Estimating Human Pose with Flowing Puppets

Silvia Zuffi, Javier Romero, Cordelia Schmid, Michael J. Black

Estimating Human Pose with Flowing Puppets

Silvia Zuffi[1,3] Javier Romero[2] Cordelia Schmid[4] Michael J. Black[2]
[1]Department of Computer Science, Brown University, Providence, RI 02912, USA
[2]Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
[3]ITC - Consiglio Nazionale delle Ricerche, Milan, Italy
[4]INRIA - Grenoble, France

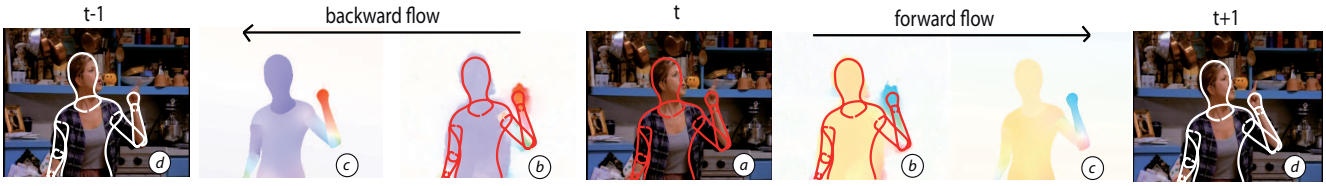

Figure 1. **Flowing puppets**. (a) Frame with a hypothesized human "puppet" model; (b) Dense flow between frame (a) and its neighboring frames; (c) The flow of the puppet is approximated by a part-based affine motion model; (d) The prediction of the puppet from (a) into the adjacent frames using the estimated flow.

Abstract

We address the problem of upper-body human pose estimation in uncontrolled monocular video sequences, without manual initialization. Most current methods focus on isolated video frames and often fail to correctly localize arms and hands. Inferring pose over a video sequence is advantageous because poses of people in adjacent frames exhibit properties of smooth variation due to the nature of human and camera motion. To exploit this, previous methods have used prior knowledge about distinctive actions or generic temporal priors combined with static image likelihoods to track people in motion. Here we take a different approach based on a simple observation: Information about how a person moves from frame to frame is present in the optical flow field. We develop an approach for tracking articulated motions that "links" articulated shape models of people in adjacent frames through the dense optical flow. Key to this approach is a 2D shape model of the body that we use to compute how the body moves over time. The resulting "flowing puppets" provide a way of integrating image evidence across frames to improve pose inference. We apply our method on a challenging dataset of TV video sequences and show state-of-the-art performance.

1. Introduction

We address the problem of estimating the 2D pose of a person in a monocular video sequence captured under uncontrolled conditions, without manual initialization. In a single frame, pose estimation is challenging and current methods tend to do poorly at estimating the pose of the limbs. Arms and hands are relatively small and can be difficult to localize due to occlusion, motion blur, accidental alignment, ambiguities, poor contrast, etc. Incorporating information from multiple frames may alleviate some of these problems but the question is how best to achieve this? Previous approaches use image evidence in individual frames and then try to infer a coherent sequence of poses by imposing priors that encode smooth motion over time. Such approaches can work well for *tracking* where an initial pose is given but, as we describe below, are difficult to use for the general pose inference problem.

Like previous work we want to exploit the inherent consistency of appearance and motion over time. However, we take a novel approach that does not rely on human motion priors. Instead, we exploit optical flow in three ways: 1) to exploit image evidence from adjacent frames, 2) to propagate information over time, and 3) to provide richer cues for pose estimation.

Our approach is enabled by recent advances in methods for dense optical flow computation and by a recently introduced 2D model of articulated human body shape, the Deformable Structures model (DS) [24]. The availability of accurate estimates of dense optical flow allows us to consider the flow as an observation, while the articulated model of 2D body shape provides a tool for modeling the regions of motion of a moving person. The question is: How can optical flow be incorporated to make the pose inference problem simpler and more accurate?

Consider the problem of estimating body pose in Fig. 1. Assume we have a hypothesis for the body at frame t (Fig. 1(a), red). In any given frame, the image evidence may be ambiguous and we would like to combine evidence from multiple frames to more robustly infer pose. Due to the complexity of human pose, we perform inference using a distribution of "particles" at each frame, where each particle represents the pose of the body. If we are lucky and have particles at frame t and $t+1$ that are both correct, then the poses in each frame explain the image evidence and the change in pose between frames is consistent with the flow. This should increase our confidence in the solution.

Unfortunately such an approach is not practical. Estimating the pose of the body in two frames simultaneously effectively doubles the size of the state space which, for articulated body models, is already high. Alternatively, if we independently estimate the pose in both frames then, given the high-dimensional space and a small set of particles, we will have to be extremely lucky to have two poses that are consistent with the image evidence in both frames and the optical flow. We need a different approach.

