposed to solve them which result in sub-optimal solutions. Furthermore, most of the existing methods model body parts as a tree structure and these methods tend to suffer from double counting issues ([23]) (which means symmetric parts, for instance left and right ankles, are easily to be mixed together). In this paper, we aim to formulate the video based human pose estimation problem in a different manner, which makes the problem solvable in polynomial time with an exact solution, and also effectively enforces the spatiotemporal constraints between body parts (which will handle the double counting issues).

One commonly employed methodology for human pose estimation in videos is the graph optimization formulation. There are two types of such formulation. **The first type** of this formulation ([19]) is to generate several human pose hypotheses in each frame and select one best hypothesis from each frame, while making sure they are consistent throughout the video. The inference in this approach is very efficient, however, due to the large variations of pose configurations, it is very difficult to get good poses with all body parts correctly estimated. Therefore, **the second type** of such formulation ([22, 34, 5]) was introduced to handle each body part separately (please see Fig.1(b)). In this formulation, hypotheses are generated for each body part in every frame. Following the spatial constraints between body parts in each frame and using temporal consistency of appearances and locations between adjacent frames, the goal is to optimally select the best hypotheses for each body part from all the frames together. This formulation is desirable, since it is able to expand sufficient diverse human pose configurations, and also, it is able to effectively model spatiotemporal constraints between body parts. Despite all the benefits of this formulation, it is an NP-hard problem due to the underlying loopy graph structure (i.e. there are many simple cycles in the graph; e.g. the simple cycles in Fig.1(b) induced by the green, blue and red edges) ([22]). Several methods were proposed to attack this NP-hard problem in different ways. To reduce the complexity induced by inter-frame simple cycles, Tokola *et.al* [34] proposed a tracking-by-selection framework, in which each body part is tracked separately and parts are then combined at a later stage. Authors in [5] proposed an approximate method which focuses on less-certain parts in order to reduce the complexity. Ramakrishna *et. al* [22] introduced a method which takes symmetric parts into account and proposed an approximate solution to handle the loopy graph. And in [27], the original model is decomposed into many sub-models which are solvable, since the sub-models have a tree-based structure. All of the above methods are insightful, however, none of them has simultaneously exploited the important constraints between body parts (e.g. symmetry of parts) and has an efficient exact solution.

Based on the discussion above, the major issue is: *How to exploit the spatial constraints between the body parts in each frame and temporal consistency through frames to the greatest possible extent, with an efficient exact solution?* Since the inference of a tree-based optimization problem has a polynomial time solution ([39, 41]), the main issue becomes (please refer to Fig.1): How to formulate the problem in order to model the useful spatial and temporal constraints between body parts among the frames without inducing simple cycles?

We propose two key ideas to tackle this issue, which approximate the original **fully connected** model into a simplified **tree-based** model. The first idea is **Abstraction**: in contrast to the standard tree representation of body parts, we introduce a new concept, *abstract body parts*, to conceptually combine the symmetric body parts (please refer to Fig.2, and details are given in Section 3.2). This way, we take advantage of the symmetric nature of the human body parts without inducing simple cycles into the formulation. The second idea is **Association**, using which we generate optimal tracklets for each abstract body part to ensure the temporal consistency. Since each abstract body part is processed separately, it does not induce any temporal simple cycles into the graph.

The proposed method is different from the state-of-the-art methods ([22, 34, 5, 27]) in the following ways: [22] exploits the symmetric nature of body parts, however, the problem is formulated as a multi-target tracking problem with mutual exclusions, which is NP-complete and only approximate solutions can be obtained by relaxation; the method in [34] is designed

to remove the temporal simple cycles from the graph shown in Fig.1(b) to track upper body parts, however, the employed junction tree algorithm will have much higher computational complexity if applied to full-body pose estimation, since there are many more simple cycles induced by symmetric body parts; compared to [5], the proposed method has no temporal simple cycles; and in contrast to [27], our method can model symmetric body part structure more accurately rather than settling for the sub-models. *Therefore, the proposed method ensures both spatial and temporal constraints without inducing any simple cycles into the formulation and an exact solution can be efficiently found by dynamic programming.*

The organization of the rest of the paper is as follows. After discussing related work in Section 2, we introduce the proposed method in Section 3. We present experimental results in Section 4 and conclude the paper in Section 5.

2. Related Work

A large body of work in human pose estimation have been reported over the last few years. Early works are focused on human pose estimation and tracking in controlled environment ([29]); there is also some important work using depth images ([28]). Single image based human pose estimation ([39, 36, 6, 37]) in unconstrained scenes has progressed dramatically in the last a few years; however, video based human pose estimation in unconstrained scenes is still in a very early stage, and some pioneer research ([22, 19, 34, 5]) has been conducted only recently.

