Puppet Flow

Silvia Zuffi
Department of Computer Science
Brown University, Providence, RI 02912, USA
ITC - CNR, Milan, Italy
zuffi@cs.brown.edu

Michael J. Black
Department of Perceiving Systems
Max Planck Institute for Intelligent Systems
72076 Tübingen, Germany
black@tuebingen.mpg.de

Abstract

We introduce Puppet Flow (PF), a layered model describing the optical flow of a person in a video sequence. We consider video frames composed by two layers: a foreground layer corresponding to a person, and background. We model the background as an affine flow field. The foreground layer, being a moving person, requires reasoning about the articulated nature of the human body. We thus represent the foreground layer with the Deformable Structures model (DS) [2], a parametrized 2D part-based human body representation. We call the motion field defined through articulated motion and deformation of the DS model, a Puppet Flow. By exploiting the DS representation, Puppet Flow is a parametrized optical flow field, where parameters are the person's pose, gender and body shape.

1. Introduction

We describe a procedure to synthesize an optical flow image describing the motion of a human body. We exploit a part-based human shape model to define a layered model of human body optical flow that we name Puppet Flow.

We assume the optical flow between two adjacent frames in a video sequence o a moving person can be represented with two layers, background and foreground. Given two frames, a person flow can be computed based only on the articulated motion and contour deformations of the person; given the person layer, the background pixels can be isolated to compute a simple affine background motion model. The composition of the person flow and the background flow is our layered representation of image motion. We call the person flow Puppet Flow. To represent the person layer we use the Deformable Structures model [2]. Figure 3 (*left, center*) shows a synthesized example where we have applied an affine motion to an image of a real scene, simulating camera motion. Then we have used two instances of the DS model to generate a foreground layer. On the right we show the corresponding flow field that is generated with our model. Note that for the foreground layer the dense flow field is not computed from image data, but only from the articulated motion and contour deformation of the DS model.

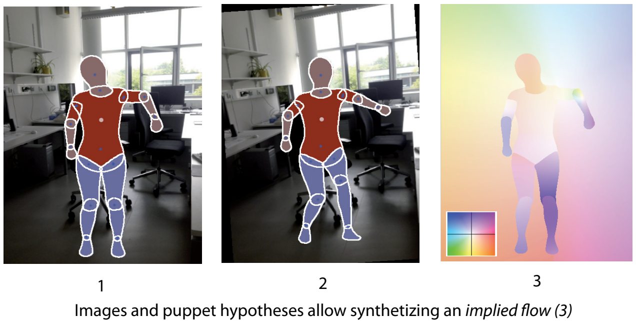
1
2
3
Images and puppet hypotheses allow synthetizing an *implied flow (3)*

Figure 1. **Image layers**. *left, center* Two frames with a synthesized foreground layer generated from the DS model; *right* The corresponding dense optical flow as computed from our model, where the background is an estimated affine motion and the foreground is computed only from the DS model without using image information.

2. Model

Puppet Flow is based on the DS model representation of the human body. DS model is a contour body model parametrized by pose and shape. It is also a part-based model, where each part is represented by a closed contour, defining part sub-layers for the foreground, that can then be defined by the union of the body parts layers (Fig. 2).

Contour points of body parts are generated in local coordinate systems by linear models in 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 the matrix of PCA basis components and $\mathbf{m}_i$ is the part mean. The PCA model is learned from 2D projections of a realistic, gender-specific, 3D human body model. The correlation between the shape coefficients $\mathbf{z}_i$ and body pose is modeled with pairwise Multivariate Gaussian distributions over the relative pose and shape parameters of connected body parts. Note that for the DS model the shape coefficients describe therefore pose-dependent shape deformations, they do not model variability of intrinsic body shape (tall, short, fat, slim) as the model is learned from a single 3D body model. Part contour points from 1 expressed in local coordinates are then converted in global coordinates given each part position and orientation in global frame, $\mathbf{c}_i, \theta_i$. More details on the DS model can be found in [2].

Given two instances of the DS model corresponding to foreground layers in adjacent frames, we define a per-pixel body motion by the warping transformation that maps the points of the first layer into the second layer. This warping function is estimated from corresponding contour points, and as for the body layer, is defined per part.

Let $\mathbf{x}_t$ be a vector of body model variables, that is $\mathbf{z}_i, \mathbf{c}_i, \theta_i$ for each body part, and puppet scale. Define $U_{t,t+1}$ the flow field as:

$$U_{t,t+1}(\mathbf{x}_t, \mathbf{x}_{t+1}) = M(\mathbf{x}_t) \circ W^{fg}_{t,t+1}(\mathbf{x}_t, \mathbf{x}_{t+1}) + \bar{M}(\mathbf{x}_t) \circ W^{bg}_{t,t+1}(\mathbf{x}_t, \mathbf{x}_{t+1}) \tag{2}$$

where $M(\mathbf{x}_t)$ is the mask for the foreground layer, that is a 2-channels image (for the horizontal and vertical dimensions of the optical flow) with value 1 for the foreground pixels and value 0 for the background pixels. The matrices $W^{fg}_{t,t+1}$ and $W^{bg}_{t,t+1}$ define per-pixel foreground and background flow, respectively.

