Supplementary Material for
Unsupervised 3D Pose Estimation for Hierarchical Dance Video Recognition

A. UID Dataset

The proposed *UID video dataset* can be found at https://drive.google.com/drive/folders/1-SdWYxIorbhQzi9Bp_HpJf25_ieMjoh5?usp=sharing. The dataset folder contains 9 sub-folders, each having ∼30 to ∼300 videos, showing one of the following 9 dance types: Ballet, Belly dance, Flamenco, Hip Hop, Rumba, Swing dance, Tango, Tap dance and Waltz.

B. Demo Videos

In Figure 6 in Section 3 of the paper, we have shown estimated 3D poses by drawing skeletal figures in the estimated poses and overlaying them on the corresponding images of the dancers. These single frame overlays help us verify the placement of the skeleton within the body parts in only four frames in a video (containing ∼100 to ∼800 frames). However, the contributions of our paper also include enforcement of temporal smoothness constraints, and estimation of complex 3D poses. Here we therefore include videos that show the overlays of the skeletons in all frames of the videos. Viewing these videos shows the temporal smoothness of pose estimates as well as their continuous alignment with the dancer's poses achieved by our method, which cannot be seen from the static depictions in the paper. Further, the pose and alignment quality can now be seen for the entire range of complexities associated with the poses assumed by the dancer throughout the video instead of with the selected few frames in the paper.

The videos we use to show our results are selected from the test set in the *UID dataset*. The selected videos can be found at https://drive.google.com/drive/folders/1X5K2U1Eq1QlcU8GmM_gHFVv75VkBzcoV?usp=sharing. The right side of each video shows videos of skeletons representing 2D projections (2D poses) $\hat{p}_t$ of the estimated 3D poses $\hat{P}_t$ by themselves. These 2D poses are estimated using the estimated 3D-to-2D projection parameter ω^{*2D}. To bring out the poses, they are shown from a closer and different viewpoint than used to capture the original video. On the left side, we show the same 2D poses, using the viewpoint used to capture the original video, and overlaid on the original video frames. We can see that the skeletons align well with the complex movements of the dancers, such as spinning, Pointe (fully extended feet), and Tour Jeté (a high turning leap). Also, we can see that transitions between poses in adjacent frames are smooth, e.g., without large, abnormal displacements between the locations of the same joint in successive frames.

Finally, names of the recognized 3D movements $\hat{y}_t^e$ of each body part $e \in E$ are shown at the bottom of the video. The recognized movements $\hat{y}_t^e$ can be seen to well match the dancers' movements in the original video.

C. Algorithms in Detail

Algorithms 1*, 2*, 3* and 4* here are the detailed versions of Algorithms 1, 2, 3 and 4. The movement and dance genre recognition is described in detail in Algorithm 5 and 6.

D. Implementation Details

Table 7 and 8 show the the values of the DH parameters, and the bounds of the joint rotation offset angles and bone length of our 34-DOF digital dancer model.

Algorithm 1*: Object Tracking

Input: a sequence of video frames $\{I_t\}_{t=0}^{T-1}$
Output: a sequence of bounding boxes $\{(x_t^i, y_t^i, w_t^i, l_t^i)\}_{t=0}^{T-1}$ of the i^{th} dancer
Initialization: select the bounding box $(x_0^i, y_0^i, w_0^i, l_0^i)$ of N dancers to track by mouth
while *new frame I_t available* **do**
  for i^{th} *dancer* **do**
    Obtain $(x_t^i, y_t^i, w_t^i, l_t^i)$ by LDES approach
    if *not overlap with others* **then**
      $\tilde{h} \leftarrow h_t^i$ // Store histogram of i^{th} dancer
      $\tilde{v} \leftarrow v_t^i$ // Store velocity of i^{th} dancer
    end
    if *overlap happens & tracking fails* **then**
      Estimate when overlap ends
    end
    if *overlap ends* **then**
      // Relocate the bounding box
      $\hat{k} = \underset{k}{\operatorname{argmax}}(correlation(\tilde{h}, h^k))$ where h^k is the histogram of the k^{th} patch along the moving direction in the cone searching region
      $(x_t^i, y_t^i, w_t^i, l_t^i) \leftarrow$ location of $\hat{k}^{th}$ patch
    end
  end
end

Algorithm 2*: Tracking Based 2D Pose Estimation

Input: a sequence of video frames $\{I_t\}_{t=0}^{T-1}$ and a sequence of bounding boxes $\{B_t^i\}_{t=0}^{T-1} = \{(x_t^i, y_t^i, w_t^i, l_t^i)\}_{t=0}^{T-1}$ of the i^{th} dancer
Output: a sequence of poses $\{\hat{p}_t^i\}_{t=0}^{T-1}$ of the i^{th} dancer
while *new frame I_t available* **do**
  Estimate poses // Perform OpenPose
  for i^{th} *dancer* **do**
    Select C poses $\{p_t^{i,c}\}_{c=0}^{C-1}$ overlapped with the bounding box B_t^i
    $\hat{c} = \underset{c}{\operatorname{argmax}}(correlation(h_t^{i,c}, h_{t-1}^i))$ where $h_t^{i,c}$ is the histogram of the pose $p_t^{i,c}$
    $\hat{p}_t^i \leftarrow p_t^{i,\hat{c}}$
  end
end