Our first solution is to estimate the pose of the body only at one frame (keeping the dimensionality under control) and to use the optical flow to check how good this solution is in neighboring frames. We refer to the body model as a "puppet" because it can be "puppeteered" by the optical flow. Given a pose at frame t we use the computed dense optical flow (Fig. 1(b)) to predict how the puppet should move into the next frame, forwards and backwards in time. The puppet flow (Fig. 1(c)), estimated from the dense optical flow, provides the prediction of the puppet in the next fame (Fig. 1(d)). We then extend our image likelihood to take into account evidence from the neighboring frames. The advantage is that inference takes place for a single puppet at a time but we are able to incorporate information from multiple frames.

Image evidence is computed in each frame using a DS model [24]. We describe upper body pose estimation but the method should be applicable to full body pose as well. This model captures the rough shape of a person and how the shape of the body parts deform with pose. We go beyond previous work to train a multi-scale image likelihood that captures the statistics of image gradients along the contour. By learning the model at multiple scales, we capture information about how real people deviate from the model.

Our second use of optical flow is in search. Our optimization uses a particle-based stochastic search method [13]. We initialize particles on each frame of the video sequence using a state-of-the-art single-frame pose estimation method [23]. We take the most likely particles in a given frame and use the puppet flow to predict their poses in adjacent frames. This enriches the particle set at neighboring frames. Inference always happens in a single frame but the two methods above serve to incorporate information from adjacent frames. We generate additional pose proposals that incorporate information about the possible location of hands based on image and flow evidence; this is our third use of flow.

We compare our method with [23] and [20] on the VideoPose2.0 dataset [20]. VideoPose2.0 is a complex and challenging benchmark for pose estimation methods in video sequences that includes very difficult sequences where the appearance of the people can be easily confounded with the background. Until now Sapp et al. [20] had the best results on this dataset but they rely on knowing the correct scale of the person and a bounding box used to rescale and crop the video frames. Here we remove these restrictions while obtaining more accurate estimates of the wrists.

In summary our work proposes a new way of integrating information over time to do human pose estimation in video. The key idea is to use the optical flow field to define "puppets" that "flow" from one time to the next, allowing us to integrate image evidence from multiple frames in a principled and effective way and to propagate good solutions in time. A good pose is one that is good in multiple frames and agrees with the optical flow.

2. Background and Related Work

Human pose tracking in video. Much of the early work in the field addresses the tracking problem. It assumes either that there is a known template [4] or the first pose is known [15]. The tracking literature is extensive and beyond our scope.

Human pose estimation in still images. There is a similarly large literature on 2D human pose estimation in static images. Again a full review here is not feasible but the most relevant recent work is based on pictorial structures (PS) models [1, 7, 8, 18, 19]. Such models are widely used but still have trouble accurately localizing human arms. Here we use the Flexible Mixtures of Parts (FMP) model [23] on each still video frame to provide an initialization. FMP is one of the most adopted methods for human pose estimation, due to its computational efficiency and ability to detect people at different scales.

Human pose from video. Surprisingly little work has addressed the combination of monocular pose estimation with tracking in uncontrolled environments. The problem is sometimes referred to as *articulated motion parsing* [20]. In early work, Ramanan et al. [17] assume there is at least one frame in the video sequence where the person is in an easy to detect pose. Based on this detection, they build a person-specific appearance model and perform independent pose estimation on each frame using this appearance model. A similar approach is used by Buehler et al. [3] who introduce temporal information by identifying key frames with reliable poses. These effectively become anchor frames

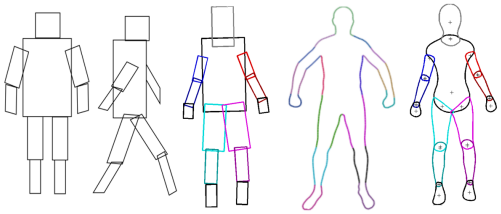

Figure 2. **2D models of human shape**. From the left: Cardboard People (in two viewpoints), Pictorial Structures (PS), Contour People (CP) and Deformable Structures (DS).

and they exploit temporal coherence to link poses in the intermediate frames. Ferrari et al. [9] formulate a spatio-temporal parsing problem and perform simultaneous pose estimation over a set of frames. In addition to single-frame potentials, the model includes temporal dependencies that represent continuity of appearance and pose. These methods rely on static image likelihoods and do not use optical flow.

Sapp et al. [20] exploit optical flow information to locate foreground contours. This integrates well with their pose estimation method, which exploits image contours and region-based likelihoods. The idea of using flow discontinuities as a cue for pose estimation dates at least to [21] on 3D body pose estimation in monocular video. In a recent work, Fragkiadaki et al. [10] exploit optical flow for segmenting body parts and propagating segmentations over time.

