Supplementary Document

MoSculp: Interactive Visualization of Shape and Time

HIDDEN MARKOV MODEL IN KEYPOINT DETECTION

As mentioned in the paper [3], we enforce temporal smoothness in the 2D keypoint detections across all frames using a Hidden Markov Model (HMM).

Specifically, given T video frames $\{x^1, \ldots, x^T\}$, we compute the maximum marginal likelihood estimate of joint i's pixel location at time t, denoted by y_i^t, as

$$\begin{aligned}
&\arg\max_{y_i^t} p\left(y_i^t \mid x^1, \ldots, x^t\right) \propto p\left(y_i^t, x^1, \ldots, x^t\right) && (1)\\
&= \int_{y_i^{t-1}} p\left(y_i^{t-1}, y_i^t, x^1, \ldots, x^t\right) \mathrm{d}y_i^{t-1} && (2)\\
&= \int_{y_i^{t-1}} p\left(y_i^t, x^t \mid y_i^{t-1}, x^1, \ldots, x^{t-1}\right) && (3)\\
&\quad p\left(y_i^{t-1}, x^1, \ldots, x^{t-1}\right) \mathrm{d}y_i^{t-1} && (4)\\
&= \int_{y_i^{t-1}} p\left(y_i^t, x^t \mid y_i^{t-1}\right) && (5)\\
&\quad p\left(y_i^{t-1}, x^1, \ldots, x^{t-1}\right) \mathrm{d}y_i^{t-1} && (6)\\
&= \int_{y_i^{t-1}} p\left(y_i^t \mid y_i^{t-1}\right) p\left(x^t \mid y_i^t\right) && (7)\\
&\quad p\left(y_i^{t-1}, x^1, \ldots, x^{t-1}\right) \mathrm{d}y_i^{t-1}, && (8)
\end{aligned}$$

where the transition probability $p\left(y_i^t \mid y_i^{t-1}\right)$ is a bivariate Gaussian centered at y_i^{t-1} with a standard deviation of three pixels, and the emission probability $p\left(x^t \mid y_i^t\right)$ is approximated as the heatmap predicted for joint i, independently at frame t. Strictly speaking, this heatmap resembles more the posterior $p\left(y_i^t \mid x^t\right)$, *i.e.*, how likely joint i lies at each pixel location given frame t. Considering the simple Bayes' rule

$$p\left(x^t \mid y_i^t\right) = \frac{p\left(y_i^t \mid x^t\right) p\left(x^t\right)}{p\left(y_i^t\right)},$$

by using the heatmaps as $p\left(x^t \mid y_i^t\right)$, we essentially assume a uniform $p\left(y_i^t\right)$, the prior distribution on joint i's location.

Notice how $p\left(y_i^t, x^1, \ldots, x^t\right)$ in Equation 1 gets expressed in terms of $p\left(y_i^{t-1}, x^1, \ldots, x^{t-1}\right)$ in Equation 8. This recurrent structure allows us to solve the HMM efficiently by message passing. More specifically, to find the y_i^t that maximizes $p\left(y_i^t, x^1, \ldots, x^t\right)$, we integrate (spatially) over frame t the product of the Gaussian distribution for smoothness, the predicted heatmap for this joint in frame t, and where this joint is believed to be in frame $t-1$, *i.e.*, $p\left(y_i^{t-1}, x^1, \ldots, x^{t-1}\right)$.

In cases where some of the person's joints are not detected locally, we linearly interpolate their locations from neighboring frames before running the HMM.

