Pose from Flow and Flow from Pose

Katerina Fragkiadaki
GRASP Laboratory
University of Pennsylvania
Philadelphia PA, 19104
katef@seas.upenn.edu

Han Hu
TNList
Tsinghua University
Beijing PRC, 100084
huh04@mails.tsinghua.edu.cn

Jianbo Shi
GRASP Laboratory
University of Pennsylvania
Philadelphia PA, 19104
jshi@seas.upenn.edu

Abstract

Human pose detectors, although successful in localising faces and torsos of people, often fail with lower arms. Motion estimation is often inaccurate under fast movements of body parts.

We build a segmentation-detection algorithm that mediates the information between body parts recognition, and multi-frame motion grouping to improve both pose detection and tracking. Motion of body parts, though not accurate, is often sufficient to segment them from their backgrounds. Such segmentations are crucial for extracting hard to detect body parts out of their interior body clutter. By matching these segments to exemplars we obtain pose labeled body segments. The pose labeled segments and corresponding articulated joints are used to improve the motion flow fields by proposing kinematically constrained affine displacements on body parts. The pose-based articulated motion model is shown to handle large limb rotations and displacements. Our algorithm can detect people under rare poses, frequently missed by pose detectors, showing the benefits of jointly reasoning about pose, segmentation and motion in videos.

1. Introduction

We study human pose detection and dense body motion estimation. With fast motion and extreme pose variation, both pose and motion estimation algorithms often fail.

Can bad motion information be useful? Our insight is that estimated body part motion, though not accurate, is often sufficient to segment body parts from their backgrounds. By matching body part segments to shape exemplars, one can improve pose estimation under large body deformations.

Can imprecise pose estimation be useful? Pose estimation, though it often fails for lower limbs, is accurate for torso and shoulders of people [18]. Such reliable detections help segmentation of body pose by adjusting motion affinities to conform with the figure-ground segmentation of the detected body joints.

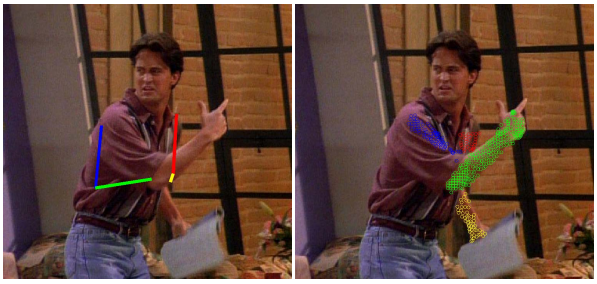

Figure 1. Pose and Flow. *Left:* Results of state-of-the-art body pose detector of [24] that combines Pb, optical flow edges, skin color and image gradient features in a structural model of human body parts in space and time. *Right:* Results of our method that mediates salient motion segmentations with body part detections for detecting human body limbs under large motion. Joint reasoning about pose, segmentation and motion proposed in our method allows 1) resolving local cue contradictions due to presence/absence of flow edges or Pb contours, 2) estimating pose under partial occlusions by reasoning about part occlusions during pose tracking, rather than during pose detection.

Our method exploits "lucky" segmentations of moving body parts to 1) index into pose space, 2) infer articulated kinematic chains in the image, 3) improve body part motion estimates using kinematic constraints. The proposed framework targets rare, widely deformed poses, often missed by pose detectors, and optical flow of human body parts, often inaccurate due to clutter and large motion.

Best practices of general object detection algorithms, such as hard mining negative examples [11], and expressive, mixture of parts representations [28], have recently led to rapid progress in human pose estimation from static images. Large number of exemplars is used to learn an alignment between articulated HOG templates and gradients in the image [28]. However, body parts at the end of the articulation chains, i.e., lower arms, are still not easily detectable. The long tails of the distribution of visual data make it hard to harvest exemplars for the rare, widely deformed part poses.

We estimate pose inversely to current detectors: our method aligns image segmentations to pose exemplars rather than learnt templates to image gradients, bypassing the need for enormous training sets. For such alignment to be possible we exploit human poses under distinct motion. Fast body part motion, despite being inaccurate, is salient and easily segmentable.

Our algorithm segments moving body parts by leveraging motion grouping cues with figure-ground segregation of reliably detected body parts, e.g., shoulders. Confident body part detections [2] induce figure-ground repulsions between regions residing in their interior and exterior, and clean up region motion affinities in places where motion is not informative. Extracted motion segments with hypothesized body joint locations (at their corners and endpoints) are matched against body pose exemplars close in body joint configuration. Resulting pose labeled segments extract occluding body part boundaries (also interior to the body), not only the human silhouette outline, in contrast to background subtraction works [15].