Algorithm 3*: 3D Pose Initialization

Input: a sequence of 2D poses $\{p_t\}_{t=0}^{N-1}$ of a dancer
Output: a sequence of 3D poses $\{\tilde{P}_t\}_{t=0}^{N-1}$ of the dancer
Set the temporal window size to be 2Δ
Denote total number of segments as $s = \lfloor \frac{N}{2\Delta} \rfloor$
for $t = \Delta$ *to* $N - \Delta$ **do**
  for k = *0 to* $K - 1$ **do**
    Try new seed for DH parameters $\Lambda^k = \{\boldsymbol{\Theta}^k, \boldsymbol{d}^k, \boldsymbol{a}^k, \boldsymbol{\alpha}^k\}$ and perspective projection parameters $\omega^k = \{f^k, c^k\}$
    for $i = t - \Delta$ *to* $t + \Delta$ **do**
      Generate 3D pose $\hat{P}_i^k = G(\Lambda^k)$
      Estimate 2D pose $\hat{p}_i^k = \Psi(\hat{P}_i^k; \omega^k)$
      Compute error $e_i^k = ||\hat{p}_i^k - p_i||_2^2$
      Optimize $\Lambda^{*k}, \omega^{*k} = \underset{\Lambda^k, \omega^k}{\operatorname{argmin}} e_i^k$
      Assign $\hat{P}_i^{*k} = G(\Lambda^{*k})$
      Update $\Lambda^k \leftarrow \Lambda^{*k}$
      Update $\omega^k \leftarrow \frac{1}{i-t+\Delta} \sum_{l=t-\Delta}^{t} \omega_l^{*k}$
    end
  end
end
Select the seed $k^* = \underset{\tilde{k}}{\operatorname{argmin}} \sum_{i=t-\Delta}^{t+\Delta} e_i^{\tilde{k}}$
Assign $\tilde{P}_t, \omega_t^{2D} \leftarrow \hat{P}_t^{*k^*}, \omega^{k^*}$

Algorithm 4*: 3D Pose Estimation

Input: a sequence of video frames $\{I_t\}_{t=0}^{T-1}$, 2D poses $\{p_t\}_{t=0}^{T-1}$ and initial 3D poses $\{\tilde{P}_t\}_{t=0}^{T-1}$ of a dancer
Output: a sequence of estimated 3D poses $\{\hat{P}_t\}_{t=0}^{T-1}$ of the dancer
while *new frame* I_t *available* **do**
  Estimate 3D pose $\hat{P}_t = \Phi(p_t; \omega^{3D})$
  Project to 2D pose $\hat{p}_t = \Psi(\hat{P}_t; \omega^{2D})$
  Compute loss $L = \alpha(||\hat{p}_t - \hat{p}_{t-1}||_2^2 + \beta||\hat{P}_t - \hat{P}_{t-1}||_2^2) + ||\hat{p}_t - p_t||_2^2 + ||\hat{P}_t - \tilde{P}_t||_2^2$
  Update $\omega^{2D} \leftarrow \omega^{2D} - \eta \frac{\partial L}{\partial \omega^{2D}}$
  Update $\omega^{3D} \leftarrow \omega^{3D} - \eta \frac{\partial L}{\partial \omega^{3D}}$
end

Algorithm 5: Movement Identification

Input: a sequence of poses $\{\bar{p}_t^e = \{(\bar{x}_t^j, \bar{y}_t^j)\}_{j=0}^{|J_e|}\}_{t=0}^{T-1}$ where T denotes the total number of frames and J_e denotes the set of body joints connected to the body part e; a sequence of corresponding movement labels $\{\tilde{y}_t^e\}_{t=0}^{T-1}$ of the body part e
Output: predicted movement labels $\{\hat{y}_t^e\}_{t=0}^{T-1}$
for *epoch* $= 0$ ***to*** $N - 1$ **do**
  $\{\hat{y}_t^e\}_{t=0}^{T-1} = \text{LSTM}(\{\bar{p}_t^e\}_{t=0}^{T-1})$
  $L = \text{BCELoss}(\{\hat{y}_t^e\}_{t=0}^{T-1}, \{\tilde{y}_t^e\}_{t=0}^{T-1})$
  Update LSTM until converge
end

Algorithm 6: Dance Classification

Input: a sequence of movement labels $\{\{\hat{y}_t^e\}_{e=0}^{|E|-1}\}_{t=0}^{T-1}$ of all the body parts $e \in E$; and the ground-truth dance genre label g of the sequence
Output: predicted dance genre label $\hat{g}$
for *epoch* $= 0$ ***to*** $N-1$ **do**
  $\hat{g} = \text{LSTM}(\{\{\hat{y}_t^e\}_{e=0}^{|E|-1}\}_{t=0}^{T-1})$
  $\tilde{L} = \text{CrossEntropyLoss}(\hat{g}, g)$
  Update LSTM until converge
end