For a DS model of R parts, let $\mathbf{x}_{t,r}$ be the DS model variables for the part of index r plus the global scale. Define $D(\mathbf{x}_t)$ as the body layers mask, that is a 2-channels image where each element assumes an index of the body part, namely the part closest to the camera (Figure 2 (2)). The foreground flow is then defined as:

$$W^{fg}_{t,t+1}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \sum_{r=1}^{R} \mathbb{1}_r(D(\mathbf{x}_t)) \circ W^{r}_{t,t+1}(\mathbf{x}_{t,r}, \mathbf{x}_{t+1,r}) \tag{3}$$

where $W^{r}_{t,t+1}(\mathbf{x}_{t,r}, \mathbf{x}_{t+1,r})$ is a part-specific synthetic flow.

The part-specific synthetic flows $W^{r}_{t,t+1}(\mathbf{x}_{t,r}, \mathbf{x}_{t+1,r})$ are given by a warping function over pixels. The warping function is a mapping between pixels at two different time steps; it is estimated by the deformation of contour points of the part r at time t into the corresponding points at time $t+1$. This can be accomplished in different ways, for example considering a polynomial mapping between the contour points, or using Thin Plate Splines (TPS). We use TPS that we compute with the method described in [1].

Consider two sets of points: a set of N landmarks on the first frame $\mathbf{p}^{(1)}$, and a set of corresponding N points on the second frame $\mathbf{p}^{(2)}$. These sets are subsets of the points along the contour of a body part r. We compute the mapping from $\mathbf{p}^{(2)}$ to $\mathbf{p}^{(1)}$, and then apply the mapping to a set of grid points to propagate the estimated mapping between landmarks to points inside the body part.

Let

$$U(r) = r^2 log(r^2) \tag{4}$$

and compute the $N \times N$ matrix K with elements

$$K_{ij} = U(\|\mathbf{p}^{(2)}_i - \mathbf{p}^{(2)}_j\|). \tag{5}$$

Define

$$P = (1_N \; \mathbf{x}^{(2)} \; \mathbf{y}^{(2)}) \tag{6}$$

a $N \times 3$ matrix, where $\mathbf{x}^{(2)}$ and $\mathbf{y}^{(2)}$ are column vectors with the coordinates of the destination landmarks. Define

$$L = \begin{pmatrix} K & P \\ P^T & 0 \end{pmatrix} \tag{7}$$

a $(N+3) \times (N+3)$ matrix. Also define

$$Y = \begin{pmatrix} \mathbf{x}^{(1)} & \mathbf{y}^{(1)} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{8}$$

a $(N+3) \times 2$ matrix, with $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ being column vectors with the coordinates of the source landmarks. Then compute

$$Q_a = L^{-1}Y \tag{9}$$

where the $(N+3) \times 2$ matrix Q_a, following [1], has the form $Q_a = (Q \mid \mathbf{a}_1 \; \mathbf{a}_x \; \mathbf{a}_y)$. Now consider a set of M points $\mathbf{p}$ on a grid. Define

$$\hat{K}_{ij} = U(\|\mathbf{p}_i - \mathbf{p}^{(2)}_j\|) \tag{10}$$

$$\hat{P} = (1_M \; \mathbf{x} \; \mathbf{y}) \tag{11}$$

$$\hat{L} = \begin{pmatrix} \hat{K} \\ \hat{P}^T \end{pmatrix} \tag{12}$$

The estimated warped grid point locations on the first frame are then given by

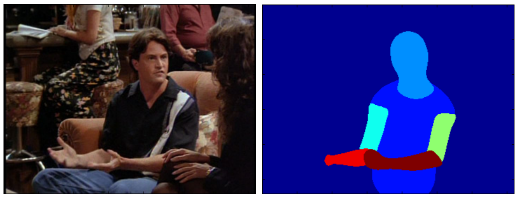

Figure 2. **DS puppet layer**. (1) Frame; (2) Corresponding puppet layer with parts ordered by fixed order. The warmer the color, the closer to the camera.

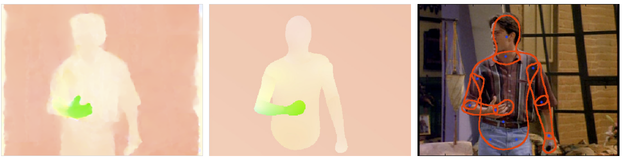

Figure 3. **Example**. (1) Dense flow; (2) Pose-parametrized flow; (3) Foreground layer by the DS puppet on the first frame.

$$\mathbf{f} = \hat{L}^T Q_a \tag{13}$$

which corresponds to computing the following expression

$$f(x,y,:) = a_{1,:} + a_{x,:}x + a_{y,:}y + \sum_{1..N} q_{i,:} U(\|(x,y) - \mathbf{p}_j^{(2)}\|). \tag{14}$$

The flow from the first frame to the second frame for the body part r is computed as

$$W^r_{t,t+1}(\mathbf{x}_{t,r}, \mathbf{x}_{t+1,r}) = \mathbf{p} - \mathbf{f}. \tag{15}$$

The background motion $W^{bg}_{t,t+1}(\mathbf{x}_t, \mathbf{x}_{t+1})$ can be expressed as an affine motion.

The composition through masking of the warping functions is the pose-parametrized Puppet Flow. Figure 3 (1) shows the dense flow computed between two frames of a person moving his left lower arm toward his left. (2) shows a corresponding optical flow field computed using Equation 2, where the background has been generated fitting an affine motion model to the background layer and the foreground is the Puppet Flow generated from the DS model, shown on one of the two frames in (3).

References

[1] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989. 2

[2] S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *CVPR*, pages 3546–3553, 2012. 1, 2