Pose segmentation hypotheses induce kinematic constraints during motion estimation of body parts. We compute coarse piece-wise affine, kinematically constrained part motion models, incorporating reliable pixel correspondences from optical flow, whenever they are available. Our hybrid flow model benefits from fine-grain optical flow tracking for elbows and slowly moving limbs of the articulation chain, while computes coarser motion estimates for fast moving ones. The resulting "articulated" flow can accurately follow large rotations or mixed displacements and rotations of body parts, which are hard to track in the standard optical flow framework. It propagates the pose segmentations in time, from frames of large motion to frames with no salient motion. We show such tracking is robust to pose partial self or scene occlusions.

We evaluate our framework on video sequences of TV shows. Our algorithm can detect people under rare poses, frequently missed by state-of-the-art pose detectors, by proposing a versatile representation for the human body that effectively adapts to the segmentability or detectability of different body parts and motion patterns.

2. Related work

We distinguish two main categories of work combining pose and motion estimation in existing literature: (i) Pose estimation methods that exploit optical flow information; and (ii) part motion estimation methods that exploit pose information. The first class of methods comprises methods that use optical flow as a cue either for body part detection or for pose propagation from frame-to-frame [12, 24]. Brox *et al.* [7] propose a pose tracking system that interleaves between contour-driven pose estimation and optical flow pose propagation from frame to frame. Fablet and Black [10] learn to detect patterns of human motion from optical flow.

The second class of methods comprises approaches that exploit kinematic constraints of the body for part motion estimation. Bregler and Malik [3] represent 3D motion of ellipsoidal body parts using a kinematic chain of twists. Ju *et al.* [17] model the human body as a collection of planar patches undergoing affine motion, and soft constraints penalize the distance between the articulation points predicted by adjacent affine models. In a similar approach, Datta *et al.* [8] constrain the body joint displacements to be the same under the affine models of the adjacent parts, resulting in a simple linear constrained least squares optimization for kinematically constrained part tracking. Rehg and Kanade [22] exploit the kinematic model to reason about occlusions.

In the "strike a pose" work of [21], stylized (canonical) human body poses are detected reliably, and are used to learn instance specific part appearance models for better pose detection in other frames. In this work, we follow a "strike a segment" approach by segmenting widely deforming body poses and propagating inferred body pose in time using articulated optical flow. Previously, Mori et. al.[19] have used image segments to extract body parts in static images of baseball players.

3. From Flow to Pose

We use segmentation to help the detection of highly deformable body poses. Stylized body poses are covered by an abundance of training examples in current vision datasets [9], and can be reliably detected with state-of-the-art detectors [2]. Highly deformable poses appear infrequently in the datasets, which reflects their low frequency in people's body pose repertoire. They are mostly transient in nature, the actor is briefly in a highly deformed pose, away from the canonical body configuration. It is precisely their transient nature that makes them easily detectable by motion flow.

There is an asymmetry of motion segmentability among the parts of the human body due to its articulated nature. Parts towards the ends of the articulated chains often deform much faster than the main torso (root of the body articulation tree). Lack of motion may cause ambiguities in motion segmentation of root body parts. However, such root parts can often be reliably detected thanks to their rigidity.

We exploit detectability and segmentability across different body poses and parts in a graph theoretic framework which combines motion-driven grouping cues of articulated parts and detection-driven grouping cues of torso like parts. We call it steering cuts, because detection-driven figure-ground repulsions of root parts correct (steer) ambiguous motion-based affinities. We segment arm articulated chains by constrained normalized cuts in the steered region graph.

Resulting segmentations with hypothesizing body joints at their corners and endpoints infer body pose by matching against pose exemplars. While detectors would need

