Mixing Body-Part Sequences for Human Pose Estimation

Anoop Cherian, Julien Mairal, Karteek Alahari, Cordelia Schmid

Mixing Body-Part Sequences for Human Pose Estimation

Anoop Cherian* Julien Mairal* Karteek Alahari* Cordelia Schmid*

Inria

Abstract

In this paper, we present a method for estimating articulated human poses in videos. We cast this as an optimization problem defined on body parts with spatio-temporal links between them. The resulting formulation is unfortunately intractable and previous approaches only provide approximate solutions. Although such methods perform well on certain body parts, e.g., head, their performance on lower arms, i.e., elbows and wrists, remains poor. We present a new approximate scheme with two steps dedicated to pose estimation. First, our approach takes into account temporal links with subsequent frames for the less-certain parts, namely elbows and wrists. Second, our method decomposes poses into limbs, generates limb sequences across time, and recomposes poses by mixing these body part sequences.

We introduce a new dataset "Poses in the Wild", which is more challenging than the existing ones, with sequences containing background clutter, occlusions, and severe camera motion. We experimentally compare our method with recent approaches on this new dataset as well as on two other benchmark datasets, and show significant improvement.

1. Introduction

Articulated human pose estimation plays a key role in many computer vision applications, including activity recognition and video understanding [30, 34]. Several factors make this task challenging, such as the diversity of appearances, changes in scene illumination and camera viewpoint, background clutter, and occlusion. In recent years, a significant effort has been devoted to estimating human poses in single images [3, 7, 24, 33]. Although these methods perform well on certain body parts, e.g., head, their performance on localizing parts corresponding to lower arms, i.e., elbows and wrists, is poor in general. The focus of this paper is to improve human pose estimation, and in particular to localize lower-arm parts accurately by modeling interactions between body parts across time.

Recent algorithms assume that articulated human poses are composed of a set of rigid body parts [3, 4, 7, 8, 11, 33], for which body-part templates are learned from training data. A probabilistic graphical model, often a Markov random field (MRF), is designed with scores provided by these templates. For single images, the MRF is usually modeled as a tree or a star-shaped graph, leading to tractable and efficient inference, as successfully done in [4, 7, 8, 33]. One way to extend such methods for estimating poses in videos is by introducing regularization on the body parts across time, e.g., by adding temporal part-part edges [10, 24, 27, 29, 31]. The resulting graphical model is no longer a tree, and inference becomes intractable. Thus, approximations are required, which can be done by changing the graph structure, e.g., ensemble of tree-structured MRFs [24, 31], or by using approximate inference methods, such as loopy belief propagation or sampling [10, 27, 29].

(a) (b) (c)

Figure 1. Human pose estimated by (a) Yang & Ramanan's method [33], our approach: (b) local temporal model for less-certain parts, and (c) mixing body-part sequences.

In this paper, we introduce a new approximation scheme adapted to the human pose estimation problem. We begin by generating a set of pose candidates in each frame with a model including temporal links with subsequent frames for the less-certain parts, namely elbows and wrists, see Figure 1(b). Since the loops in the corresponding MRF are isolated, we show that inference can be performed efficiently with the use of distance transforms [9]. We then compute the n-best poses [5, 18] in each frame to obtain a diverse set of candidate poses (Section 3.3). Next we introduce an effective method to smooth these poses temporally. We decompose the human pose into limbs and track them to generate body-part sequences. We then recompose the complete pose by mixing these part sequences (Figure 1(c), Section 3.4). This procedure explores a set of poses that is exponential in K, the size of the candidate set, in polynomial time ($\mathcal{O}(NTK^2)$, where N is the number of body parts and T is the number of frames).

*LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France.

We experimentally demonstrate the effectiveness of our algorithm on two state-of-the-art datasets: VideoPose [24] and MPII Cooking Activities [21]. While these are interesting preliminary benchmarks, they do not have the same level of difficulty that occurs in recent datasets for pose estimation in single images (i.e., datasets where pose labels are only provided for individual isolated frames), such as FashionPose [6]. For instance, VideoPose sequences have very limited occlusions, are all shot indoors, pre-processed to stabilize motion and to align the head location. The videos in the Cooking Activities dataset are recorded in a single indoor environment with a static camera. We introduce a more challenging dataset called *Poses in the Wild*, which is available on-line [1]. It contains 30 sequences from three Hollywood movies. It is representative of real-world scenarios, with background clutter, body-part occlusions, and severe camera motion. Our method improves over prior art on these three datasets by approximately 12% in localizing wrists (Section 4).