Body representation. The representation of the body for monocular articulated motion parsing has not received much attention but, we argue, is critically important. While there are a wide range of 3D body representations, here we focus on 2D, where the range of models is remarkably limited. The most common approach represents the body as rectangular (or polygonal) regions [1, 4, 7, 8, 15, 18] (Fig. 2). The pose of the body can then be represented either explicitly by a kinematic tree [4, 15] or by a probabilistic collection of parts [1, 8].

In contrast to polygonal parts, the Contour People [11] and Deformable Structures [24] models are derived from a realistic 3D model of body shape and better capture the 2D shape of the person including perspective effects, foreshortening, and non-rigid deformations with pose. The DS model is like a pictorial structures model in that its parts are connected only probabilistically and the configuration of parts is inferred using belief propagation.

Note that in [6] body pose is displayed in a way that looks like a DS model but is not. They take the rectangular body parts of a standard PS model and smooth them with the probability distribution for the part. This produces attractive contours but these are unrelated to the model or the inference.

Alternatively Andriluka et al. [1] work with a traditional rectangular part model but use shape contexts to learn a model that captures information about the shape of the parts within these regions. This is not an explicit model of shape and is consequently not appropriate for our task. Guan et al. [12] propose a 2D model of clothed body shape but the model is not articulated.

For our application, there are two properties that a representation must satisfy. First, it must be able to represent occlusion. Pure part-based models like pictorial structures, have no notion of what is in front of what. To make sense of the optical flow in the scene this is necessary and is supported by our model (though crudely here). Second, it should approximate the shape of the body. As we will see, to "flow" the puppet in time requires that we associate observed optical flow with body parts. If the parts match the size and shape of parts in the image, this is easy. Rectangular parts would make this much harder. Consequently, here we take the idea of the DS model but, instead of a distributed collection of parts, our state space represents the full pose of the model; this allows us to model occlusions. We augment the DS model with a scale parameter to capture the overall size of the person in the image. The DS model is learned for random poses and cameras. We introduce prior knowledge on camera location and pose in TV shows by redefining the mean DS shape as the average shape in the Buffy training set [9] annotated with the DS model. We also go beyond previous work to learn a new, multi-scale, image likelihood that captures image statistics along the contour of the puppet.

3. Model

We briefly summarize the DS model and refer the reader to [24] for details. DS is a gender-specific part-based probabilistic model, where contour points of body parts are represented in local coordinate systems by linear models of the form

$$\begin{bmatrix}\mathbf{p}_i \\ \mathbf{y}_i\end{bmatrix} = \mathbf{B}_i\mathbf{z}_i + \mathbf{m}_i \tag{1}$$

where $\mathbf{p}_i$ are contour points, $\mathbf{y}_i$ are joint points, $\mathbf{z}_i$ are PCA (Principal Component Analysis) coefficients, $\mathbf{B}_i$ is a matrix of principal components, and $\mathbf{m}_i$ is the mean part contour. Let $\mathbf{l}_i = (\mathbf{c}_i, \theta_i, \mathbf{z}_i)$, where $\mathbf{c}_i$ is the center of the part i and θ_i is the part orientation. The correlation between the *shape coefficients*, $\mathbf{z}_i$, and body pose parameters captures how shape varies with pose and is modeled with pairwise Multivariate Gaussian distributions over the relative pose and shape coefficients of connected body parts. The probability of a model instance is factored as:

$$p(\mathbf{l}|\pi_{DS}) \propto \prod_{(i,j)\in E} p_{ij}(\mathbf{l}_i, \mathbf{l}_j|\pi_{ij}) \tag{2}$$

where E is the set of pairs of connected parts and π_{DS} represents the model parameters. The DS model does not in-

Figure 3. **DS puppet layer**. (1) Frame; (2) Corresponding puppet layer with parts ordered by fixed order. The warmer the color, the closer to the camera.

clude a scale variable in the potentials, but a scale factor can be specified and it is used to convert the model from the DS model space to image pixel coordinates.

Let $\mathbf{x}_t$ be a vector of DS model variables and the scale at time t (i.e. $\mathbf{x}_t = [\mathbf{l}_t, s_t]$), let I_t be the image frame at time t, and $U_{t,t+1}$ the dense optical flow between images I_t and I_{t+1}. We define the posterior distribution over the DS model variables and scale for each frame in the sequence of N frames as:

$$p(\mathbf{X}|\mathbf{I},\mathbf{U},\pi_{DS}) \propto \prod_{t=1}^{N-1} p(I_{t+1}|\hat{\mathbf{x}}_{t+1}) p(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_t, U_{t,t+1}) \prod_{t=1}^{N} p(I_t|\mathbf{x}_t) \prod_{t=1}^{N} p(\mathbf{l}_t|\pi_{DS}) \prod_{t=1}^{N} p(s_t|\pi_s) \quad (3)$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$, $\mathbf{I} = [I_1, ..., I_N]$, $\mathbf{U} = [U_{1,2}, ..., U_{N-1,N}]$, $p(\mathbf{l}|\pi_{DS})$ is from Eq. (2), $p(s_t|\pi_s)$ is a prior on scale, $p(I_t|\mathbf{x}_t)$ is the static image likelihood for the frame at time t, $p(I_{t+1}|\hat{\mathbf{x}}_{t+1})$ is the static image likelihood for the frame at $t+1$, evaluated for $\hat{\mathbf{x}}_{t+1}$, which is the "flowing puppet" of $\mathbf{x}_t$ given the flow $U_{t,t+1}$ (see below). Here our likelihood uses flowing puppets in the forward direction, but our formulation is general and can be extended to consider flowing puppets generated with backward flow and for more than one time step.