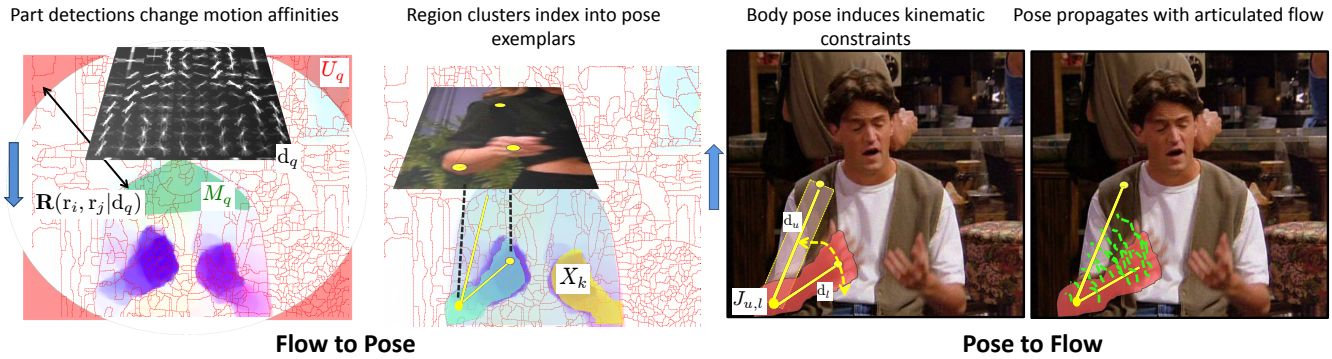

Figure 2. *Left:* Mediating motion grouping with part detections. Region motion affinities in $\mathbf{A}$ change according to confident body part detections that induce repulsions $\mathbf{R}$ between regions assigned to their foreground and background. Region clusters index into pose exemplars according to hypothesized joint locations at their endpoints. *Right:* Pose labelled segmentations propose coarse motion models coupled at the articulation joint. Coarse motion proposals compute an articulated optical flow field that can deal with large part rotations.

many training examples to learn to extract a deformed pose from background clutter [16], our pose segmentations are already freed from their backgrounds. Their good segmentation score makes them non-accidental and we use contour matching between exemplars and segmentations to select the right kinematic chain configurations.

3.1. Motion-Driven Affinities

We pursue a single frame segmentation approach from "lucky" frames that contain non-zero motion, rather than a multi-frame segmentation [13]. Multi-frame segmentation methods exploit optical flow trajectory correspondences that integrate motion estimates across multiple frames and can segment parts reliably even in frames with no motion. Large per frame deformations of lower body limbs though, often prevent such correspondences to be reliably established: in coarse-to-fine optical flow schemes, motion that is larger than the spatial extent of the moving structure cannot be recovered, since the structure is lost at the coarser levels of the image pyramid [4]. As such, we will integrate per frame optical flow estimates on region spatial support to segment frames with large motion as measured from a box around a shoulder activation.

We describe the motion of an image region in two ways: i) with the set of point trajectories, if any, overlapping with the region mask, ii) with an affine model fitted to the optical flow displacements of the region pixels. Affine motion fitting allows motion representation in places of ambiguous optical flow anchoring and sparse trajectory coverage. It only takes into account per frame motion estimates and in that sense it is weaker than multi-frame trajectory affinities.

Given a video frame I_t of video sequence I , let $\mathcal{P}$ denote the set of image pixels and let $\mathcal{R} = \{\mathrm{r}_i, i = 1 \cdots n_R\}$ denote the set of image regions. We will use r_i to refer to both the region r_i and its corresponding pixel set. Let $\mathcal{T} = \{\mathrm{tr}_a, a = 1 \cdots n_T\}$ denote the set of point trajectories of video sequence I. Between each pair of trajectories tr_a, tr_b we compute motion affinities $\mathbf{A}_T(\mathrm{tr}_a, \mathrm{tr}_b)$ encoding their long range motion similarity [6]. Each region r_i is characterized by i) an affine motion model $\mathbf{w}_i^R : \mathcal{P} \rightarrow \mathbb{R}^2$, fitted to its optical flow estimates, that for each pixel outputs a predicted displacement vector (u, v), and ii) a set of point trajectories T_i overlapping with its pixel mask. We set motion affinities between each pair of regions r_i, r_j to be:

$$\mathbf{A}(\mathrm{r}_i, \mathrm{r}_j) = \begin{cases} \dfrac{\sum_{a \in T_i, b \in T_j} \mathbf{A}_T(\mathrm{tr}_a, \mathrm{tr}_b)}{|T_i||T_j|} & \text{if } \frac{|T_i|}{|r_i|}, \frac{|T_j|}{|r_j|} > \rho \\ \dfrac{\sum_{p \in r_i \cup r_j} \exp(-\frac{1}{\sigma}||\mathbf{w}_j^R(p) - \mathbf{w}_i^R(p)||_2)}{|r_i \cup r_j|} & \text{o/w,} \end{cases}$$

where $|S|$ denotes cardinality of set S and ρ a density threshold that depends on the trajectory sampling step. The first case measures mean trajectory affinity between regions, used if both regions are well covered by trajectories. The second case measures compatibility of region affine models, being high in case the two regions belong to the projection of the same 3D planar surface.