2. Related Work

One of the main challenges for pose estimation in videos is to handle the temporal coupling of parts across frames. This results in models that are highly inter-connected (i.e., loopy graphs with high tree-width) and are thus intractable to perform inference on. Previous works have resorted to approximate inference to address this issue. For instance, the methods in [27, 29] use a sampling approach. More recent methods [10, 14] have used loopy belief propagation instead. Such approaches typically come with high computational costs for marginal improvements in performance in general. Sapp *et al*. [24] propose a strategy where the model is represented as a convex combination of tree-structured graphs linked with dual variables, and solve it with a dual decomposition algorithm. It shows better performance over other approximations, but remains computationally expensive on long sequences.

Some of the earlier approaches detect a pose in a few frames, and track it in the rest of the sequence [25, 26]. A few other methods adopt a tracking-by-detection scheme to estimate poses in videos [2, 5, 12, 16, 18, 20]. Specifically, they compute the pose in some (or in all the) frames, and track it over time. Canonical poses [20] or part-level segmentations [12, 16] have been used to extract the initial pose in a frame. A diverse set of poses [5, 18] instead of one candidate in each frame has also been used. These methods smooth the entire articulated pose over time using the candidate(s), which are typically no more than a few hundred in number. An alternative strategy is to track individual parts, and to explore a set of poses, which is exponential in the size of the candidate set [13, 15, 19]. Ramakrishna *et al*. [19] present a top-down approach in this context. They compute the optimal track for a part, and use it to condition the tracking problem on the neighboring part(s). Our approach also tracks parts individually to exploit a large pose candidate set, but imposes regularization along the limbs (connecting two body parts) instead of the parts alone, as shown in Figure 3.

More recently, Zuffi *et al*. [35] have proposed a scheme where poses across two consecutive frames are coupled using optical flow. Although this method showed promising results, it is limited to frame-to-frame refinements. In comparison, our approach additionally optimizes pose-part locations over entire sequences. Tokola *et al*. [28] explore the exponentially large search space for finding optimal parts with an ad hoc approach to find part tracks (i.e., locations of a part over the entire sequence), whereas we present a method to find optimal tracks. As shown in the experimental evaluation in Section 4, our proposed approach outperforms both these methods.

3. Proposed Approach

Our work relies on the deformable mixture-of-parts model proposed for single images in [33] due to its good performance and computational efficiency. In this section, we first briefly present this technique, and then introduce our approach for video sequences.

3.1. Pose Estimation in Single Images

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph with vertices $\mathcal{V}$ and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ representing the structure of a human pose. Each vertex corresponds to a body part (i.e., head, shoulders, elbows, wrists), and each edge represents a connection between two of these parts; see Figure 2(a). We define a pose p with respect to this graph $\mathcal{G}$ as a set of 2D coordinates representing the positions of the different body parts in an image as:

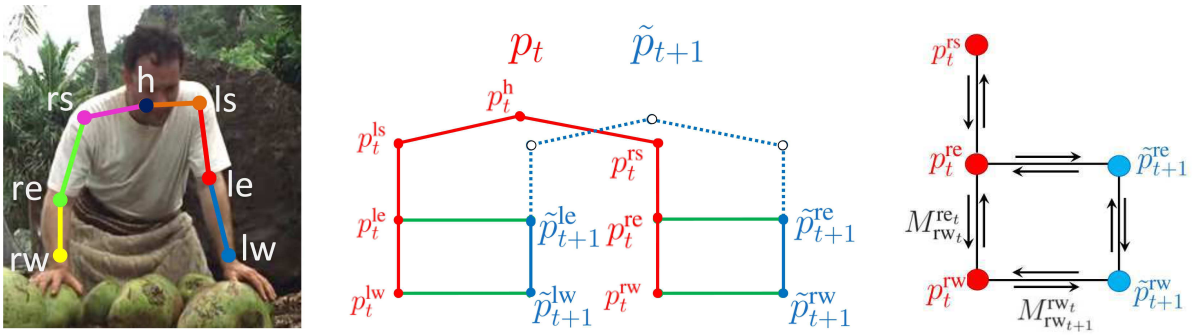
$$p = \left\{ p^u = (x^u, y^u) \in \mathbb{R}^2 : \forall u \in \mathcal{V} \right\}.$$

The formulation of [33] uses a mixture of body part models. Every part can be associated with one of M possible "types", and choosing a type configuration determines the pose model. Thus, estimating the pose involves not only choosing the part positions p, but also the type for each body part [33]. We use this framework in the paper, but omit the details in the following presentation to simplify the notation.