3.1. Flowing puppets

Given a DS puppet defined by the variables $\mathbf{x}_t$, and given the dense flow $U_{t,t+1}$, the corresponding flowing puppet for frame $t+1$ is generated by propagating $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ through the flow. The conditional probability distribution $p(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_t, U_{t,t+1})$ expresses the noisy generative process for the flowing puppet $\hat{\mathbf{x}}_{t+1}$.

Exploiting the part-based representation of the DS model, we define a layered model of the body parts with a fixed depth ordering. We assume the torso is the most distant part, then comes the head, the right and the left upper arms, then the right and left lower arms. Figure 3 shows an example of the layer map, where warm colors indicate parts that are closer to the camera. Given the visibility mask for each body part, we consider the corresponding pixels in the optical flow map $U_{t,t+1}$. This is where the DS body shape representation becomes important. Figure 1(c) shows a puppet, $\mathbf{x}_t$, overlaid on the forward and backward optical flow fields (i.e., computed from t to $t+1$ and from t to $t-1$). We fit an affine motion model to the optical flow vectors within each body part. The resulting puppet flow field is illustrated in Fig. 1(c); this is our estimate for how the puppet should move from frame to frame. We then apply the estimated affine motion to the joints of each part, resulting in predicted puppets, $\hat{\mathbf{x}}_{t-1}$ and $\hat{\mathbf{x}}_{t+1}$, at the adjacent frames (Fig. 1(d), white). Our current process of generating the flowing puppet does not include a noise model, thus the probability distribution $p(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_t, U_{t,t+1})$ is simply a delta function centered on the predicted puppet.

3.2. Image likelihood

The static image likelihood is defined by three terms: a contour-based term $p_c(I_t|\mathbf{x}_t)$ that encourages alignment of the DS puppet contour with the edges in the image, a color term $p_s(I_t|\mathbf{x}_t)$ that encodes the knowledge that lower arms and hands are likely to be skin, and a hand likelihood $p_h(I_t|\mathbf{x}_t)$ computed from a hand probability map generated from a hand detector:

$$p(I_t|\mathbf{x}_t) = p_c(I_t|\mathbf{x}_t) p_s(I_t|\mathbf{x}_t) p_h(I_t|\mathbf{x}_t). \quad (4)$$

The DS model we use is learned from a 3D model that does not include hand pose variations, consequently our 2D model does not have separate hand parts with their own articulation parameters. Instead, hands are included in the model of the shape of the lower arm. To define a region-based likelihood for hand color, we simply consider the image area surrounded by the points in the lower arm that correspond to hand contours.

Similar to [24], we define the contour-based likelihood $p_c(I_t|\mathbf{x}_t)$ as the output of an SVM classifier with a feature vector of HOG descriptors [5] that are steered to the contour orientation and computed along the model contour. In order to obtain a likelihood model that is more robust to scale variations, we compute the features at different scales. Figure 4 shows an example for the upper arm: We use a 3-level pyramid; HOG cells are placed at contour points (blue), inside the contour (red), and outside (green). We learn the contour-based image likelihood from the Buffy dataset [9].

The skin-color likelihood $p_s(I_t|\mathbf{x}_t)$ is defined by a histogram of skin chroma and hue color coordinates in the CIE L*C*h* color space. We define a log probability map for the image, and compute log skin color likelihood for the hands, and the lower arms without hands, as the average log probability value in the part region. The skin color likelihood is learned from a dataset of skin colors downloaded from the web as in [24].

The hand likelihood $p_h(I_t|\mathbf{x}_t)$ is based on a hand probability map generated by a hand detector using optical flow.

Figure 4. **Contour-based likelihood features**. HOG descriptors are steered to the contour orientation and computed at contour points (blue), inside (red) and outside the contour (green) in a 3-level pyramid.

Figure 5. **Hand detection**. Example of output from the hand detector trained on optical flow. Image (left), optical flow (center), and hand probability map defined from running a flow-based hand detector on the flow (right).

The hand detector is defined as in [20] by computing the gradient magnitude of the flow, then learning an SVM classifier. The hand probability map is built as the max response from the detector at each image location over a discrete set of hand orientations. The detector is learned from training images from the VideoPose2.0 dataset [20] where we have manually annotated hands with oriented bounding boxes. Figure 5 shows an example of the hand probability map.