For image based human pose estimation in unconstrained scenes, most work has been focused on pictorial structure models ([2, 3, 13, 24]) for quite long time and the performance has been promising. In [39], a flexible mixture-of-parts model was proposed to infer the pose configurations, which showed very impressive results. A new scheme was introduced in [14] to handle a large number of training samples, which resulted in significant increase in pose estimation accuracy. Authors in [31, 40, 21] attempted to estimate 3D human poses from a single image. The high order dependencies of body parts are exploited in [16, 11, 12, 32, 30, 38, 15, 33, 20, 23]. The authors in [33] proposed a hierarchical spatial model with an exact solution, while [20] achieves this by defining a conditional model, and [23] employs an inference machine to explore the rich spatial interactions among body parts. A novel, non-linear joint regressor model was proposed in [6], which handles typical ambiguities of tree based models quite well. More recently, deep learning ([36, 35, 18, 4]) has also been introduced for human pose estimation.

For video based human pose estimation in unconstrained scenes, some early research adopted the tracking-by-detection framework ([1, 17, 25]). More recently, some methods ([5, 34, 42, 9, 26, 27]) have mainly focused on upper body pose estimation and other methods ([19, 22]) have focused on full body pose estimation.

The method proposed in this paper aims to estimate full body poses in the video, without reducing its ability to model symmetric body parts ([19]); and it has lower computational complexity compared to [34], and gives an exact solution using dynamic programming compared to approximate solutions in [22, 5].

3. Tree-based Optimization for Human Pose Estimation in Videos

We formulate the video based human pose estimation problem into a *unified* tree-based optimization framework, which can be solved efficiently by dynamic programming. In view of the major steps shown in Fig.3, we introduce the general notions of relational and hypothesis graphs, and related problem formulation and solutions in Section 3.1; we discuss the new concept: 'abstract body parts' in comparison to 'real body parts' in section 3.2 and show how to generate body part hypotheses in each frame in Section 3.3; we introduce tracklets generation in Sections 3.4 and 3.5, and finally show how to obtain the optimal poses in Section 3.6.

3.1. Relational Graph vs. Hypothesis Graph

In computer vision, and several other disciplines, many problems ([10]) can be abstracted as follows. Assume there is a set of entities $\mathcal{E} = \{e^i|_{i=1}^N\}$, where each entity can only be in one of the many states $\mathcal{S} = \{s^k|_{k=1}^M\}$, with the unary scoring functions $\{\Phi(e^i, s^k)|e^i \in \mathcal{E}, s^k \in \mathcal{S}\}$, which gives the likelihood that an entity e^i is in state s^k. And there is a binary compatibility function for each pair of entities $\{\Psi(e^i, e^j, s^k, s^l)|e^i, e^j \in \mathcal{E}, s^k, s^l \in \mathcal{S}\}$, which represents the compatibility of entity e^i in state s^k and entity e^j in state s^l. The goal then is to determine the best states for each entity such that all of them have high unary scores and they are also compatible with each other. This problem can be modeled as a graph optimization problem formulated by relational and hypothesis graphs, which is described next.

A **relational graph**, $G_r = (V_r, E_r)$, represents the relationship of a set of entities which are represented by entity nodes $\{v_r^i|_{i=1}^{|V_r|}\}$, and the relationships between pairs of entities are represented by edges E_r . Examples of relational graph are shown in Fig.2, Fig.4(a) and Fig.5(a). Corresponding to a relational graph G_r, a **hypothesis graph**, $G_h = (V_h, E_h)$, can be built. For an entity node v_r^i in V_r, a *group* of hypothesis nodes $V_{h(i)} = \{v_{h(i)}^k|_{k=1}^{|V_{h(i)}|}\}$ are generated to form the hypothesis graph, so $V_h = \bigcup_{i=1}^{|V_r|} V_{h(i)}$. The hypothesis nodes represent the possible states of each entity, and in this paper they represent possible locations of body parts. Hypothesis edges, $E_h =$