The single-image pose estimation problem is then formulated as the minimization of the following *cost* $C(I, p)$ for a pose p and an image I:

$$C(I, p) := \sum_{u \in \mathcal{V}} \phi_u(I, p^u) + \sum_{(u,v) \in \mathcal{E}} \psi_{u,v}(p^u - p^v), \quad (1)$$

where $\phi_u(I, p^u)$ is an appearance term for the body part u at the position p^u in I, and $\psi_{u,v}(p^u - p^v)$ is a deformation cost



(a) (b) (c)

Figure 2. (a) Our graphical model for human pose in a single image is shown with the body parts (**h**ead, **l**eft and **r**ight **s**houlders, **e**lbows and **w**rists) and their spatial connections. (b) The graphical model used to refine the pose estimate p_t in image I_t with a dummy pose $\tilde{p}_{t+1}$ (shown in solid blue lines), which contains only the wrist and the elbow parts in image I_{t+1}. The temporal links between these two poses are shown in green. (c) An illustration showing messages between some of the body parts.

for body parts (u, v), which is often compared to the energy model of a spring. Both ϕ_u and $\psi_{u,v}$ have underlying linear filters that are learned by using a structured SVM formulation. When $\mathcal{G}$ is a tree, the exact minimizer of (1) can be found in polynomial time with dynamic programming [33].

3.2. Pose Estimation in Videos

Given a video sequence $\mathcal{I} = (I_1, I_2, \ldots, I_T)$, it is common to introduce temporal links between every pair of frames I_t and I_{t+1} in the sequence, in order to impose temporal consistency in the estimation of the pose positions $p_1, p_2, \ldots, p_T$. This is achieved by adding a temporal edge between every pair of nodes p_t^u and p_{t+1}^u, leading to the following cost function:

$$C(I_T, p_T) + \sum_{t=1}^{T-1} C(I_t, p_t) + \lambda_1 \theta(p_t, p_{t+1}, I_t, I_{t+1}), \quad (2)$$

where θ is a consistency term between the poses in two consecutive frames and λ_1 is a regularization parameter. We measure the consistency between p_t and p_{t+1} by comparing p_{t+1} with p_t adjusted with optical flow as follows:

$$\theta(p_t, p_{t+1}, I_t, I_{t+1}) = \sum_{u \in \mathcal{V}} \|p_{t+1}^u - p_t^u - f_t(p_t^u)\|_2^2, \quad (3)$$

where $f_t(p_t^u)$ is the optical flow between frames I_t and I_{t+1} evaluated at the position p_t^u. Indeed, this approach is quite natural and similar formulations have been proposed [5, 18, 24]. Our work mainly differs from these approaches in the way we address the problem of minimizing (2), which is intractable and requires some approximations.

The temporal edges introduce loops in the graph, which leads to an intractable inference problem. It would be possible to use an approximate method like loopy belief propagation, whose complexity is exponential in the size of the maximal clique in the graph [17]. We have found such a strategy too slow to be practical for pose estimation. Instead, we propose a two-stage approach.

The first step consists of generating a set of candidate poses in each frame. We achieve this by minimizing an approximation of (2) in combination with the n-best algorithm [18]. Specifically, we build on the approach of [18] by introducing frame-frame temporal smoothness among some of the body parts. In the second step, we decompose the candidate poses into limbs, and generate limb sequences across time. We then recompose the complete, accurate pose my mixing these body-part sequences. This strategy shows a better performance than simply optimizing (2) over the candidate poses as in [18] because it explores a larger set of poses. We now detail these two steps.

3.3. Generating Candidate Poses

In this step, we focus on generating a set of K candidate poses in each frame I_t. One approach for this task is to use the cost $C(I_t, p_t)$ in (1) for estimating poses in the frame I_t, and compute the K *best* and diverse solutions, as proposed in [18]. In other words, we find diverse pose configurations that yield low cost in each frame independently, regardless of the temporal smoothness. We have observed that this strategy tends to be inaccurate for parts that are difficult to estimate, such as wrists, as shown in Section 4.

We propose a method to refine the estimation of a pose p_t in a single image frame I_t using a *dummy* pose $\tilde{p}_{t+1}$ that contains only the wrist and elbow parts in the frame I_{t+1}. We define this task as optimizing the following cost function:

$$C(I_t, p_t) + \tilde{C}(I_{t+1}, \tilde{p}_{t+1}) + \tilde{\lambda}_1 \sum_{u \in \mathcal{W}} \|\tilde{p}_{t+1}^u - p_t^u - f_t(p_t^u)\|_2^2, \quad (4)$$

where $\mathcal{W} \subset \mathcal{V}$ represents the left and right wrists and elbows, $\tilde{\lambda}_1$ is a regularization parameter. The cost $\tilde{C}$ is defined as C in (1), except that only terms corresponding to wrists and elbows are considered, i.e., it contains the appearance terms ϕ_u for these parts and the deformation costs $\psi_{u,v}$ between them.