Joint	Θ	d	a	α	**Joint**	Θ	d	a	α	**Joint**	Θ	d	a	α
0	$180+\theta_0$	0	0	90	12	$90+\theta_{11}$	$b_4 h$	0	90	24	$0+\theta_{23}$	0	0	90
1	$-90+\theta_0$	0	0	90	13	$90+\theta_{12}$	0	$0.6b_4 h$	0	25	$0+\theta_{24}$	$b_8 h$	0	-90
2	$90+\theta_1$	0	$b_0 h$	-90	14	$0+\theta_{13}$	0	0	-90	26	$-90+\theta_{25}$	0	$0.1b_8 h$	0
3	$0+\theta_2$	0	0	90	15	$-90+\theta_{14}$	0	0	90	27	$0+\theta_{26}$	0	0	-90
4	$90+\theta_3$	0	0	90	16	$90+\theta_{15}$	$-b_3 h$	0	90	28	$-90+\theta_{27}$	0	0	90
5	$90+\theta_4$	$b_1 h$	0	90	17	$0+\theta_{16}$	0	0	-90	29	$0+\theta_{28}$	$b_6 h$	0	-90
6	$90+\theta_5$	0	0	90	18	$90+\theta_{17}$	$-b_4 h$	0	90	30	$90+\theta_{29}$	0	$-b_7 h$	0
7	$90+\theta_6$	0	$b_1 h$	0	19	$-90+\theta_{18}$	0	$-0.6b_4 h$	0	31	$0+\theta_{30}$	0	0	90
8	$0+\theta_7$	0	0	90	20	$0+\theta_{19}$	0	0	-90	32	$0+\theta_{31}$	$-b_8 h$	0	-90
9	$90+\theta_8$	0	0	90	21	$-90+\theta_{20}$	0	0	90	33	$-90+\theta_{32}$	0	$-0.1b_8 h$	0
10	$90+\theta_9$	$b_3 h$	0	90	22	$0+\theta_{21}$	$-b_6 h$	0	-90					
11	$0+\theta_{10}$	0	0	-90	23	$90+\theta_{22}$	0	$b_7 h$	0					

Table 7. The DH parameters $\Lambda = \{\boldsymbol{\Theta}, \boldsymbol{d}, \boldsymbol{a}, \boldsymbol{\alpha}\}$ for the 34-DOF human model as shown in Figure 3. The joint rotation angle $\boldsymbol{\Theta}$ along z axis, the distance $\boldsymbol{d}$ along z axis, and the offset distance $\boldsymbol{a}$ along x axis are determined by the joint rotation offsets $\boldsymbol{\theta} = (\theta_0, ..., \theta_{32})$ and bone lengths $\boldsymbol{b} = (b_0, ..., b_6)$, where their bounds are defined in Table 8.

Rotation	θ_1	θ_2	θ_3	θ_7	θ_8	θ_9	θ_{10}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{19}	θ_{20}	θ_{21}	θ_{22}	θ_{23}	θ_{26}	θ_{27}	θ_{28}	θ_{29}	θ_{30}
min	$-\frac{\pi}{8}$	$-\frac{\pi}{4}$	$-\frac{\pi}{4}$	$-\frac{\pi}{1.6}$	$-\frac{\pi}{4}$	$-\pi$	$-\frac{\pi}{2}$	$-\frac{\pi}{1.6}$	$-\frac{\pi}{4}$	$-\pi$	$-\frac{\pi}{2}$	$-\frac{\pi}{2}$	$-\pi$	$-\frac{\pi}{2}$	0	$-\frac{\pi}{4}$	$-\frac{\pi}{2}$	$-\pi$	$-\frac{\pi}{2}$	0	$-\frac{\pi}{4}$
max	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$	$\frac{\pi}{1.6}$	$\frac{\pi}{1.6}$	0	0	$\frac{\pi}{1.6}$	$\frac{\pi}{1.6}$	0	0	$\frac{\pi}{4}$	$\frac{\pi}{1.3}$	$\frac{\pi}{2}$	$\frac{\pi}{1.3}$	$\frac{\pi}{2}$	$\frac{\pi}{4}$	$\frac{\pi}{1.3}$	$\frac{\pi}{2}$	$\frac{\pi}{1.3}$	$\frac{\pi}{2}$

Bone	neck b_0	head b_1	shoulder b_2	uparm b_3	lowarm b_4	hip b_5	upleg b_6	lowleg b_7	toe b_8
Average	0.25	0.08	0.06	0.17	0.17	0.04	0.21	0.21	0.04
Std	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Table 8. The bounds of the joint rotation offset angles $\boldsymbol{\theta} = (\theta_0, ..., \theta_{32})$ and bone length ratios $\boldsymbol{b} = (b_0, ..., b_6)$ defined for our digital dancer model.