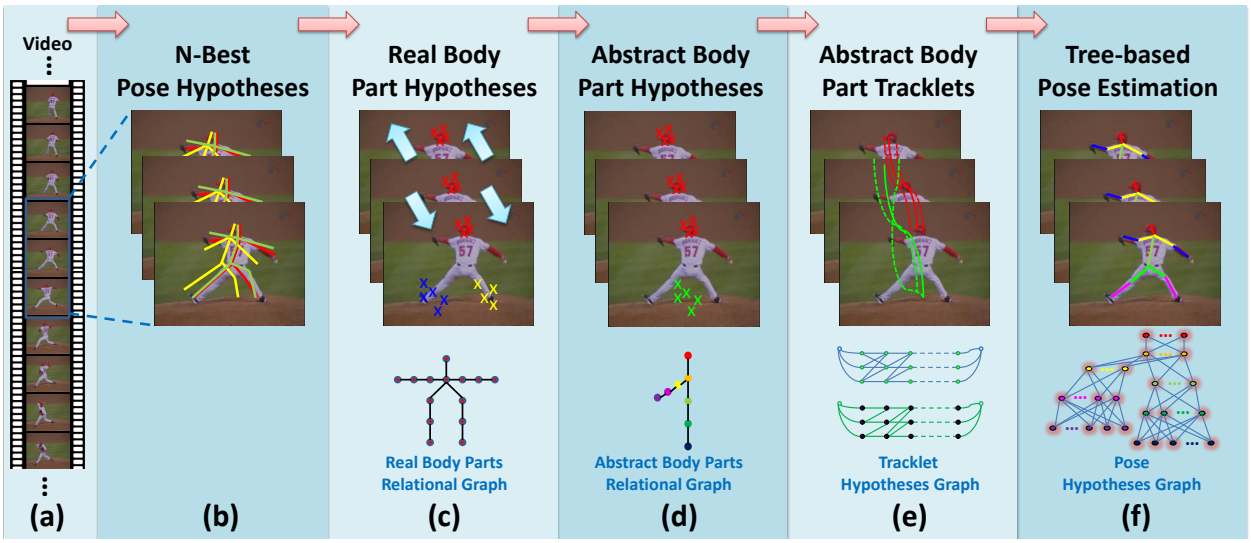

Figure 3. An outline of the proposed method. (a) shows the original video frames; in (b) the N-Best method ([19]) was employed to generate a set of diverse poses for each single frame; in (c), by using the results from (b), real body part hypotheses are generated for each body part in each frame and propagated to the adjacent frames; in (d), real body parts are combined into abstract body parts and the hypotheses are also combined accordingly in order to remove the intra-frame simple cycles (i.e. **the simple cycles with blue and green edges** in Fig.1(b)); in (e), tracklets are generated for abstract body parts (including single body parts and coupled body parts) using the abstract body part hypotheses generated in (d); in (f), the pose hypotheses graph is built, each node is a tracklet corresponding to the abstract body part, and the best pose estimation is obtained by selecting the best hypotheses for the parts from the graph.

$\{(v^k_{h(i)}, v^l_{h(j)}) | v^k_{h(i)} \in V_{h(i)}, v^l_{h(j)} \in V_{h(j)}, (v^i_r, v^j_r) \in E_r\}$, are built between each pair of hypothesis nodes from different *group*s following the structure of G_r. An unary weight, Φ, can be assigned to each hypothesis node, which measures the likelihood of the corresponding entity to be in the state of this hypothesis node; and a binary weight, Ψ, can be assigned to each hypothesis edge, which measures the compatibility of the pair of hypothesis nodes connected by the edge. Examples of hypothesis graph are shown in Fig.4(b,c) and Fig.5(b). The methodology is to select one hypothesis node for each entity, in order to maximize the combined unary and binary weights. This is a graph optimization problem and the general form is NP-hard; however, *if the relational graph is a tree (including the degenerate case of a single branch), the problem is no longer NP-hard and efficient dynamic programming based polynomial time solutions exist.*

For a tree-based relational graph, G_r, and the corresponding hypothesis graph, G_h, the objective function for a set of arbitrary selected nodes $s = \{s^i|_{i=1}^{|V_r|}, s^i \in V_h\}$ is:

$$\mathcal{M}(s) = \sum_{s^i \in V_h} \Phi(s^i) + \lambda \cdot \sum_{(s^i, s^j) \in E_h} \Psi(s^i, s^j), \quad (1)$$

in which λ is the parameter for adjusting the binary and unary weights, and the goal is to maximize $\mathcal{M}(s)$: $s^* = \arg\max_s(\mathcal{M}(s))$. Let the algorithm proceed from the leaf nodes to the root, and let $\mathcal{F}(i,k)$ be the maximum achievable combined unary and binary weights of kth hypothesis for ith entity. $\mathcal{F}(\cdot,\cdot)$ satisfies the following recursive function:

$$\mathcal{F}(i,k) = \Phi(v^k_{h(i)}) + \sum_{v^j_r \in kids(v^i_r)} \max_l \Big(\lambda \cdot \Psi(v^k_{h(i)}, v^l_{h(j)}) + \mathcal{F}(j,l)\Big). \quad (2)$$

Based on this recursive function, the problem can be solved efficiently by dynamic programming ([39, 41]), with a computation complexity of $\mathcal{O}(|V_r| \cdot N)$, in which N is the max number of hypotheses for each node in V_r.

3.2. Real Body Parts vs. Abstract Body Parts

We use the term *real body parts* to represent body parts which are commonly used in the literature. And we use **abstract body parts**, which is a new concept introduced in this paper, to facilitate the formulation of the proposed method (as shown in Fig.2). In contrast to the real body part definitions, there are two types of the abstract body parts in this paper: **single part** and **coupled part**. **Single parts** include *HeadTop* and *HeadBottom*. **Coupled parts** include *Shoulder*, *Elbow*, *Hand*, *Hip*, *Knee* and *Ankle*. Note that, for coupled parts, we use one part to represent two symmetric real body parts, for instance *Ankle* is employed to represent

the abstract part which is actually the combination of the *left* and *right* ankles. The motivation of abstract body parts is to remove simple cycles in the body part relational graph, and at the same time maintaining the ability of modeling the symmetric body parts. For example, in Fig.1(b), in each frame, the green and blue edges are used to model the body part relationships, and at the same time there are many simple cycles in a given frame. After introducing the abstract body parts in Fig.1(c), the symmetric parts are combined, and as a result, none of the frames contain simple cycles. However, there are still simple cycles between frames, which will be handled by the abstract body part tracklets in Section 3.4 and 3.5.