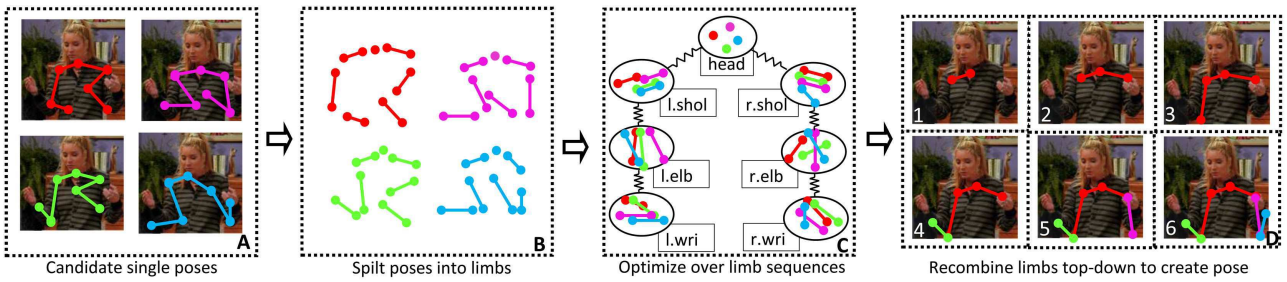


Figure 3. Illustration of our limb recombination scheme. From left to right: Block-A: An image and four candidate poses, where only a part of each pose is well-aligned with the person. Block-B: We divide each candidate pose into limb parts. Block-C: We allow the recombination of limbs from different pose candidates with constraints between two limbs that have a joint in common. Block-D: An example where recombination builds an accurate pose, which is not in the original candidate set. See text for more details.

In Figure 2(b) we show the graphical model corresponding to this step. It contains two isolated loops—a setting where exact inference can be performed with loopy belief propagation [32]. This algorithm proceeds as a sequence of message passing steps. In each step, a message M_v^u is passed from node u to node v in the graph. It is then used to update the message from v to other nodes it is connected to. This procedure is repeated for all the nodes until convergence, i.e., none of the messages change after an update iteration is performed. On our graph, we begin by sending messages from the leaf nodes to the root, and then from the root node to the rest. After convergence, we assign each node to the label corresponding to the minimum marginal at that node. This procedure can be implemented efficiently on our graph with the distance transform technique [9].

As shown in the experiments, our approach for generating candidate poses by minimizing (4) instead of simply minimizing $C(I_t, p_t)$ performs better with no significant increase in computational cost.

3.4. Recombining Limbs with Variable Splitting

After generating a set of K candidate poses for every frame (denoted by $\mathcal{P}_t$ for frame I_t), a simple strategy is to optimize our global objective function (2) over this set as:

$$\min_{p_t \in \mathcal{P}_t, \forall t} C(I_T, p_T) + \sum_{t=1}^{T-1} C(I_t, p_t) + \lambda_1 \theta(p_t, p_{t+1}, I_t, I_{t+1}). \tag{5}$$

This can be solved efficiently with dynamic programming in $\mathcal{O}(TK^2)$ operations, as done in [18] for example. However, we have observed that the constraint $p_t \in \mathcal{P}_t$ is an important limitation of such a strategy. On the one hand, it has the positive effect of making (5) tractable, but on the other hand, having only K different possible poses in every frame can be problematic. The $\mathcal{P}_t$ may contain a "good" candidate pose, but the method is unable to deal with situations where it is not the case, thus motivating us to propose an approximate scheme exploring a larger set of poses than $\mathcal{P}_t$.

Our main idea is to allow the recombination of limbs from candidate poses in $\mathcal{P}_t$ in order to create new poses, yielding a new set $\bar{\mathcal{P}}_t$ that is exponentially larger than $\mathcal{P}_t$. We will then minimize (5) approximately over $\bar{\mathcal{P}}_t$. Before going into details, we start by sketching our approach. As shown in Figure 3: (A) We break each pose p in $\mathcal{P}_t$ into limbs $l^{u,v} = (p^u, p^v)$, where (u, v) is in $\mathcal{E}$. (B) We decompose (5) into a sum of costs for every limb sequence. (C) We allow the recombination of limbs from different poses as follows: consider two limbs $l^{u,v} = (p^u, p^v)$ and $l^{v,w} = (p'^v, p'^w)$ obtained respectively from two poses p and p' in $\mathcal{P}_t$. The two limbs share the same body part v, and thus p^v should be close to p'^v, such that the two individual limbs can be considered as a good approximation of the combination (p^u, p^v, p'^w). This is achieved by adding a pairwise cost $\gamma(l^{u,v}, l^{v,w}) = \lambda_2 \|p^v - p'^v\|_2^2$ to our formulation, which can be interpreted as attaching a spring between the two limbs (Figure 3-C). (D) We finally estimate the pose by recombining limbs in a top-to-bottom fashion, approximately minimizing the resulting objective function.