JOINT OPTIMIZATION OF SHAPE, POSE, AND GLOBAL TRAJECTORY OVER TIME

Here, we provide the mathematical and implementation details of our formulation.

$$\begin{aligned}
\mathcal{L}\left(\{T^t\}, \{\theta^t\}, \beta\right) &= \sum_{t=1}^{N} \mathcal{L}_{\text{data}}\left(T^t, \theta^t, \beta\right) + \alpha_1 \mathcal{L}_{\text{prior}}\left(\theta^t, \beta\right)\\
&+ \alpha_2 \sum_{t=1}^{N-1} \mathcal{L}_{\text{temporal}}\left(T^t, T^{t+1}, \theta^t, \theta^{t+1}, \beta\right). \qquad (9)
\end{aligned}$$

As mentioned in the paper, this formulation can be seen as an extension of SMPLify [1], a single-image 3D human pose and shape estimation algorithm, to videos. Therefore, we will elaborate only on the newly added term—the temporal smoothness prior $\mathcal{L}_{\text{temporal}}$.

This term encourages the 3D model to be temporally smooth; it penalizes changes in the human's global translations (Equation 11), local vertex locations (Equation 12), and pose parameters (Equation 13). More specifically,

$$\begin{aligned}
\mathcal{L}_{\text{temporal}} &= \lambda_1 \mathcal{L}_{\text{global}} + \lambda_2 \mathcal{L}_{\text{local}} + \lambda_3 \mathcal{L}_{\text{rotation}} && (10)\\
\mathcal{L}_{\text{global}}\left(T^t, T^{t+1}\right) &= \left\|T^t - T^{t+1}\right\|_2^2 && (11)\\
\mathcal{L}_{\text{local}}\left(\theta^t, \theta^{t+1}, \beta\right) &= \left\|V\left(\theta^t, \beta\right) - V\left(\theta^{t+1}, \beta\right)\right\|_F^2 && (12)\\
\mathcal{L}_{\text{rotation}}\left(\theta^t, \theta^{t+1}\right) &= \left\|\begin{bmatrix}\cos\left(\theta^t\right) - \cos\left(\theta^{t+1}\right)\\ \sin\left(\theta^t\right) - \sin\left(\theta^{t+1}\right)\end{bmatrix}\right\|_2^2, && (13)
\end{aligned}$$

where $V(\cdot)$ are the vertices' local 3D coordinates given the pose and shape, and the λ's are constant weights that roughly match the orders of magnitude of the three losses.

Intuitively, $\mathcal{L}_{\text{global}}$ requires the body's global trajectory to be smooth; $\mathcal{L}_{\text{local}}$ further requires vertices of the human mesh to translate smoothly; and $\mathcal{L}_{\text{rotation}}$ imposes additional rotational

UIST '18 October 14–17, 2018, Berlin, Germany

ACM ISBN 978-1-4503-5948-1/18/10.

DOI: https://doi.org/10.1145/3242587.3242592

smoothness in the parameter space, which is necessary for producing natural pose evolution.

Optimization

We adopt a two-stage optimization procedure, in which we first ignore the temporal loss, optimizing only the per-frame loss: $\mathcal{L}_{\text{data}}+\alpha_1\mathcal{L}_{\text{spatial}}$. This runs in a fully parallelized fashion and provides good initializations to, thereby speeding up, the subsequent joint optimization.

More importantly, this two-step procedure allows us to address effectively the "pose-flipping problem": tendency of the joint optimization getting stuck with flipped facing directions when the person is captured in a side view throughout, often due to the inherent ambiguity (see *Run, Forrest, Run!* for an example). To avoid such local minima, the algorithm tries both directions that the human model could face, when the left-right shoulder or hip joints are less than 100 pixels apart (a heuristic), and then initializes the joint optimization with the pose that gives a lower $\mathcal{L}_{\text{prior}}$. We minimize this loss using a non-linear least squares approach [2].

REFERENCES

1. Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision*.
2. Matthew M. Loper and Michael J. Black. 2014. OpenDR: An Approximate Differentiable Renderer. In *European Conference on Computer Vision*.
3. Xiuming Zhang, Tali Dekel, Tianfan Xue, Andrew Owens, Qiurui He, Jiajun Wu, Stefanie Mueller, and William T. Freeman. 2018. MoSculp: Interactive Visualization of Shape and Time. In *ACM Symposium on User Interface Software and Technology*.