3.2. Detection-Driven Repulsions

Each detection d_q in a detection set $\mathcal{D} = \{\mathrm{d}_q, q = 1 \cdots n_D\}$ implicitly induces figure-ground repulsive forces between the regions associated with its interior and exterior. Let mask M_q denote the pixel set overlapping with d_q [2]. We show mask M_q of a shoulder detection in Figure 2. Let $x_q^F, x_q^B \in \{0,1\}^{n_R \times 1}$ denote foreground and background region indicators for detection d_q and let U_q denote the pixel set outside a circle of radius that upper-bounds the possible arm length, as estimated from shoulder distance, shown also in Figure 2. We have:

$$x_q^F(i) = \delta\left(\frac{|r_i \cap M_q|}{|r_i|} > 0.9\right), \quad i = 1 \cdots n_R,\ q = 1 \cdots n_D$$
$$x_q^B(i) = \delta\left(\frac{|r_i \cap U_q|}{|r_i|} > 0.5\right), \quad i = 1 \cdots n_R,\ q = 1 \cdots n_D,$$

where δ is the Dirac delta function being 1 if its argument is true and 0 otherwise.

Repulsions are induced between foreground and background regions of each detector response:

$$\mathbf{R}(\mathrm{r}_i, \mathrm{r}_j|\mathcal{D}) = \max_{q|\mathrm{d}_q \in \mathcal{D}} x_q^F(i)x_q^B(j) + x_q^B(i)x_q^F(j).$$

Let $\mathcal{S}(\mathcal{D})$ denote the set of repulsive edges:

$$\mathcal{S}(\mathcal{D}) = \{(i,j) \text{ s.t. } \exists\, \mathrm{d}_q \in \mathcal{D},\ x_q^F(i)x_q^B(j) + x_q^B(i)x_q^F(j) = 1\}.$$

3.3. Steering Cut

We combine motion-driven affinities and detection-driven repulsions in one region affinity graph by canceling motion affinities between repulsive regions. In contrast to previous works that sample from a bottom-up segmentation graph [23] and post process segments with detection fit scores, we *precondition* the graph affinities to incorporate model knowledge in $\mathcal{D}$ and allow the segmentation graph to correct itself in ambiguous places. Given region motion affinities $\mathbf{A}$ and detection-driven repulsions $\mathbf{R}$, we have:

$$\mathbf{A}^{\text{steer}}(\mathrm{r}_i, \mathrm{r}_j|\mathcal{D}) = (\mathbf{1} - \mathbf{R}(\mathrm{r}_i, \mathrm{r}_j|\mathcal{D})) \cdot \mathbf{A}(\mathrm{r}_i, \mathrm{r}_j). \quad (1)$$

Inference in our model amounts to selecting the part detections $\mathcal{D}$ and clustering the image regions $\mathcal{R}$ into groups that ideally correspond to the left and right upper arms, left and right lower arms, torso and background. In each video sequence, inferring the most temporally coherent shoulder detection sequence given poselet shoulder activations in each frame worked very well in practice, since people are mostly upright in the TV shows we are working with, which makes their shoulders easily detectable. As such, instead of simultaneously optimizing over part selection and region clustering as proposed in [14], we fix the detection set $\mathcal{D}$ during region clustering.

Let $X \in \{0,1\}^{n_R \times K}$ denote the region cluster indicator matrix, X_k denote the kth column of X, respectively, and K denote the total number of region clusters. Let $\mathbf{D}_{\mathbf{A}^{\text{steer}}}$ be a diagonal degree matrix with $\mathbf{D}_{\mathbf{A}^{\text{steer}}}(i,i) = \sum_j \mathbf{A}^{\text{steer}}(i,j)$. We maximize the following constrained normalized cut criterion in the steered graph:

Steering Cut:

$$\begin{aligned} \max_X . \quad & \epsilon(X|\mathcal{D}) = \sum_{k=1}^{K} \frac{X_k^T \mathbf{A}^{\text{steer}}(\mathcal{D}) X_k}{X_k^T \mathbf{D}_{\mathbf{A}^{\text{steer}}(\mathcal{D})} X_k} \\ \text{s.t.} \quad & X \in \{0,1\}^{n_R \times K}, \quad \sum_{k=1}^{K} X_k = \mathbf{1}_{n_R}, \\ & \forall (i,j) \in \mathcal{S}(\mathcal{D}), \quad \sum_{k=1}^{K} X_k(i) X_k(j) = 0. \end{aligned} \quad (2)$$

The set of constraints in the last row demand regions connected with repulsive links in $\mathcal{S}(\mathcal{D})$ to belong to different clusters. We call them "not merge" constraints. Our cost function without the "not merge" constraints is equivalent to a Rayleigh quotient after a change of variables $Z = X(X^T \mathbf{D}_{\mathbf{A}^{\text{steer}}} X)^{-\frac{1}{2}}$ and relaxing Z to the continuous domain, and is typically solved by the corresponding spectral relaxation.