Formally, the above approach consists of approximating the objective (5) by a sum of costs over the limbs:

$$\sum_{(u,v)\in\mathcal{E}} S^{u,v}(l^{u,v}_{1\ldots T}) + \lambda_2 \sum_{t=1}^{T} \gamma(l_t^{\mathrm{pa}(u),u}, l_t^{u,v}), \tag{6}$$

where $l^{u,v}_{1\ldots T}$ represents a limb sequence $(l_1^{u,v}, \ldots, l_T^{u,v})$, the function γ is the cost defined in the previous paragraph, and $\mathrm{pa}(u)$ represents the parent node of the body part u in the tree. Note that to simplify the notation, we associate the head to a limb (h, h) with $\mathrm{pa}(h) = h$, where h in $\mathcal{V}$ is the root of the tree. The score $S^{u,v}(l^{u,v}_{1\ldots T})$ for a limb (u, v) contains all the pairwise terms in (5) involving p_t^u and p_t^v, as well as all the terms involving p_t^v's only. To further simplify the computations and improve the speed of the procedure, we approximate the non-temporal terms involving p_t^u and $p_t^u - p_t^v$ by the cost $C(I_t, p_t)$ computed in Section 3.3.

We then proceed in a top-to-bottom fashion; we start by estimating the head sequence, which is usually the most reliable body part in pose estimation, by minimizing the corresponding function $S^{h,h}$ over the set of head candidates. This can be done in $\mathcal{O}(K^2T)$ operations with dynamic programming. In the next step, we estimate

the limbs connected to the head by minimizing the cost $S^{h,v}(l^{h,v}_{1...T}) + \lambda_2 \sum_{t=1}^{T} \gamma(l_t^{\text{pa}(h),h}, l_t^{h,v})$. Again, this is done in $\mathcal{O}(K^2T)$ operations. We proceed recursively until the wrists are estimated.

The procedure is approximate, but turns out to be effective in practice, as shown in Section 4. It improves upon (5) by exploring a larger set $\bar{\mathcal{P}}_t$ instead of $\mathcal{P}_t$.

3.5. Practical Extensions

We now present a few simple variations of our model that have shown to improve the quality of our pose estimation.

Temporal Regularization along Dense Limb Positions. The joint positions are typically sufficient to entirely characterize a pose p, but temporal regularization can be added to different points along the limbs to make the estimation more robust. We use this strategy, and define a set of three equidistant positions $p_t^{u'}$ along the limbs in our implementation. We add the corresponding regularization terms $\theta(p_t^{u'}, p_{t+1}^{u'}, I_t, I_{t+1})$ for these additional keypoints to the cost function (5), sharing the same regularization parameter. We compute the maximum optical flow within a 12×12 patch around each of these positions as well as the keypoints to obtain a robust estimate of the flow.

Enriched Wrist Model. Wrists being the most difficult part to estimate, we have found it useful to enrich their model by using a motion a priori. In many cases the wrist is moving, i.e., it is rare that a person's arms are not in motion. We can use this fact to encourage the alignment of a part position to a region with high motion, leading to an additional regularization term, $\lambda_3 \frac{1}{F_t} \sum_{u'} |f_t(p_t^{u'})|$, in the objective (5). Here, $f_t(p_t^{u'})$ is the flow of the wrist and the additional keypoints between the wrist and the elbow (see above), and F_t is the maximum absolute flow between (I_t, I_{t+1}) used for normalization.

Limiting Spatial Dynamics. In a few situations, we have observed the optical flow to be unreliable. For example, when the flow is associated with a background object. To prevent any large motion due to an inaccurate flow vector, we have found it reasonable to encourage the positions p_t^u and p_{t+1}^u to be close, and to add the regularization term $\lambda_4 \left\| p_t^u - p_{t+1}^u \right\|_2^2$ to our objective function (5).

4. Experiments

In this section, we first describe the three datasets used, followed by the implementation details, and then present our comparison with the state of the art.

4.1. Datasets

VideoPose. This dataset was introduced in [23] for evaluating upper-body pose estimation in videos. It consists of 26 sequences ($\sim$ 750 frames) for training and 18 sequences ($\sim$ 500 frames) for testing. All the frames are annotated with the following body parts: torso, shoulders, elbows, wrists. We follow the evaluation scheme of [12], i.e., test on the 17 sequences from the Friends TV series and compare the results for localizing elbows and wrists.