4. Inference

The DS model state space consists of pose and shape parameters. To reduce the number of variables during inference, we define the configuration of a body as only the location of the joint points $\mathbf{y}_t$, marginalizing out the shape parameters, and thus reducing the number of variables. We use a set of 11 joints points, namely nose, neck, right and left shoulders, belly button, right and left hips, elbows and wrists. From these joints, we can easily compute the DS puppet parameters, $\mathbf{l}_i = (\mathbf{c}_i, \theta_i, \mathbf{z}_i)$, where the shape coefficients $\mathbf{z}_i$ represent the expected shape for each part. The state space for a puppet in a frame is then $\mathbf{x}_t = [\mathbf{y}_t, s_t]$, where s_t is the puppet scale.

We convert the probabilistic formulation, Eq. (3), into an energy

$$E(\mathbf{X}) = \lambda_{DS} \sum_{t=1}^{N} E_{DS}(\mathbf{y}_t) + \lambda_c \sum_{t=1}^{N} E_c(\mathbf{x}_t) + \lambda_s \sum_{t=1}^{N} E_s(\mathbf{x}_t) + \lambda_h \sum_{t=1}^{N} E_h(\mathbf{x}_t) + \lambda_c \sum_{t=1}^{N-1} E_c(\hat{\mathbf{x}}_{t+1}) + \lambda_s \sum_{t=1}^{N-1} E_s(\hat{\mathbf{x}}_{t+1}) + \lambda_h \sum_{t=1}^{N-1} E_h(\hat{\mathbf{x}}_{t+1})$$

where $E_{DS}(\mathbf{y}_t) = -\log p(\mathbf{l}_t|\pi_{DS})$ (see Eq. (2)), $E_c(\mathbf{x}_t)$, $E_s(\mathbf{x}_t)$ and $E_h(\mathbf{x}_t)$ are the energy terms associated with the contour-based, the skin-color based, and the hand-detector-based image likelihood on the current frame, respectively. $E_c(\hat{\mathbf{x}}_{t+1})$, $E_s(\hat{\mathbf{x}}_{t+1})$ and $E_h(\hat{\mathbf{x}}_{t+1})$ are the negative log likelihoods of the puppet in frame t propagated to the frame $t+1$ through the dense optical flow $U_{t,t+1}$. We use a uniform prior for scale, as bounds for the scale parameters are set in the optimizer. The weights for the energy terms are estimated with grid search on a validation set.

We seek a maximum a posteriori estimate and minimize the energy using a novel iterative approach based on frame optimization and propagation. For the frame optimization, we adopt Particle Swarm Optimization (PSO) [13]. PSO searches the parameter space by perturbing the particles. Perturbing the vertices can produce implausible puppets so we first convert the pose into a joint angle representation, do this perturbation in joint angle space, convert back to joint positions, and then to the expected DS model to obtain contours and regions. Inspired by [14] we employ a hierarchical strategy optimizing first the torso, head and right arm, then the whole puppet. In addition, we resample the particles according to their energy. PSO requires setting bounds for the variables to optimize: We estimated bounds for scale, limb angles, and limb lengths from the VideoPose2.0 training set.

The process of optimization and propagation is visually described in Fig. 6. We start by initializing a set of P particles on each frame (Fig. 6, first row). Then the video sequence is scanned forward and backward to propagate the best M particles from a frame to the next using the flow (Fig. 6, second row). Each frame in the sequence is then optimized in turn, using PSO, starting from the first frame, and proceeding forward for all the frames then backwards. Figure 6, third row, shows examples of particles after PSO. After optimizing pose in each frame, the best M particles are propagated to the neighbors, forward and backward, using the flow (Fig. 6, fourth row). After propagation, each frame has P+2M particles, but only the best P particles are retained for the frame. This process of optimization and propagation iterates for a defined number of runs, R. In our experiments we used P=40, M=5, R=8. We run the optimization with 3 different seeds for the random number generator and select