We solve the constrained normalized cut in Eq. 2 by propagating information from confident (figure-ground seeds, saliently moving regions) to non-confident places, by iteratively merging regions close in embedding distance and recomputing region affinities, similar in spirit to the multiscale segmentation in [25]. Specifically, we iterate between:

1. Computing embedding region affinities $\mathbf{W} = V \Lambda V^T$, where (V, Λ) are the top K eigenvectors and eigenvalues of the row normalized region affinity matrix $\mathbf{D}^{-1}_{\mathbf{A}^{\text{steer}}} \mathbf{A}^{\text{steer}}$. Embedding affinities in $\mathbf{W}$ are a globally propagated version of the local affinities in $\mathbf{D}^{-1}_{\mathbf{A}^{\text{steer}}} \mathbf{A}^{\text{steer}}$.

2. Merging regions $\mathrm{r}_{\tilde{i}}, \mathrm{r}_{\tilde{j}}$ with the largest embedding affinity, $(\tilde{i}, \tilde{j}) = \underset{(i,j) \notin \mathcal{S}(\mathcal{D})}{\arg\max}\ \mathbf{W}(i,j)$. We update $\mathbf{A}^{\text{steer}}$ with the motion affinities of the newly formed region.

Matrix $\mathbf{A}^{\text{steer}}$ shrinks in size during the iterations. In practice, we would merge multiple regions before recomputing affinities and the spectral embedding of $\mathbf{A}^{\text{steer}}$. Iterations terminate when motion affinities in $\mathbf{A}^{\text{steer}}$ are below a threshold. We extracted region clusters X_k with high normalized cut scores $\frac{X_k^T \mathbf{A}^{\text{steer}} X_k}{X_k^T \mathbf{D}_{\mathbf{A}^{\text{steer}}} X_k}$ even before the termination of iterations. While upper arms are very hard to delineate from the torso interior, lower arms would often correspond to region clusters, as shown in Figure 2. Foreground and background shoulder seeds help segmenting lower limbs by claiming regions of torso foreground and background, which should not be linked to the lower limb cluster. This is necessary for reliably estimating the elbow from the lower limb endpoint, as described in Section 3.4.

We compute steered cuts in graphs from multiple segmentation maps $\mathcal{R}$ by thresholding the global Pb at 3 different thresholds. Note that in coarser region maps, a lower limb may correspond to one region.

3.4. Matching Pose Segmentations to Exemplars

For each region cluster $X_k \in \{0,1\}^{n_R}$ we fit an ellipse and hypothesize joint locations J_k^1, J_k^2 at the endpoints of the major axis. Using J_k^1, J_k^2 and detected shoulder locations, we select pose exemplars close in body joint configuration as measured by the partial Procrustes distance between the corresponding sets of body joints (we do not consider scaling). We compute a segment to exemplar matching score according to pixelwise contour correspondence between exemplar boundary contours and segment boundary contours, penalizing edgel orientation difference. For

Figure 3. Articulated flow. *Left:* A video sequence ordered in time with fast rotations of left lower arm. *Right:* Motion flow is displayed as a) color encoded optical flow image, and b) the warped image using the flow. We compare the proposed articulated flow, Large Displacement Optical Flow (LDOF) [5] and coarse-to-fine variational flow of [4]. The dashed lines in the warped image indicate the ideal position of the lower arm. If the flow is correct, the warped arm will be aligned on the dashed line. Standard optical flow cannot follow fast motion of the lower arm in most cases. LDOF, which is descriptor augmented, recovers correctly the fast motion in case of correct descriptor matches. However, when descriptors capture the hand but miss the arm, hand and arm appear disconnected in the motion flow space (2nd row). Knowing the rough body articulation points allows to restrict our motion model to be a kinematic chain along the body parts. The resulting articulated motion flow is more accurate.

this we adapted Andrew Goldberg's implementation of Cost Scale Algorithm (used in the code package of [1]) to oriented contours. We also compute a unary score for each segmentation proposal, independent of exemplar matching, according to i) chi-square distance between the normalized color histograms of the hypothesized hand and the detected face, and ii) optical flow magnitude measured at the region endpoints J_k^1, J_k^2, large motion indicating a hand endpoint. We combine the 3 scores with a weighted sum.