MPII Cooking Activities. This dataset was proposed in [21] for recognizing cooking activities in video sequences. It contains approximately 1500 frames from 16 video clips, where each frame is annotated with upper-body parts. The frames are captured with a static camera, and all the sequences are recorded in the same kitchen.

Poses in the Wild. We introduce a challenging dataset named *Poses in the Wild*. It contains 30 sequences, with about 30 frames each, extracted from the Hollywood movies "Forrest Gump", "The Terminal", and "Cast Away". We manually annotated all the frames with upper-body poses. In contrast to the VideoPose and Cooking Activities datasets, it contains realistic poses in outdoor scenes, with background clutter, severe camera motion and body-part occlusions. The dataset is publicly available at [1].

4.2. Implementation Details

Training. The main components of the pose model are the appearance term for each body part and the deformation cost between a pair of parts (ϕ_u and $\psi_{u,v}$ in (1) respectively). We learn the underlying filters for these terms using the method of [33] for estimating poses in single images, as described in Section 3.1. Following [33], we augment our set of parts (keypoints) with keypoints corresponding to: (i) the midpoints of the lower and upper arms, (ii) the center of the torso, and (iii) the midpoint between the head and the center of the torso. Thus, we have a 13-part body model, where each part is associated with one of the eight HOG templates in the mixture model [33].

For our experiments on VideoPose dataset we train our model on the VideoPose training set used in [12, 24]. For our experiments on the Cooking Activities and the Poses in the Wild datasets, the model is trained with all the images annotated with upper-body parts (about 4.5K) in the FLIC dataset [22]. This dataset contains a larger diversity of poses than VideoPose.

Evaluation Metric. We use the keypoint localization error [24] to measure the accuracy of different methods. Given the best pose estimation per frame, it measures the percentage of keypoints localized within a given distance from the ground truth per keypoint type. We show results with distances in the 15-40 pixel range.

Hyperparameters. The hyperparameters are fixed to the same values for all the experiments and datasets. We set $\tilde{\lambda}_1 = 10^{-4}$, which yields the same order of magnitude for the optical flow and the cost terms in (4). The hyperparameter λ_1 in (2) is set to 1, giving the same importance to