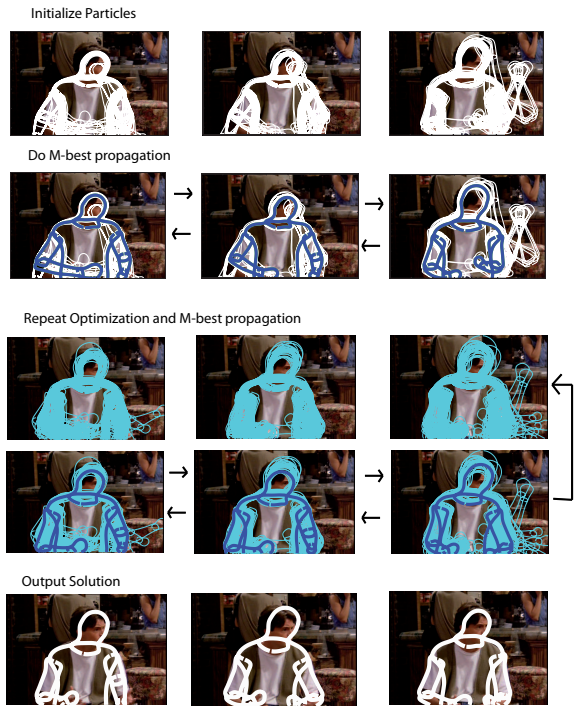

Figure 6. **Particle-based optimization**. Particles are initialized on each frame (first row), then the M best are propagated through the flow forward and backward (second row). For a defined number of iterations particles are then locally optimized, then the M best are propagated to the neighbors (third and fourth row). Then the best particle on each frame is returned as the solution (last row).

the solution with the minimum energy.

We use the Flexible Mixtures of Parts (FMP) model [23] to provide the optimization with a good starting guess for the solution. We use the code provided by the authors with their model trained on the Buffy dataset. The FMP model generates a "stickman" as output. In order to map the FMP stickman to the DS model at a proper scale, we learn a regression function between the scale of the stickman and DS from manually annotated frames of the VideoPose2.0 training set. To further help the optimizer, we generate additional initial puppets relocating the wrists of the FMP solution to likely hands locations. We exploit both image cues and motion cues for hand detection. We train and use a hand detector based on the method described in [16], which uses image features like statistics of image gradients and colors, to provide initial guesses for the hand location. We also exploited the hand detector trained on optical flow described in Sec. 3.2 to generate proposals for wrist positions.

5. Experiments

The VideoPose2.0 dataset [20] contains 1225 frames from two popular TV shows (Friends and Lost) corresponding to 44 clips. The dataset is divided into 706 training frames and 519 test frames in 18 clips. For consistency with [20], we use the dataset with only every other frame of the original video sequences. It contains frames at the original size, and frames that have been cropped and rescaled to have the person in the middle of the frame to meet the needs of the pose estimation method of [20]. In contrast to [20] we use the original frames, since we model scale explicitly and estimate during optimization. We annotate the clips for gender and use a DS model of the appropriate gender. The dense flow is computed with the method of [22] in both the forward and backward time direction.

Results are reported as in [20] as the percentage of joints that have a distance lower than a threshold in pixels from the ground truth (Fig. 7). As a baseline we report results for Yang et al. [23], which we use during initialization. Note that it performs rather poorly but was not trained for this dataset. We compare different variants of our method (FP for Flowing Puppets). First, we report results without and with median filtering as applied in [20]. Second, to show the benefit of our optimization strategy, we show results obtained without exploiting the dense flow for propagation and likelihood (FP, -flow). We report performance better than the state-of-the-art for wrists, significantly improving over [23], and at the performance level of [20] for elbows. We can observe that our approach performs significantly better than the static image detector of [23]. A recent paper [10] also performs pose estimation on the VideoPose2.0 dataset, but for testing they select a set of the clips that is different from the one specified in the dataset; a direct comparison is not possible. Figure 8 shows several examples of correctly predicted body pose, with the DS puppet overlaid on the image and on the optical flow. Figure 9 shows some representative failure cases.

6. Conclusions

Given recent improvements in the accuracy of optical flow estimation, we argue that it is a useful source of information for human pose estimation in video. Here we use flow in a novel way to make predictions about the pose of the body in neighboring frames. Given a representation of body shape, we use the optical flow forwards and backwards in time from a given frame to predict how the body should move, creating what we call a flowing puppet. If the body pose is correctly estimated in the current frame, and the flow is accurate, then our method should accurately predict the pose in neighboring frames. We use this to construct an extended energy function that incorporates image evidence from multiple frames. We also use our flowing puppets to propagate good candidate poses during optimization and to hypothesize putative hand locations.

The approach improves accuracy and robustness relative to a baseline method that does not use puppet flow. If the pose in one frame is ambiguous, it may not be in neigh-

