Persistent-Transient Duality in Human Behavior Modeling

Hung Tran, Vuong Le, Svetha Venkatesh, Truyen Tran
Applied AI Institute, Deakin University, Geelong, Australia
{tduy,vuong.le,svetha.venkatesh,truyen.tran}@deakin.edu.au

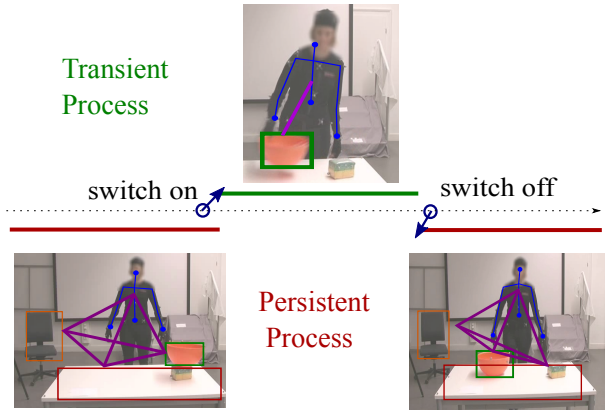


Figure 1. We model the *Persistent-transient duality* of human behaviors. In the default *Persistent process* (lower row), the human subjects consider the whole scene from the global view with dense relational connections (purple links). For an interactive action, they switch on a *Transient process* (upper row) that zooms into a small set of local objects that are directly interacted. When done with such task, they switch back to the default Persistent mode.

Abstract

We propose to model the persistent-transient duality in human behavior using a parent-child multi-channel neural network, which features a parent persistent channel that manages the global dynamics and children transient channels that are initiated and terminated on-demand to handle detailed interactive actions. The short-lived transient sessions are managed by a proposed Transient Switch. The neural framework is trained to discover the structure of the duality automatically. Our model shows superior performances in human-object interaction motion prediction.

1. Introduction

Human behavior is highly contextualized, reacting to the rapid changes in the situation. This fast-changing nature requires a model to adapt quickly in structure, representation, and inference mechanism to follow the true patterns of the behavior. Such requirement is critical in the problem of human-object interaction (HOI) motion prediction, where the subjects follow an overall plan but occasionally deviate from it to solve emerging tasks [1].

Motion models on this task reflect the changing situations by gradually adapting their relational structure. However, with a fixed inference mechanism, they cannot account for the discrete switching between distinctive mechanisms and fail to keep up with the movement patterns. For example, when the human subject deviates from the overall path to interact with an object, these models will continue to consider the interacted object as an equal member of the scene and miss its importance as the action's direct recipient.

We address this limitation by factorizing the human behaviors into two processes: a *slow-changing persistent process*, which maintains a continuous default dynamic, and a *fast-changing transient process*, which has an adaptive lifecycle and a personalized structure that reflects the human's perspective in emerging events (See Fig. 1). The two processes are modeled into a parent-child multi-channel neural network called *Persistent-Transient Duality*. The *Persistent channel* is a recurrent relational network operating on the global scene spatially and throughout the session temporally. The *Transient channels* instead have a contextualized graphical structure constructed on the spot whenever the human subjects shift the priority toward interacting with other entities. The life cycles of these channels are managed by a neural *Transient Switch*, which can learn to anticipate when a Transient channel will be needed and trigger it in time.

Our model establishes the SOTA performance on the HOI subset of the KIT Whole-Body Human Motion Database.

2. Method

2.1. Preliminaries

We consider the problem of modeling the sequential behaviors of N entities in a dynamic system, where each entity i is represented by a class label c_i and sequential features $X_i = \{x_i^t\}_{t=1}^{T}$. After observing T steps, we predict the features in the next L steps, $Y_i = \{y_i^t\}_{t=T+1}^{T+L+1}$. The entity classes include the human and object (c_i="human"/"object"), which decide the entity's feature spaces and behaviors.