3.3. Body Part Hypotheses in a Single Frame

N-Best human pose estimation approach ([19]) is applied to each video frame to generate N best full body pose hypotheses. N is usually a large number (normally $N > 300$). And for each real body part, the body part hypotheses are body part locations extracted from the N-best poses. The body part hypotheses are sampled by an iterative non-maximum suppression (NMS) scheme based on the detection score map. Detection score is a combination of max-marginal ([19]) and foreground score,

$$\Phi_s(p) = \alpha\Phi_M(p) + (1-\alpha)\Phi_F(p), \tag{3}$$

in which Φ_s is the detection score, Φ_M is the max-marginal derived from [19], Φ_F is the foreground score obtained by the background subtraction ([7]), and p is the location of the body part.

The abstract body part hypotheses for a single part are the same as its corresponding real body part hypotheses. And the abstract body part hypotheses for a coupled part are the permutation of its corresponding left and right body part hypotheses.

3.4. Single Part Tracklets

Based on the abstract body part hypotheses generated in Section 3.3, we want to obtain several best single part and coupled part tracklets through the video frames. The problem is how to select one hypothesis from each frame, ensuring that they have high detection scores and are consistent throughout the frames. Following the definitions in Section 3.1, the relational graph for this problem is shown in Fig.4(a), and the hypothesis graphs for single parts and coupled parts are shown in Fig.4 (b) and (c) respectively.

Based on the single part hypotheses, a single part tracklet hypothesis graph is built (Fig.4(b)) for each single part (*headTop* and *headBottom*). In this graph, each node represents a single part hypothesis and the detection score $\Phi_s(p)$ from Eqn.3 is used to assign the node an unary weight. Edges are added between every pair of nodes from the adjacent frames. Binary weights are assigned to

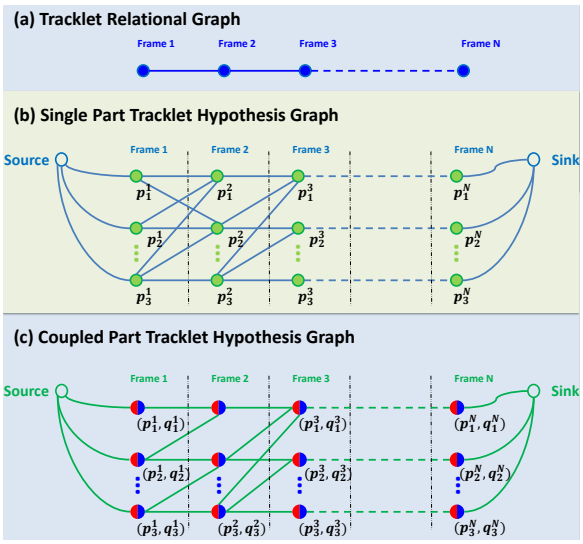

Figure 4. **Tracklet graphs**. (a) Shows the relational graph for the abstract body part tracklet generation. (b) Shows the tracklet hypothesis graph for single body parts. Each node represents one hypothesis location of the body part in a specific frame, and edges show the similarity between the connected body part hypotheses in adjacent frames. (c) Shows the tracklet hypothesis graph for coupled parts. Each node represents a coupled body part hypothesis, which is the combination of the corresponding symmetric body parts (that is why each node is colored into two halves). The edges represent the similarities between connected coupled body parts in adjacent frames. Note that, (b) and (c) are only illustrations, and for simplicity, not all edges are shown.

the edges which represent similarities between hypotheses in adjacent frames. The binary weight is defined as a combination of optical flow predicted location distance and the Chi-square distance of HOG features as follows:

$$\Psi_s(p^f, p^{f+1}) = \exp\left(-\frac{\chi^2(\Upsilon(p^f), \Upsilon(p^{f+1}) \cdot \|\hat{p}^f - p^{f+1}\|_2^2)}{\sigma^2}\right), \tag{4}$$

where p^f and p^{f+1} are two arbitrary hypotheses from frames f and $f+1$, $\Upsilon(p)$ is the HOG feature vector centered at location p, $\hat{p}^f$ is the optical flow predicted location for p^f in frame $f+1$, and σ is a parameter. The goal is to select one node from each frame to maximize the overall combined unary and binary weights. Given an arbitrary selection of nodes from the graph $s_s = \{s_s^i|_{i=1}^F\}$ (F is the number of frames), the objective function is given by

$$\mathcal{M}_s(s_s) = \sum_{i=1}^{F} \Phi_s(s_s^i) + \lambda_s \cdot \sum_{i=1}^{F-1} \Psi_s(s_s^i, s_s^{i+1}), \tag{5}$$

where λ_s is the parameter for adjusting the binary and unary weights, and $s_s^* = \arg\max_{s_s}(\mathcal{M}(s_s))$ gives the optimal solution. It is clear that the relational graph of this problem

is a **degenerate tree** (i.e. single branch tree, please see Fig.4(a)), and as shown in Section 3.1, the problem can be solved using dynamic programming efficiently. After the optimal solution is obtained, the selected nodes are removed from the graph and the next optimal solution can be obtained. This process can be iterated for multiple times in order to get several tracklets from the graph.