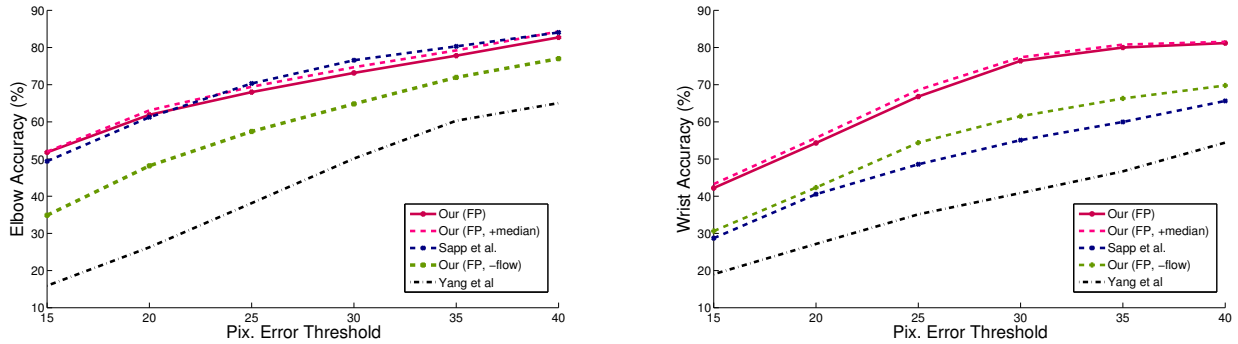

Figure 7. **Results**. Accuracy of elbow (left) and wrist detection (right) for different threshold distances from ground truth. We significantly improve over our baseline (Yang and Ramanan [23]) and over the state-of-the-art (Sapp et al. [20]) in wrist detection. FP stands for Flowing Puppet.

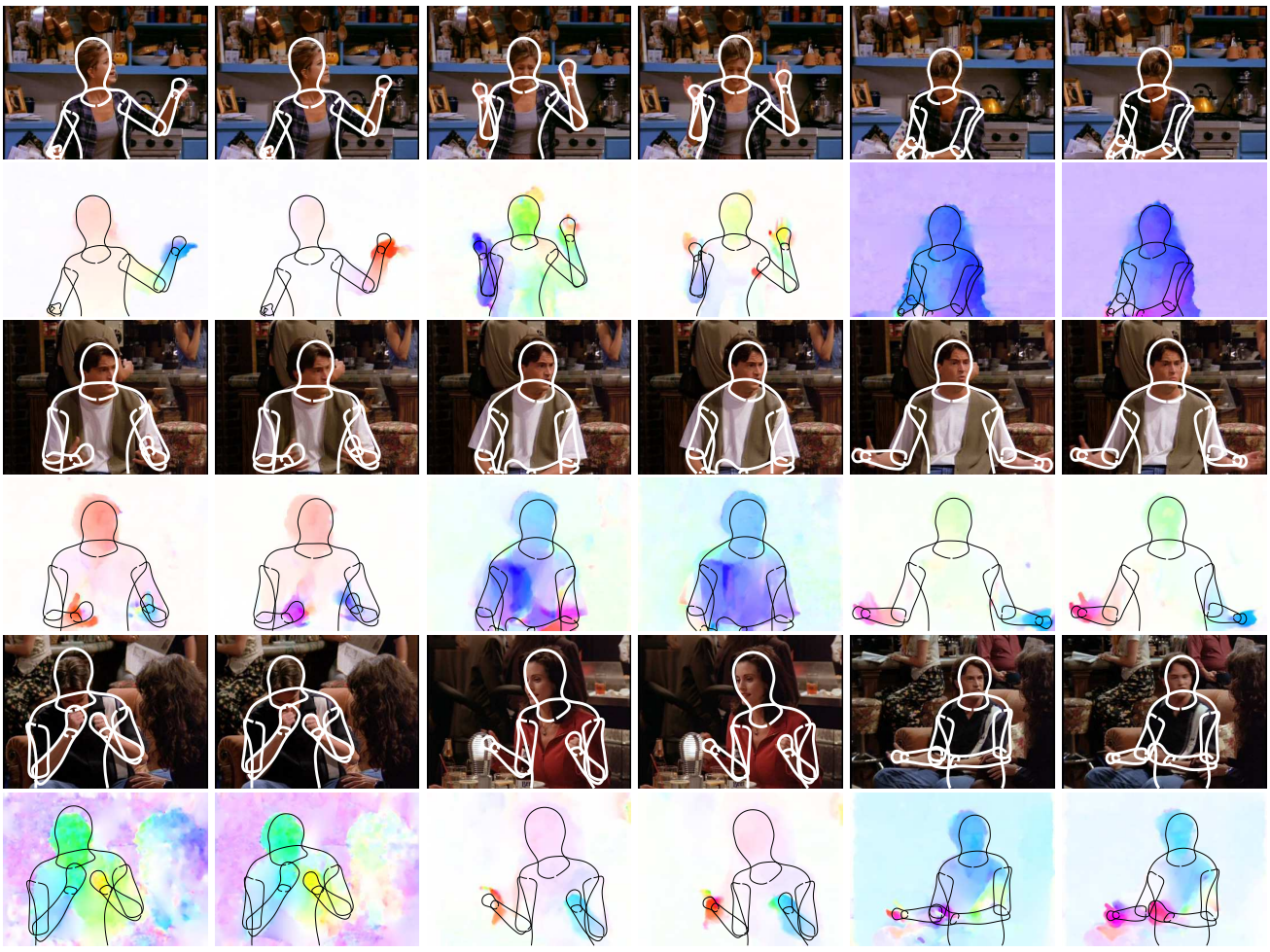

Figure 8. **Estimated body pose**. Successful detection results from 9 test clips are shown (2 frames per clip). Images are shown with the estimated puppet overlaid in white. Below each image is the estimated forward flow field color coded as in [2] with the puppet overlaid in black.