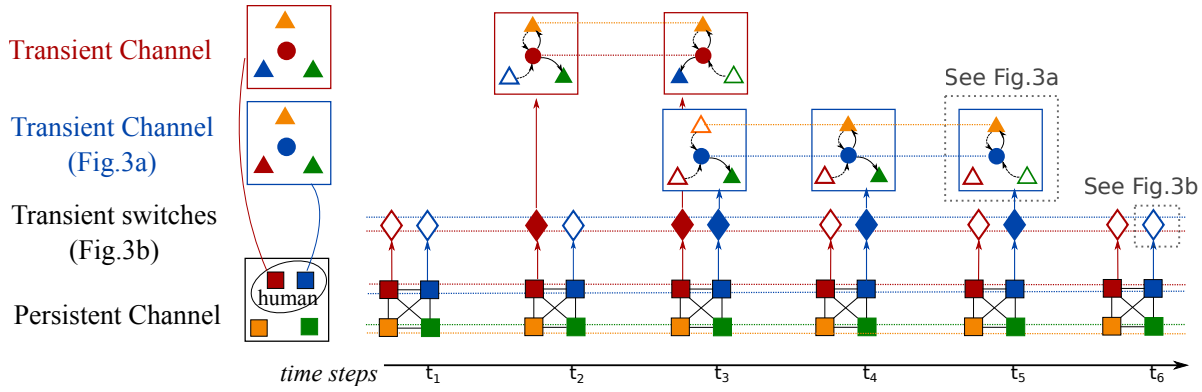


Figure 2. The architecture of *Persistent-transient Duality* Networks (PTD). The *persistent channel* contains fully-connected recurrent graph networks of all entities (squares) similar to current SoTA works. We introduce the new *Transient channel* to zoom into the local context of each human (circles) when they interact with surrounding entities (triangles). The transient channels are initialized and terminated on-demand, controlled by the neural *Transient Switches* (diamonds).

In this paper, we use a customized attention function defined over the query $q \in R^d$ and the identical key/value pairs $V = \{v_j\}_{j=1}^N \in R^{d \times N}$:

$$\text{Attn}\left(q, V\right) = \sigma\left(\sum_{j=1}^{N} \text{softmax}_j\left(\boldsymbol{W}_\alpha^T\left[\boldsymbol{W}_q q; \boldsymbol{W}_v v_j\right]\right)\boldsymbol{W}_v v_j\right)$$

where $[\cdot;\cdot]$ is the concatenation, σ is a non-linear activation function, $\boldsymbol{W}_v$, $\boldsymbol{W}_q$, and $\boldsymbol{W}_\alpha$ are learnable weights.

2.2. The Persistent-Transient Duality

We model the persistent-transient duality of human behavior by a hierarchical neural network called *Persistent-Transient Duality Networks (PTD)* (See Fig. 2). The network has three main components: The Persistent Channel, the Transient Channels, and the Transient Switch.

2.3. Persistent Channel

The Persistent channel oversees the global view of the scene, including the humans and other entities. It has the form of a recurrent relational network where entities interact in the spatio-temporal space spanned by the video. The temporal evolution of each entity is modeled as:

$$h_i^{\mathcal{P},t} = \text{RNN}_{c_i}\left(\left[z_i^{\mathcal{P},t}, m_{i,\mathcal{T}\to\mathcal{P}}^t\right], h_i^{\mathcal{P},t-1}\right), \quad (1)$$

where RNN_{c_i} is a recurrent unit that corresponds to the class of the i^{th} entity, maintaining hidden states $h_i^{\mathcal{P},t-1}$ and consuming input vector $z_i^{\mathcal{P},t}$. This input is formed as $z_i^{\mathcal{P},t} = [x_i^t; m_i^t]$, where x_i^t is the entity's feature and m_i^t is the message aggregated from the entity's neighbor through attention, $m_i^t = \text{Attn}\left(u_i^t, \left\{u_j^t\right\}_{j\neq i}\right)$ with $u\_^t = \left[x\_^t; h\_^{\mathcal{P},t-1}\right]$.

The unit also uses an optional transient-persistent message $m_{i,\mathcal{T}\to\mathcal{P}}^t$ which is non-zero if a corresponding *transient process* (Sec. 2.4) is currently active.

The Persistent channel generates two outputs from its hidden state: the future prediction $\hat{y}_i^{\mathcal{P},t}$ and the message $m_{i,\mathcal{P}\to\mathcal{T}}^t$ to the Transient channel:

$$\hat{y}_i^{\mathcal{P},t} = \text{MLP}\left(h_i^{\mathcal{P},t}\right), \quad m_{i,\mathcal{P}\to\mathcal{T}}^t = \text{MLP}\left(h_i^{\mathcal{P},t}\right). \quad (2)$$

The channel's prediction output $\hat{y}_i^{\mathcal{P},t}$ are combined with those from the Transient channel as detailed in Sec. 2.6.