3.5. Coupled Part Tracklets

The relational graph for the coupled part tracklets generation is the same as for the single part; however, the nodes and edges are defined differently. In this case, each hypothesis node is composed of the locations of a pair of symmetric parts (e.g. left and right ankles). Fig.4(c) shows an illustration of the graph. Such design aims to model the symmetric relationship between coupled parts, including mutual location exclusions and appearance similarity in order to reduce double counting. As discovered in previous research ([22]), double counting is a key issue which severely hinders the pose estimation. Theoretically, tree based model ([39]) lacks the ability to model spatial relationship of the coupled parts (e.g. left and right ankles). Furthermore, as discussed in Section 1, attempting to model such spatial relationship would inevitably induce simple cycles in the graph, which would severely increase the computational complexity. By introducing the coupled parts, this could be effectively dealt with. In the coupled part tracklet hypothesis graph, each node $r = (p, q)$ represents a composition of a pair of symmetric parts p and q. Unary weights are assigned to the nodes which represent the detection confidence and the compatibility between the two symmetric parts, and the weight is defined as:

$$\Phi_c(r) = \frac{(\Phi_s(r.p) + \Phi_s(r.q)) \cdot (\Lambda(r.p)^T \cdot \Lambda(r.q)))}{1 + e^{-|r.p - r.q|/\theta}}, \quad (6)$$

where Φ_s is from Eqn.3, $r.p$ and $r.q$ respectively represent the left and right components of the coupled part r, $\Lambda(p)$ is the normalized color histogram of a local patch around p, the denominator is the inverse of a sigmoid function which penalizes the overlap of the symmetric parts, and θ is the parameter that controls the penalty. The binary weights of the edges are computed as

$$\Psi_c(r^f, r^{f+1}) = \Psi_s(r.p^f, r.p^{f+1}) + \Psi_s(r.q^f, r.q^{f+1}), \quad (7)$$

where Ψ_s is from Eqn. 4.

Similarly, the goal is to select one node (which is a composition of a pair of symmetric parts) from each frame to maximize the overall combined unary and binary weights. Given an arbitrary selection of nodes from the graph $s_c = \{s_c^i|_{i=1}^F\}$ (where F is the number of frames), the objective function is

$$\mathcal{M}_c(s_c) = \sum_{i=1}^{F} \Phi_c(s_c^i) + \lambda_c \cdot \sum_{i=1}^{F-1} \Psi_c(s_c^i, s_c^{i+1}), \quad (8)$$

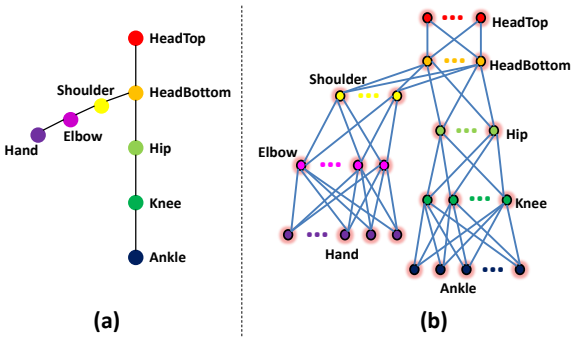

Figure 5. **Pose graphs**. (a) is the pose relational graph. Each node represents one abstract body part and edges represent the relationship between the connected body parts. (b) is the pose hypothesis graph. Each node is a tracklet for the part, and edges represent the spatial compatibility of connected nodes.

where λ_c is the parameter to adjust the binary and unary weights, and $s_c^* = \arg\max_{s_c}(\mathcal{M}(s_c))$ gives the optimal solution. As discussed in Section 3.1, the problem can also be solved by dynamic programming efficiently, and iterated for multiple times to get several tracklets.

3.6. Optimal Pose Estimation using Part Tracklets

Since the best tracklets for each abstract body parts are obtained by the methods introduced in Section 3.4 and 3.5, the next step is to select the best ones which are compatible. The relational graph, $G_T = (V_T, E_T)$, for this final tracklet based optimal pose estimation is shown in Fig.5(a). Each node represents an abstract body part, and the edges model the spatial relationships between them. Following the definitions of the abstract body parts and using the part tracklets generated for these abstract body parts, a pose hypothesis graph can be built to get the optimal pose (as shown in Fig.5(b)). In this graph, each node represents an abstract body part tracklet and edges represent the spatial constraints. For each hypothesis tracklet node, s, depending on if it corresponds to a single part or a coupled part, $E_s(s)$ from Eqn.5, or $E_c(s)$ from Eqn.8 is used as its unary weight $\Phi_T(s)$. Let $\Psi_d(p_i, q_j) = \omega_{i,j} \cdot \psi(p_i - q_j)$ be the relative location score in [39] ($\omega_{i,j}$ and ψ are defined the same as in [39]), the binary weight between a pair of adjacent single part tracklet nodes $s_s = \{s_s^i|_{i=1}^F\}$ and $t_s = \{t_s^i|_{i=1}^F\}$ is

$$\Psi_T(s_s, t_s) = \sum_{i=1}^{F} \Psi_d(s_s^i, t_s^i), \quad (9)$$

the binary weight between a single part tracklet node $s_s = \{s_s^i|_{i=1}^F\}$ and an adjacent coupled part tracklet node $t_c = \{t_c^i|_{i=1}^F\}$ is

$$\Psi_T(s_s, t_c) = \sum_{i=1}^{F} (\Psi_d(s_s^i, t_c^i.p) + \Psi_d(s_s^i, t_c^i.q)), \quad (10)$$

and the binary weight between a pair of adjacent coupled tracklet part nodes $s_c = \{s_c^i|_{i=1}^F\}$ and $t_c = \{t_c^i|_{i=1}^F\}$ is

$$\Psi_T(s_c, t_c) = \sum_{i=1}^{F} (\Psi_d(s_c^i.p, t_c^i.p) + \Psi_d(s_c^i.q, t_c^i.q)). \quad (11)$$