Confidently matched pose segments recover body parts that would have been missed by the pose detectors due to overwhelming surrounding clutter or misalignment of pose. We select the two segmentations with the highest matching scores, that correspond to left and right arm kinematic chains. Each kinematic chain is comprised of upper and lower arms $\mathrm{d}_u, \mathrm{d}_l$ connected at the elbow body joint $J_{u,l}$, as shown in Figure 2.

4. From Pose to Flow

We use the estimated body pose to help motion estimation of lower limbs. Human body limbs are hard to track accurately with general motion estimation techniques, such as optical flow methods, due to large rotations, deformations, and ambiguity of correspondence along their medial axis (aperture problems). These are challenges even for descriptor augmented flow methods [5, 27] since descriptor matches may "slide" along the limb direction.

We incorporate knowledge about articulation points and region stiffness in optical flow. Articulation points correspond to rotation axes and impose kinematic constraints on the body parts they are connected to. They can thus suggest rotations of parts and predict occlusions due to large limb motion.

4.1. Articulated Flow

We use our pose labelled segmentations to infer dense displacement fields for body parts, which we call articulated flow fields. Given an arm articulated chain (left or right), let M_u, M_l denote the masks of the corresponding upper and lower arms d_u, d_l, linked at the elbow location $J_{u,l}$. Let $\mathbf{w} = (\mathrm{u}, \mathrm{v})$ denote the dense optical flow field. Let $\mathbf{w}_u^D, \mathbf{w}_l^D$ denote affine motion fields of parts d_u and d_l i.e. functions $\mathbf{w}_u^D : M_u \rightarrow \mathbb{R}^2$. Let $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}$, $\epsilon = 0.001$ denote the frequently used convex robust function, and $\phi_u(\mathbf{x}) = \exp(-|(I_2(\mathbf{x} + \mathbf{w}_u^D(\mathbf{x})) - I_1(\mathbf{x}))|^2/\sigma)$ the pixelwise confidence of the affine field $\mathbf{w}_u^D$. The cost function for our articulated optical flow reads:

$$
\begin{aligned}
\min_{\mathbf{w}, \mathbf{w}_u^D, \mathbf{w}_l^D} \quad & E(\mathbf{w}, \mathbf{w}_u^D, \mathbf{w}_l^D) = \int_\Omega \Psi(|I_2(\mathbf{x} + \mathbf{w}(\mathbf{x})) - I_1(\mathbf{x})|^2) d\mathbf{x} \\
& + \gamma \int_\Omega \Psi(|\nabla \mathrm{u}(\mathbf{x})|^2 + |\nabla \mathrm{v}(\mathbf{x})|^2) d\mathbf{x} + \\
& \beta \sum_{e \in \{u,l\}} \int_{M_e} \phi_e(\mathbf{x}) \Psi(|\mathbf{w}(\mathbf{x}) - \mathbf{w}_e^D(\mathbf{x})|^2) d\mathbf{x} \\
& + \sum_{e \in \{u,l\}} \int_{M_e} \Psi(|I_2(\mathbf{x} + \mathbf{w}_e^D(\mathbf{x})) - I_1(\mathbf{x})|^2) d\mathbf{x} \\
\text{s.t.} \quad & \mathbf{w}_u^D(J_{u,l}) = \mathbf{w}_l^D(J_{u,l}).
\end{aligned}
\tag{3}
$$

Figure 4. *Top Row:* Pose propagation with articulated optical flow. *Bottom Row:* Pose propagation with affine motion fitting to the optical flow estimates of [5]. Green outline indicates frames with pose detection and red outline indicates frames with the propagated pose. Limb motion is often too erratic to track with standard optical flow schemes, which drift to surroundings under wide deformations.

The first two terms of Eq. 3 correspond to the standard pixel intensity matching and spatial regularization in optical flow [4]. For brevity we do not show the image gradient matching term. The third term penalizes deviations of the displacement field $\mathbf{w}$ from the affine fields $\mathbf{w}_u^D, \mathbf{w}_l^D$, weighted by the pixelwise confidence of the affine displacements $\phi_u(\mathbf{x}), \phi_l(\mathbf{x})$. The forth term measures the fitting cost of the affine fields. The constraint requires the affine displacements predicted for the articulated joint by the two affine fields to be equal.