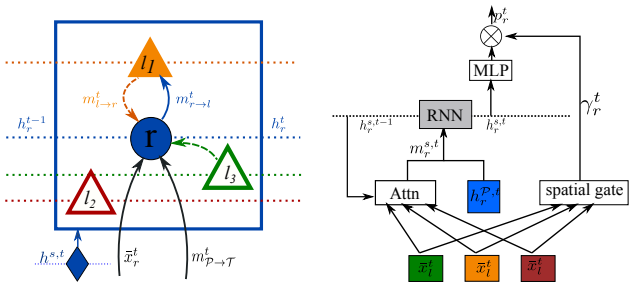
This persistent channel is equivalent in modeling with the major state-of-the-art HOI-M by using relational recurrent models [2]. Our key novelty is the consideration of the second side of the duality - the Transient process, presented in the next section.

2.4. Transient Channel

Within the persistent-transient duality, the Transient process allows the model to zoom in at relevant context and take the local viewpoint of the human entity when it starts to interact with objects. This human-specific process is implemented by a Transient channel. The egocentric property of this channel separates it from the global view of its parent persistent channel and reflects in three aspects of *feature representation*, *computational structure*, and *inference logic*.

The egocentric graph structure reflects the relations between the active subject and the surrounding passive entities. We define the *Transient graph* for a human entity of index i at time t to be $\mathcal{G}_i^t = (\mathcal{V}_i^t, \mathcal{E}_i^t)$. For a particular human, subscript i will be omitted for conciseness.

The egocentric characteristic of $\mathcal{G}^t$ reflects in its star-like structure: the nodes $\mathcal{V}^t$ includes a single *center node* r for the considering human, and *leaf nodes* of indices $\{l\}_{l\neq r}$ for other entities. The dynamic edges $\mathcal{E}^t$ connect the center with the leaves in two directions: *inward edges* $e_{l\to r}^t$ reflect which objects the human pay attention to, and the *outward*



(a) Transient Channel (b) Transient Switch

Figure 3. The Transient Channel (a) and Transient Switch (b)

edges $e^t_{r\to l}$ represents the objects are being manipulated by the human. These edges are determined at each time step by thresholding the center-leaf distances d^t_{lr}, making the graph's topology evolve within one single Transient session.

The Egocentric representation of the entities are computed by transforming the geometrical features into the egocentric coordinate system corresponding to the viewpoint of the human center node r, $\bar{x}\_^t = f_{\text{ego}}\left(x\_^t, x^t_r\right) = x\_^t - \text{centroid}(x^t_r)$. This change of system puts various patterns of the human's motion into the same aligned space, filters out the irrelevant global information, and facilitates efficient inference of the egocentric model.

The Egocentric inference is made by updating the RNN hidden state $h^t\_$ of each node in the transient graph structure (see Fig. 3a). In detail, for the *center node*, the inward messages from its leaves are aggregated into: $m^t_{l\to r} = \text{Attn}\left(\left[\bar{x}^t_r; h^{t-1}_r\right], \left\{\left[\bar{x}^t_l; h^{t-1}_l\right]\right\}_{e^t_{l\to r}\in\mathcal{E}^t}\right)$. It is then combined with the egocentric features $\bar{x}^t_r$ and the message from the persistent channel $m^t_{\mathcal{P}\to\mathcal{T}}$ (Eq. (2)) to update the RNN:

$$z^t_r = \left[\bar{x}^t_r; m^t_{l\to r}; m^t_{\mathcal{P}\to\mathcal{T}}\right], \quad h^t_r = \text{RNN}_r\left(z^t_r, h^{t-1}_r\right). \tag{3}$$

For *leaf nodes*, they only update their states if the center node interacts with them, indicated by the outward edge

$$z^t_l = \left[\bar{x}^t_l; m^t_{r\to l}\right], h^t_l = \begin{cases}\text{RNN}_l\left(z^t_l, h^{t-1}_l\right) & \text{if } e^t_{r\to l}\in\mathcal{E}^t \\ h^{t-1}_l & \text{otherwise}\end{cases}, \tag{4}$$

where $m^t_{r\to l}$ is the outward message from the center to its leaves, calculated from its hidden state through an MLP.