Now, the goal is to select only one tracklet for each abstract body part in order to maximize the combined unary (detection score) and binary (compatible score) weights. Given an arbitrary tree selected from the hypothesis graph $s_T = \{s_T^i|_{i=1}^{|V_T|}\}$, the objective function is given by

$$\mathcal{M}_T(s_T) = \sum_{v_T^i \in V_T} \Phi_T(s_T^i) + \lambda_T \cdot \sum_{(v_T^i, v_T^j) \in E_T} \Psi_T(s_T^i, s_T^j), \quad (12)$$

where λ_T is a parameter for adjusting the binary and unary weights, and as discussed in Section 3.1, the optimal solution $s_T^* = \arg\max_{s_T}(\mathcal{M}(s_T))$ can also be obtained by the dynamic programming algorithm efficiently. The body part locations in each frame are extracted from this final optimal solution.

4. Experiments

4.1. Datasets

We evaluated our method on three publicly available datasets:

Outdoor Pose Dataset: this dataset was collected by the authors of [22], which contains six video sequences from outdoor scenes. There are a lot of self-occlusions of the body parts and annotations of more than 1,000 frames are provided by the authors.

Human Eva-I: this dataset ([29]) contains human activities in indoor controlled conditions. The activities are synchronized with a ground truth of 3D motion capture data, which can be converted into 2D joint locations. In order to have a fair comparison with [22], we use 250 frames from the sequences: *S1_Walking*, *S1_Jog*, *S2_Jog* captured by camera 1.

N-Best Dataset, this dataset was collected by the authors of [19], which has four sequences in total. As a fair comparison to [22], we also report results on sequences *walkstraight* and *baseball*.

4.2. Evaluation Metrics

Similar to [22], we use PCP and KLE to evaluate our results. Probability of a Correct Pose (PCP) [8] is a standard evaluation metric which measures the percentage of correctly localized body parts within a threshold. Keypoint Localization Error (KLE) [22] measures the average Euclidean distance from the ground truth to the estimated keypoints, normalized by the size of the head in each frame.

Outdoor Dataset [22]								
Metric	**Method**	**Head**	**Torso**	**U.L**	**L.L**	**U.A.**	**L.A.**	**Average**
PCP	[22]	0.99	0.86	0.95	0.96	0.86	0.52	0.86
	[19]	0.99	0.83	0.92	0.86	0.79	0.52	0.82
	[5]	0.87	0.97	0.68	0.89	0.78	0.52	0.79
	Ours(Baseline)	0.92	1.00	0.84	0.73	0.68	0.47	0.77
	Ours(Abt. Only)	0.99	1.00	0.89	0.77	0.72	0.53	0.82
	Ours(Asc. Only)	0.99	1.00	0.87	0.76	0.79	0.56	0.83
	Ours	0.99	1.00	**1.00**	**0.97**	**0.91**	**0.66**	**0.92**
KLE	[22]	0.39	0.58	0.48	0.48	0.88	1.42	0.71
	[19]	0.44	0.58	0.55	0.69	1.03	1.65	0.82
	[5]	0.31	0.72	0.91	**0.36**	0.44	0.72	0.58
	Ours(Baseline)	0.58	0.45	0.61	0.78	0.75	1.11	0.71
	Ours(Abt. Only)	0.16	0.23	0.48	0.69	0.55	0.78	0.48
	Ours(Asc. Only)	0.16	**0.20**	0.47	0.64	0.44	0.71	0.44
	Ours	0.19	0.22	**0.35**	0.37	**0.41**	**0.61**	**0.36**
Human Eva-I Dataset [29]								
Metric	**Method**	**Head**	**Torso**	**U.L**	**L.L**	**U.A.**	**L.A.**	**Average**
PCP	[22]	0.99	1.00	0.99	**0.98**	**0.99**	0.53	0.91
	[19]	0.97	0.97	0.97	0.90	0.83	0.48	0.85
	[5]	0.99	1.00	0.90	0.89	0.96	0.62	0.89
	Ours(Baseline)	1.00	1.00	0.93	0.62	0.44	0.24	0.71
	Ours(Abt. Only)	1.00	1.00	0.98	0.66	0.43	0.30	0.73
	Ours(Asc. Only)	1.00	1.00	0.94	0.62	0.45	0.27	0.71
	Ours	1.00	1.00	**1.00**	0.94	0.93	**0.67**	**0.92**
KLE	[22]	0.27	0.48	0.13	0.22	1.14	1.07	0.55
	[19]	0.23	0.52	0.24	0.35	1.10	1.18	0.60
	[5]	**0.13**	**0.40**	0.23	0.16	**0.14**	0.24	0.22
	Ours(Baseline)	0.17	0.40	0.34	0.45	0.66	0.84	0.48
	Ours(Abt. Only)	0.17	0.41	0.29	0.41	0.66	0.75	0.45
	Ours(Asc. Only)	0.17	0.39	0.33	0.42	0.63	0.74	0.45
	Ours	0.16	0.42	0.13	**0.15**	0.20	0.24	0.22
N-Best Dataset [19]								
Metric	**Method**	**Head**	**Torso**	**U.L**	**L.L**	**U.A.**	**L.A.**	**Average**
PCP	[22]	1.00	0.69	0.91	0.89	0.85	0.42	0.80
	[19]	1.00	0.61	0.86	0.84	0.66	0.41	0.73
	[5]	1.00	1.00	0.91	0.90	0.69	0.39	0.82
	Ours(Baseline)	1.00	1.00	0.92	0.87	0.87	0.52	0.86
	Ours(Abt. Only)	1.00	1.00	0.91	0.89	0.87	0.65	0.89
	Ours(Asc. Only)	1.00	1.00	0.93	0.91	0.87	0.55	0.88
	Ours	1.00	1.00	0.92	**0.94**	**0.93**	**0.65**	**0.91**
KLE	[22]	0.53	0.88	0.67	1.01	1.70	2.68	1.25
	[19]	0.54	0.74	0.80	1.39	2.39	4.08	1.66
	[5]	0.15	0.23	0.31	0.37	0.46	1.18	0.45
	Ours(Baseline)	0.15	0.19	0.36	0.49	0.32	0.84	0.39
	Ours(Abt. Only)	0.15	0.19	0.31	0.43	0.34	0.60	0.34
	Ours(Asc. Only)	0.15	0.17	0.27	0.42	**0.29**	0.68	0.33
	Ours	0.15	0.17	**0.24**	0.37	0.30	**0.60**	**0.31**