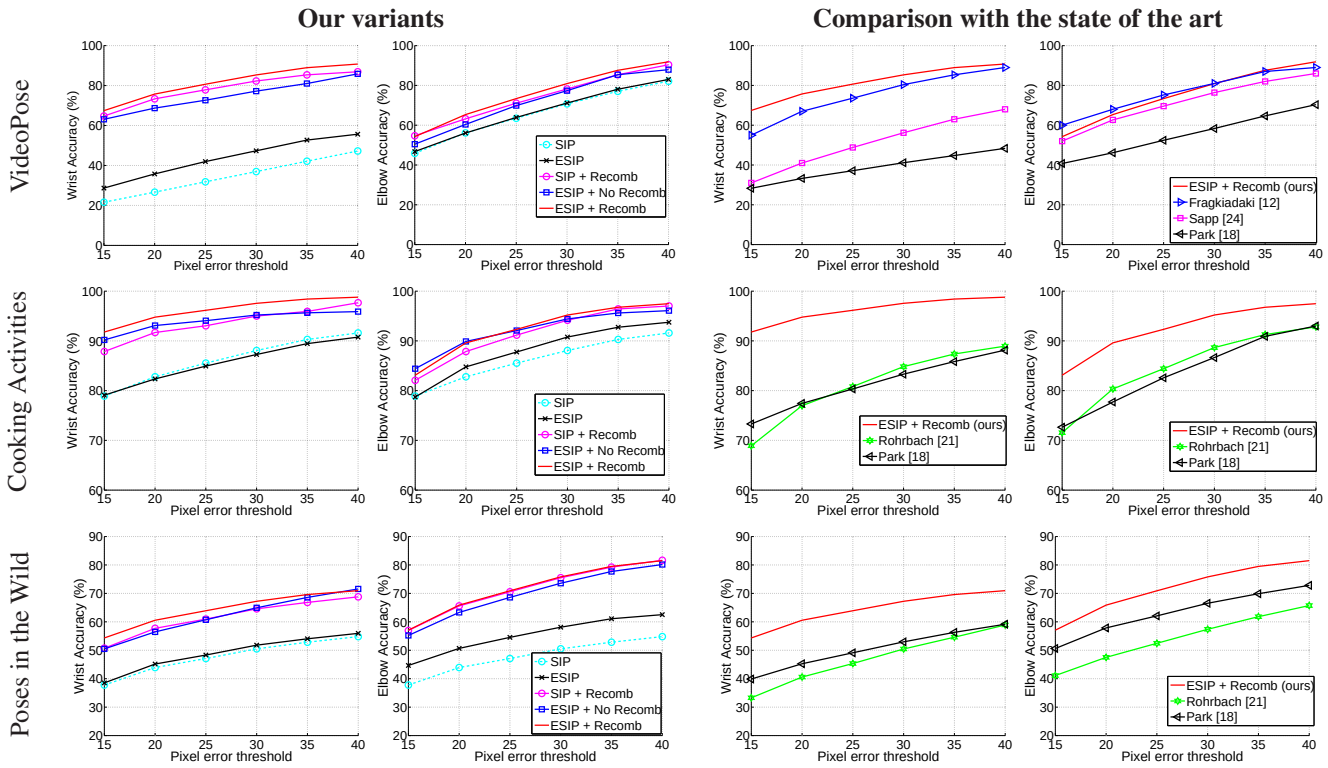


Figure 4. The first two columns compare the different steps of our approach. SIP, ESIP: We estimate the pose in each frame independently. ESIP is with the refinement using a dummy pose (§3.3, Figure 2(b)) and SIP is without (§3.1). SIP+Recomb, ESIP+No Recomb, ESIP + Recomb: We estimate poses using temporal information. Recomb refers to our method for recombining limbs to find accurate poses (§3.4). In the last two columns we compare our best approach (ESIP + Recomb) with the state of the art.

the image-based pose cost and the flow consistency between poses in consecutive frames. The scalar λ_2 in (6) is set to 5. The enriched wrist model is weighted by $\lambda_3 = 2$, and the term λ_4 for limiting spatial dynamics is set to 0.1.

4.3. Evaluating the Model Components

In Figure 4, we evaluate the performance of various components of our method in the first two columns. We compare the best pose produced by the single image pose (SIP) estimation algorithm of [33] against our extended single image pose (ESIP) model, which refines the pose in a frame using the elbow and wrist locations in the next frame (Section 3.3). We then evaluate variants of our temporal pose recombination method (Recomb; Section 3.4). The effect of using it with SIP (SIP + Recomb) and ESIP (ESIP + Recomb) are shown. In addition to this, we evaluate ESIP without using the recombination method (ESIP + No Recomb), i.e., using the entire poses in the candidate set with temporal smoothing.

The significance of our recombination scheme can be analyzed in the context of two pose models: SIP and the proposed ESIP. For example, on the Poses in the Wild dataset (Figure 4, row 3), recombination with SIP improves wrist localization by 12% and with ESIP, which already shows 10% improvement on the SIP model, by a further 4% (15 pixel error). In other words, using one of the two improvements (SIP + Recomb, ESIP + NoRecomb) shows a similar performance (12% gain over baseline SIP), and combining them together (ESIP + Recomb) gives an additional 4% gain.

Dataset	Shoulders	Elbows	Wrists
VideoPose	84.0	54.2	67.4
Cooking Activities	91.5	83.1	90.7
Poses in the Wild	62.7	57.0	54.3

Table 1. Body-part localization accuracies for our full model (ESIP + Recomb) with an error threshold of 15 pixels.

4.4. Comparison with the State of the Art

In the last two columns in Figure 4, we compare our full model ESIP + Recomb to four recent pose estimation algorithms for video sequences: Fragkiadaki *et al*. [12], Sapp *et al*. [24], Park *et al*. [18], and Rohrbach *et al*. [21]. On the VideoPose dataset, we directly compare with the scores reported in [12] for three of these methods [12, 18, 24] since we follow their experimental setup. On the other two datasets, we compare with the methods in [18, 21]. We re-implemented the temporal regularization scheme as described in [18] using the publicly available components of the n-best algorithm. For the comparison with [21], we used an implementation provided by the authors.

Figure 5. Sample results comparing our method with the n-best algorithm [18]. For each dataset, we show our result (ESIP + Recomb) in the first row and the result from [18] in the second row.

Method	Elbows	Wrists
Ours (ESIP+Recomb)	54.2	64.6
Tokola *et al*. [28]	49.0	37.0
Zuffi *et al*. [35]	52.0	42.0

Table 2. Comparison of elbow and wrist localization accuracies (15 pixel error threshold) on the VideoPose dataset. Note that our results are different to those in Table 1, which are shown on the version of VideoPose dataset used in [12], and has one sequence less than the version used in [28] and [35].