The updated hidden states are used generate the transient predictions $\hat{y}^{\mathcal{T},t}_i$ and the messages sent to the persistent process $m^t_{\mathcal{T}\to\mathcal{P}}$ (used in Eq. (1)):

$$\hat{y}^{\mathcal{T},t}\_ = f^{-1}_{\text{ego}}\left(\text{MLP}\left(h^t\_\right)\right), m^t_{\mathcal{T}\to\mathcal{P}} = \phi\left(\text{MLP}\left(h^t_r\right)\right), \tag{5}$$

where f^{-1}_{ego} is the function converting the egocentric back to global coordinates. The transient predictions $\hat{y}^{\mathcal{T},t}\_$ are combined with those from the Persistent process in Sec. 2.6.

2.5. Switching Transient Processes

The life cycles of the Transient processes are managed based on the situation of human's activity by a neural *Transient Switch* (See Fig. 3b).

The switch first considers the current persistent state and the surrounding environment of the center entity r to update its switch RNN $h^{s,t}_r$:

$$h^{s,t}_r = \text{RNN}_s\left(\left[h^{\mathcal{P},t}_r, m^{s,t}_r\right], h^{s,t-1}_r\right), \tag{6}$$

where $m^{s,t}_r = \text{Attn}\left(h^{s,t-1}_r, \{\bar{x}^t_l\}_{l\neq r}\right)$, and $\bar{x}^t_l = f_{\text{ego}}\left(x^t_l, x^t_r\right)$ defined in Sec. 2.4.

The RNN unit is important in maintaining the switch's sequential properties, making it a state-full machine that can handle patterns of on and off switchings and avoid spurious decisions caused by noises. The switch-on probability p^t_r is:

$$\hat{p}^t_r = \gamma^t_r \cdot \text{sigmoid}\left(W h^{s,t}_r\right), \tag{7}$$

where W is learnable weights. The discount factor $\gamma^t_r \in [0,1]$ responds to the distance from the subject to the nearest neighbor: $\gamma^t_r = \exp\left(-\beta \cdot \min\left\{\|d^t_{lr}\|_2\right\}_{l\neq r}\right)$, where d^t_{lr} are the center-leaf geometrical distances and β is a learnable decay rate. This factor acts as a disruptive shortcut gate that modulates the switching decision based on the spatial evidence of the interaction.

Finally, the binary switch decision $\hat{s}^t_r$ is decided by thresholding the switch score: the switch is on ($\hat{s}^t_r = 1$) when $\hat{p}^t_r >= 0.5$, and is off otherwise. When it changes from 0 to 1, a new Transient process is created at time t for person r. This transient process will keep running until the switch turns off, then the persistent process again becomes the single operator.

2.6. Future prediction

In PTD, future motions are predicted by unrolling the model into the future of L time steps. At each future time step t, the predictions from persistent channel $\hat{Y}^{\mathcal{P},t}$ (Eq. (2)) and those from Transient channel $\hat{Y}^{\mathcal{T},t}$ (Eq. (5)) are combined with the priority on the Transient predictions. For a human entity, if its Transient channel is activated, the Transient prediction will be chosen; otherwise, the Persistent prediction will be used. For an object entity, if it receives an active outward Transient edge, it will take that channel's prediction. If it receives multiple outward edges, it uses the prediction from the channel with the highest transient score $\hat{p}^t_r$. Otherwise, it uses the persistent prediction by default.

2.7. Model Training

The model is trained end-to-end with two losses: prediction loss and switch loss, $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{switch}}$.

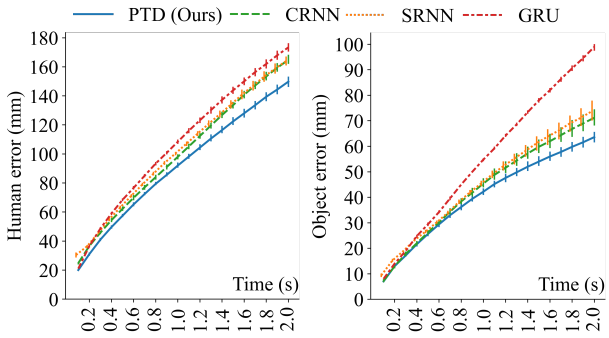


Figure 4. Quantitative performances and model size in HOI-M.