Table 1. Comparisons with the state-of-the-art methods on three publicly available datasets. Note that PCP is an accuracy measure, so the larger the better, with a max of 1; and KLE is an error measure, so the smaller the better.

4.3. Results

We compare the proposed method with three state-of-the-art video based human pose estimation methods: N-Best method ([19]), Symmetric Tracking method ([22]), and Mixing Body-part method ([5]); we did not compare with some upper body pose estimation/tracking methods ([27, 34]), since they focus on the modeling of hands/elbows by motion and appearance features but do not handle other body parts. Since [5] was designed for upper-body pose estimation, we re-implemented its algorithm by reusing most of their implementation and extended it to a full-body detection model. Quantitative results are shown in Table 1, and qualitative results are shown in Fig.6. Note that the figures for Symmetric Tracking method are reproduced from figures in [22], since the code is not publicly available.

We also show detailed results to analyze the

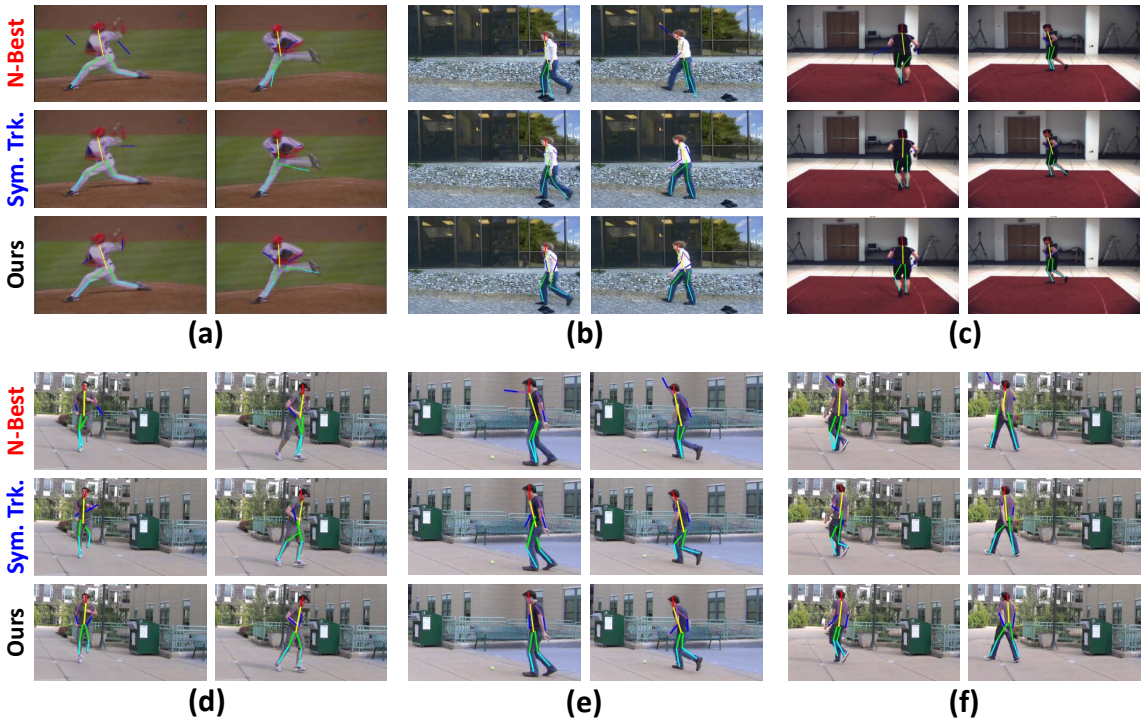

Figure 6. Examples and comparisons with the state-of-the-art methods: N-Best [19], Symmetric Tracking (Sym. Trk.) [22] and ours. (a) and (b) are from N-Best Dataset; (c) is from Human Eva I dataset; and (d)-(f) are from Ourdoor Pose Dataset. Body parts are shown in different colors. Please see more results from `http://crcv.ucf.edu/projects/human_pose_estimation/`.

contributions of each step of the proposed method. In Table 1,'Ours(Baseline)' shows the results for the proposed method without 'Abstraction' and 'Association'; 'Ours(Abt. Only)' shows the results for only applying the 'Abstraction' step of our method; and 'Ours(Asc. Only)' shows the results for only using the 'Association' step of the proposed method. From these results we found that 'Abstraction' is more important than 'Association' in the proposed method, due to the fact that it contributes more to the quantitative improvement.

Limitations: The proposed method relies on N-Best method ([19]); therefore, if N-Best method can not generate any correct candidates for a specific body part, it is not possible to obtain improved results by the proposed method.

4.4. Implementation Details

We process 15 consecutive frames each time. For Eqn.5 and 8, we normalized the unary and binary weights in each frame between 0 and 1. We use $\alpha = 0.5$ in Eqn.3, and $\lambda_c = \lambda_s = \lambda_T = 1$ for Eqn.5,8 and 12. For σ in Eqn.4 and θ in Eqn.6, we use 10% of the median height (normally 15-30 pixels) of N-Best poses ([19]) obtained from the step in Section 3.3. For each real body part (Section 3.3), we generate 20 hypotheses, and for each abstract body part we select the top 10 tracklets (Section 3.4 and 3.5).

4.5. Computation Time

We performed experiments on a desktop computer with Intel Core i7-3960X CPU at 3.3GHz and 16GB RAM. On average, to process one frame (typical frame size: 600×800, we resize the larger frames), the Matlab implementation took $0.5s$ to generate the body part hypotheses and weights, $0.5s$ to build the graph and compute the tracklets, and it took $0.1s$ to build the pose hypothesis graph (Section 3.6) and get the optimal solution.

5. Conclusions and future work