We solve our articulated flow model in Eq. 3 by computing coarse affine models for upper and lower arms and then injecting their affine displacements as soft constraints in an optical flow computation for the kinematic chain. For computing the two kinematically constrained affine fields we use "hybrid" tracking: for upper arms or the background, standard optical flow displacements are often reliable, since their motion is not erratic. We use such flow displacements to propagate foreground and background of the arm kinematic chain from the previous frame, and compute an affine motion field for the upper arm $\mathbf{w}_u^D$. Such propagation constrains i) the possible displacement hypotheses of the articulation point $J_{u,l}$, and ii) the possible affine deformations of the lower limb d_l. We enumerate a constrained pool of affine deformation hypotheses for the lower limb: it cannot be part of the background and should couple at the articulation joint with $\mathbf{w}_u^D$. We evaluate such hypotheses according to a figure-ground Gaussian Mixture Model on color computed in the initial detection frame, and Chamfer matching between the contours inside the hypothesized part bounding box and the body part contours of the previous frame, transformed according to each affine hypothesis. The highest scoring deformation hypothesis is used to compute our lower limb affine field $\mathbf{w}_l^D$. Notably, we also experimented with the method of [8] but found that it could not deal well with self-occlusions of the arms, frequent under wide deformation, as also noted by the authors.

Given part affine fields $\mathbf{w}_u^D, \mathbf{w}_l^D$, Eq. 3 is minimized with respect to displacement field $\mathbf{w}$ using the coarse-to-fine nested fixed point iteration scheme proposed in [26]. The affine displacements $\mathbf{w}_u^D, \mathbf{w}_l^D$ receive higher weights at coarse pyramid levels and are down-weighted at finer pyramid levels as more and more image evidence is taken into account, to better adapt to the fine-grain details of part motion, that may deviate from an affine model. We show results of the articulated flow in Figure 3. Articulated flow preserves the integrity of the fast moving lower arm and hand. In descriptor augmented optical flow of [26] the motion estimate of the arm "breaks" in cases of missing reliable descriptor match to capture its deformation. Standard coarse-to-fine flow misses the fast moving hand whose motion is larger that the its spatial extent.

We propagate our body segmentations in time using articulated optical flow trajectories, as shown in Figure 4. The fine grain trajectories can adapt to the part masks under occlusion while the coarse affine models prevents drifting under erratic deformations. We compare with affine fitting to standard flow estimates in Figure 4. Ambiguities of limb motion estimation due to self occlusions, non-discriminative appearance and wide deformations cause flow estimates to drift, in absence of pose informed kinematic constraints.

5. Experiments

We tested our method on video clips from the popular TV series "Friends", part of the dataset proposed in [24]. We selected 15 video sequences with widely deformed body pose in at least one frame. Each sequence is 60 frames long. The characters are particularly expressive and use a lot of interesting gestures in animated conversations.

In each video sequence, we infer the most temporally coherent shoulder sequence using detection responses from the poselet detector [2]. This was able to correctly delineate the shoulder locations in each frame. We held out a pose exemplar set from the training set of the Friends dataset, to match our steered segmentation proposals against. For each exemplar we automatically extract a set of boundary contours lying inside the groundtruth body part bounding boxes of width one fifth of the shoulder distance. We evaluate our full method, which we call "flow$\rightarrow$ pose $\rightarrow$ flow", as

Figure 5. *Top Row:* Our method. *Middle Row:* Sapp *et al.* [24] *Bottom Row:* Park and Ramanan [20].

well as our pose detection step only, without improving the motion estimation, but rather propagating the pose in time by fitting affine motion models to standard optical flow [5]. We call this baseline "flow → pose".

We compare against two state-of-art approaches for human pose estimation in videos: 1) the system of Sapp *et al.* [24]. It uses a loopy graphical model over body joint locations in space and time. It combines multiple cues such as Probability of Boundary, optical flow edges and skin color for computing unary and pairwise part potentials. It is trained on a subset of the Friends dataset. It assumes the shoulders positions known and focuses on lower arm detection. 2) The system of Park and Ramanan [20]. It extends the state-of-the-art static pose detector of [28] for human pose estimation in videos by keeping N best pose samples per frame and inferring the most coherent pose sequence across frames using dynamic programming. We retrained the model with the same training subset of Friends as [24] but the performance did not improve due to the low number of training examples.

Our performance evaluation measure is distance of the detected elbows and wrists from groundtruth locations, same as in [24]. We show in Figures 6 and 7 the percentage of correct wrists and elbows as we vary the distance threshold from groundtruth locations. The flow→pose→flow and pose→flow methods perform similarly in tracking the detected elbows since upper arms do not frequently exhibit erratic deformations. The two methods though have a large performance gap when tracking lower arms, whose wide frame-to-frame deformations cause standard optical flow to drift. This demonstrates the importance of improving the motion estimation via articulation constraints for tracking the pose in time.