boring frames. Thus by integrating evidence over time we may be able to resolve such ambiguities. This represents a novel approach to temporal estimation of body pose in video. In particular there is no temporal prior to restrict the motion. Instead we rely on flow to "link" information across time. We tested the method on the challenging

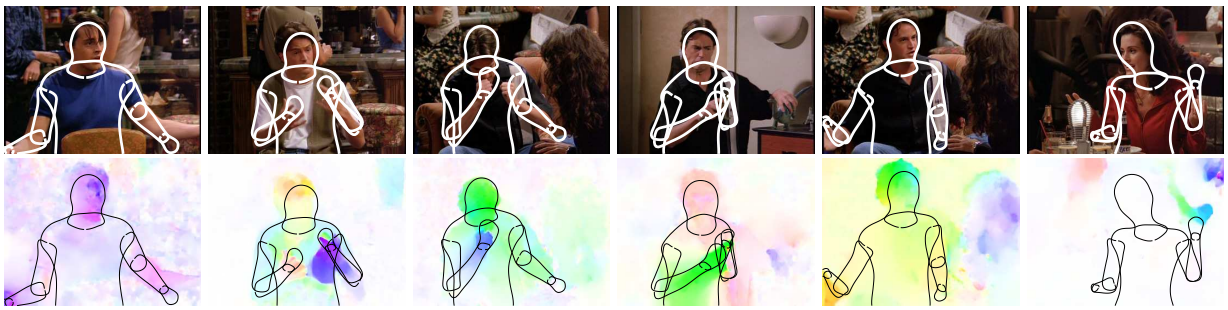

Figure 9. **Estimated body pose**. Examples of failure cases. In all cases the image evidence supports incorrect poses.

VideoPose2.0 dataset and found that we improved over previous results in the difficult problem of localizing the wrists.

This work opens up many possibilities for future research. Other properties of the flow, such as motion boundaries could be used. Reasoning about depth ordering of the parts should be added. Some sort of temporal reasoning could be included to propagate information beyond the neighboring frames, for example using particle smoothing or a Viterbi-like algorithm. More accurate models of body shape, hair, and clothing might also improve the results. Note that, while VideoPose2.0 contains background motions, more work should be done to evaluate robustness to multiple moving people and other scene motion. Finally, flowing puppets could be used to build a temporally consistent appearance model across several frames, which could provide stronger image evidence.

Acknowledgements. CS was supported in part by the ERC advanced grant Allegro. We thank Ben Sapp for helpful feedback on his code.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR*, pp. 1014–1021, 2009. 2, 3

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011. 7

[3] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing. *IJCV*, 95(2):180–197, 2011. 2

[4] T.-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, pp. 239–245, 1999. 2, 3

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, p. 886-893, 2005. 4

[6] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. *BMVC*, pp. 1–11, 2009. 3

[7] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 99(2):190–214, 2012. 2, 3

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2, 3

[9] V. Ferrari, M. Marin-Jiminez, and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR*, pp. 1–8, 2008. 3, 4

[10] K. Fragkiadaki, H. Hu, J. Shi. Pose from flow and flow from pose estimation. *CVPR*, pp. 2059–2066, 2013. 3, 6

[11] O. Freifeld, A. Weiss, S. Zuffi, and M. Black. Contour people: A parametrized model of 2D articulated human shape. *CVPR*, pp. 639–646, 2010. 3

[12] P. Guan, O. Freifeld, and M. Black. A 2D human body model dressed in eigen clothing. *ECCV*, pp. I:285–298, 2010. 3

[13] Š. Ivekovič, E. Trucco, and Y. Petillot. Human body pose estimation with particle swarm optimisation. *Evol. Comput.*, 16(4):509–528, 2008. 2, 5

[14] V. John, E. Trucco, and Š. Ivekovič. Markerless human articulated tracking using hierarchical particle swarm optimisation. *Image Vision Comput*, 28(11):1530 – 1547, 2010. 5

[15] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *IEEE Face and Gesture Recog.*, pp. 38–44, 1996. 2, 3

[16] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. *BMVC*, pp. 1–11, 2011. 6

[17] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *CVPR*, pp. 1:271–278, 2005. 2

[18] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65–81, 2007. 2, 3

[19] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. *CVPR*, pp. 422–429, 2010. 2

[20] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. *CVPR*, pp. 1281–1288, 2011. 2, 3, 5, 6, 7

[21] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *Int. J. Robot. Res.*, 22(6):371–391, 2003. 3

[22] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. *PAMI*, 34(9):1744–1757, 2012. 6

[23] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. *CVPR*, pp. 1385–1392, 2011. 2, 6, 7

[24] S. Zuffi, O. Freifeld, and M. Black. From pictorial structures to deformable structures. *CVPR*, pp. 3546–3553, 2012. 1, 2, 3, 4