We have proposed a tree-based optimization method for human pose estimation in videos. Our contribution is mostly focused on reformulating the problem to remove the simple cycles from the graph, and at the same time maintain the useful connections at the greatest possible extent, in order to transform the original NP-hard problem into a simpler tree based optimization problem, for which the exact solution exists and can be solved efficiently. The proposed formulation is general and it has a potential to be employed in solving some other problems in computer vision.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People tracking-by-detection and people detection-by-tracking. In *CVPR*, 2008.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[4] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, pages 1736–1744, 2014.

[5] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *CVPR*, 2014.

[6] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013.

[7] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *ECCV*. 2000.

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

[9] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *CVPR*, 2013.

[10] H. Ishikawa. Exact optimization for markov random fields with convex priors. *T-PAMI*, 25(10):1333–1336, 2003.

[11] H. Jiang. Human pose estimation using consistent max covering. *T-PAMI*, 33(9):1911–1918, 2011.

[12] H. Jiang and D. R. Martin. Global pose estimation using non-tree models. In *CVPR*, 2008.

[13] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[14] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[15] L. Karlinsky and S. Ullman. Using linking features in learning non-parametric part models. In *ECCV*. 2012.

[16] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005.

[17] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004.

[18] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, pages 2337–2344, 2014.

[19] D. Park and D. Ramanan. N-best maximal decoders for part models. In *ICCV*, 2011.

[20] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013.

[21] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*. 2012.

[22] V. Ramakrishna, T. Kanade, and Y. Sheikh. Tracking human pose by tracking symmetric parts. In *CVPR*, 2013.

[23] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*. 2014.

[24] D. Ramanan. Learning to parse images of articulated bodies. In *Advances in neural information processing systems*, pages 1129–1136, 2006.

[25] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, 2005.

[26] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[27] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[28] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56:116–124, 2013.

[29] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2010.

[30] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006.

[31] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *CVPR*, 2012.

[32] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR*, 2012.

[33] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*. 2012.

[34] R. Tokola, W. Choi, and S. Savarese. Breaking the chain: liberation from the temporal markov assumption for tracking human poses. In *ICCV*, 2013.

[35] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014.

[36] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[37] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 2013.

[38] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*. 2008.

[39] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[40] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained monocular 3d human pose estimation by action detection and cross-modality regression forest. In *CVPR*, 2013.

[41] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, pages 628–635, 2013.

[42] S. Zuffi, J. Romero, C. Schmid, and M. J. Black. Estimating human pose with flowing puppets. In *ICCV*, 2013.