The baseline system of [24] uses optical flow edge as a cue for part detection. We attribute its worse performance

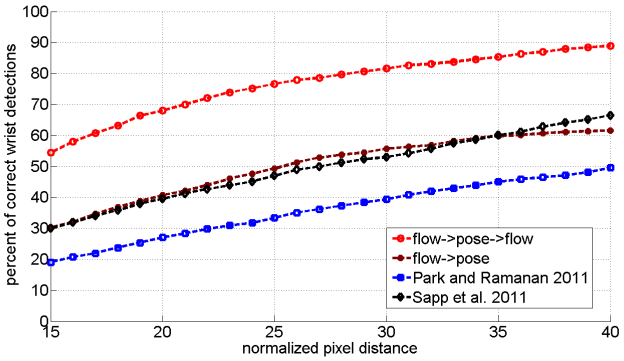

Figure 6. Evaluation of wrist locations. Our system outperforms the baseline systems by a large margin.

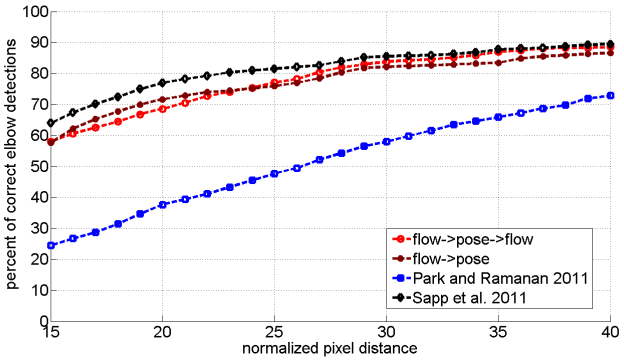

Figure 7. Evaluation of elbow locations. The system of [20] does not use motion and has significantly lower performance than our method and [24].

for wrist detection to two factors: 1) it learns a single weight combination for optical flow edges, Pb and image gradi-

ents for each part or pair of parts, which may create contradictions in case of absence of motion. 2) Optical flow edges may not align well with the body part boundaries due to optical flow "bleeding" effect. Our method detects the most saliently moving arms in the available frames, so by construction does not have contradictions between presence and absence of motion. We recover from mis-alignments of optical flow with part boundaries by computing a flow based region segmentation, rather than using optical flow as a raw feature into part detection.

Our method provides accurate spatial support for the body parts, robust to intra-body and scene occlusions. In contrast to standard pose detectors, and also our baseline systems, our method does not require all body parts to be present in each frame. The lack of specified wrist and elbow detectors makes our wrist and elbow localization occasionally poor (see last column of Figure 5) while lying inside the body part.

6. Conclusion

We proposed an approach that detects human body poses by steering cut on motion grouping affinities of lower limbs and figure-ground repulsions from shoulder detections. We focus on detecting rare, transient in nature poses, often under-represented in the datasets and missed by pose detectors. Our steering segmentations extract lower limbs from their surrounding intra-body and background clutter. Arm articulated chains resulting from matching such segmentations to exemplars, are used to provide feedback to dense body motion estimation about articulation points and region stiffness. Resulting flow fields can deal with large per frame deformations of body parts and propagate the detected pose in time, during its deforming posture. Our flow to pose to flow process is able to infer poses under wide deformations that would have been both too hard to detect and too hard to track otherwise.

References

[1] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009.

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, 1998.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[5] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *TPAMI*, 2010.

[6] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*. 2010.

[7] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints. In *ECCV*, 2006.

[8] A. Datta, Y. A. Sheikh, and T. Kanade. Linear motion estimation for systems of articulated planes. In *CVPR*, 2008.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88, 2010.

[10] R. Fablet and M. J. Black. Automatic detection and tracking of human motion with a view-based representation. In *ECCV*, 2002.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32, 2010.

[12] V. Ferrari, M. Marn-Jimnez, and A. Zisserman. 2D human pose estimation in TV shows. In D. C. et al., editor, *Statistical and Geometrical Approaches to Visual Motion Analysis*, LNCS, pages 128–147. Springer, 1st edition, 2009.

[13] K. Fragkiadaki and J. Shi. Exploiting motion and topology for segmenting and tracking under entanglement. In *CVPR*, 2011.

[14] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi. Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In *ECCV*, 2012.

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

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

[17] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *FG*, 1996.

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

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

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

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

[22] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, 1995.

[23] B. C. Russell, A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

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

[25] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. In *CVPR*, 2000.

[26] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*. 2010.

[27] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. *TPAMI*, 34, 2012.

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