4.5. Discussion

Our complete model (ESIP + Recomb) outperforms the state-of-the-art method [12] by nearly 12% on the VideoPose dataset for estimating wrists (15 pixel error). On the Cooking Activities dataset, ESIP + Recomb shows 11% (elbows), 18% (wrists) and 12% (elbows), 23% (wrists) improvement over [18] and [21] respectively. In the case of the Poses in the Wild dataset, ESIP + Recomb is over 15% better than the baseline SIP model, and also shows 7% (elbows), 14% (wrists) and 16% (elbows), 21% (wrists) improvements over [18] and [21] respectively. The improvements over [18], although significant (7%, 14%), are less pronounced (compared to those on the Cooking Activities dataset), as the color-based tracking in [18] works better on our dataset. Qualitative comparisons of our method with the n-best method [18] for the three datasets are shown in Figure 5.

We summarize the results of our complete model, ESIP + Recomb, for shoulders, elbows and wrists (15 pixel error) in Table 1. Table 2 compares ESIP + Recomb with two recent works [28, 35] on the VideoPose dataset, and shows that our model is significantly better at localizing wrists. We present a few more qualitative results on the Poses in the Wild dataset in Figure 6, including two typical failure cases which are due to background clutter and occlusion.

We also analyzed the influence of the practical extensions presented in Section 3.5. Removing the three extensions from the objective function (5) reduced the accuracy of localizing elbows and wrists slightly by 3% and 2% on the Poses in the Wild dataset. On the Cooking Activities dataset, it reduced by 3% for both elbow and wrist localization. The only significant change is for estimating wrist locations on VideoPose, where the accuracy reduced by 16% (5% for estimating elbows). This is likely due to this dataset being motion-stabilized, which results in high-motion regions corresponding to body parts such as wrists (as encouraged by our enriched wrist model). Note that the performance on VideoPose without the practical extensions is still better than [18, 24], and comparable to [12], which also uses motion-based terms.

Computation Time. To achieve a fair balance between efficiency and accuracy, we use 300 poses in our candidate sets[1] for all the datasets, as it seems to contain many good poses while yielding a low computational cost. Our refined single image pose model takes about 3 seconds per image, and our recombination scheme takes about 20 seconds for 100 frames in MATLAB with a 3.6GHz Intel processor using a single core.

5. Conclusion

We presented a novel algorithm for human pose estimation in videos that achieves state-of-the-art accuracy at a reasonable computational cost. Our approach consists of two steps: (i) an extended single image pose model using optical flow cues between consecutive frames, and (ii) a flexible scheme for splitting poses into limbs, generating limb sequences across time, and recomposing them to generate better poses. Furthermore, we proposed a new challenging dataset, Poses in the Wild, containing real-world scenarios unavailable in other datasets.

Acknowledgments. This work was supported in part by the European integrated project AXES, the MSR-Inria joint project and the ERC advanced grant ALLEGRO. We thank the authors of [21] for their implementation.

[1] Varying K from 100 to 900, we observe an increase in accuracy on all datasets until about K=600, where it saturates.

(a) Correct detections (b) Background clutter (c) Occlusion

Figure 6. Sample results on the Poses in the Wild dataset with our approach ESIP + Recomb. From left to right, we show: three examples where we estimate an accurate pose and two typical failure cases. They are due to: (i) background clutter, and (ii) occlusion, which is not modeled in our framework.

References

[1] http://lear.inrialpes.fr/research/posesinthewild/.

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

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

[4] A. Balan and M. Black. An adaptive appearance model approach for model-based articulated object tracking. In *CVPR*, 2006.

[5] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-best solutions in Markov random fields. In *ECCV*, 2012.

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

[7] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 2012.

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[9] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 8(19), 2012.

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

[11] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Computers*, 100(1):67–92, 1973.

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

[13] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *CVPR*, 2010.

[14] M. W. Lee and R. Nevatia. Human pose tracking in monocular sequence using multilevel structured models. *PAMI*, 31(1):27–38, 2009.

[15] V. Morariu, D. Harwood, and L. Davis. Tracking people's hands and feet using mixed network and/or search. *PAMI*, 35(5):1248–1262, 2013.

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

[17] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 1999.

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

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

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

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

[22] B. Sapp and B. Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[23] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

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

[25] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2d image motion. In *ECCV*, 2000.

[26] C. Sminchisescu and A. Jepson. Variational mixture smoothing for non-linear dynamical systems. In *CVPR*, 2004.

[27] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *IJRR*, 22(6):371–391, 2003.

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

[29] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 2008.

[30] Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *JMLR*, 2012.

[31] D. Weiss, B. Sapp, and B. Taskar. Sidestepping intractable inference with structured ensemble cascades. In *NIPS*, 2010.

[32] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural computation*, 2000.

[33] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *PAMI*, 2012.

[34] A. Yao, J. Gall, G. Fanelli, and L. van Gool. Does human action recognition benefit from pose estimation? In *BMVC*, 2011.

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