The **Prediction loss** $\mathcal{L}_{\text{pred}}$ measures the mismatch between predicted values $\hat{Y}$ and groundtruth Y, $\mathcal{L}_{\text{pred}} = ||\hat{Y}^{T:T+L} - Y^{T:T+L}||_2^2$.

The **Switch loss** $\mathcal{L}_{\text{switch}}$ is used to supervise the Transient Switch and is implemented as: $\mathcal{L}_{\text{switch}} = \text{BCE}\left(\hat{P}^{1:T+L}, P^{1:T+L}\right)$, where $\hat{P}^t$ are the switch scores (Eq. (7)) of all human entities, and $P^t$are binary groundtruth switch scores collected at time step t.

3. Experiment

Experiment settings. Following the standard protocol [2], we extract a subset of the Whole-Body Human Motion Database (WBHM) [4], which includes scenes that contain body poses of at least one human entity, multiple movable objects, and a stationary object such as "table". This set includes 233 videos with 20 entity classes. The selected features include 3D skeleton poses of 18 joints for human entities and 3D bounding boxes for objects, sampled at 10Hz, consistent with the compared methods [2].

We compare PTD with CRNN [2], Structural RNN [3], and the simple GRU, whose implementations are redone for consistency. We follow the common settings of observing 10 time steps (1 second) and predicting future human and object motions for the next 20 time steps (2 seconds). The models are conventionally trained on 80% of the videos and evaluated on the remaining 20%.

Quantitative evaluation. The performances are measured by prediction errors of joint positions in the Euclidean distance (in mm). The means and standard deviations from five independent runs are plotted in Fig. 4 and show that PTD consistently outperforms the state-of-the-art, especially in long-term prediction.

Visual analysis. We further verify the benefit of the duality by visualizing the internal output predictions and graph structures of PTD compared to CRNN [2]. The upper row of Fig. 5 shows that PTD could learn to switch from the Persistent dense graph to the Transient egocentric graph when

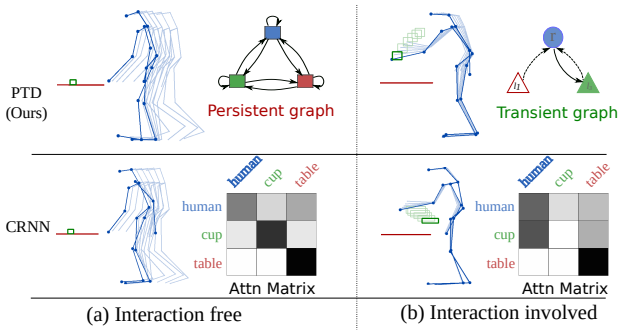


Figure 5. Persistent-transient duality in HOI-M. When the situation changes from interaction-free (a) to interaction-involved (b), PTD *(Upper row)* switches on its Transient channel with egocentric structures and handles the interaction accurately; In contrast, CRNN [2] *(Lower row)* uses a single mechanism, resulting in the sluggish adaptation of the attention map, leading to inaccurate predictions.

the situation changes from interaction-free (a) to interaction-involved (b). Particularly, in (b), the Transient graph reflects the interactions correctly thanks to it being trained only on targeted interaction samples free of noises.

In contrast, CRNN [2] (lower row) holds on to a single global mechanism and does not evolve adequately for the swift change in the true relational topology, resulting in inaccurate and unrealistic interactions.

4. Conclusion

In this work, we have introduced a new concept of the Persistent-Transient duality to model the interleaving of global dynamics and the short-lived interactions in human behavior. We model this conceptual duality into a parent-child multi-channel network that can switch between the two processes seamlessly. The superior performance of our model on the HOI subset of WBHM confirms the effectiveness of this duality in human behavior modeling.

References

[1] Adam P Baker, Matthew J Brookes, Iead A Rezek, Stephen M Smith, Timothy Behrens, Penny J Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *Elife*, 3:e01867, 2014. 1

[2] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 2.3, 3, 3, 5

[3] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 3

[4] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. 3