A Multimodal Predictive Agent Model for Human Interaction Generation

Murchana Baruah and Bonny Banerjee
Institute for Intelligent Systems, and Department of Electrical and Computer Engineering,
The University of Memphis, Memphis, TN 38152, USA
{mbaruah, bbnerjee}@memphis.edu

Abstract

Perception and action are inextricably tied together. We propose an agent model which consists of perceptual and proprioceptive pathways. The agent actively samples a sequence of percepts from its environment using the perception-action loop. The model predicts to complete the partial percept and propriocept sequences observed till each sampling instant, and learns where and what to sample from the prediction error, without supervision or reinforcement. The model is implemented using a multimodal variational recurrent neural network. The model is exposed to videos of two-person interactions, where one person is the modeled agent and the other person's actions constitute its visual observation. For each interaction class, the model learns to selectively attend to locations in the other person's body. The proposed attention-based agent is the first of its kind to interact with and learn end-to-end from human interactions, and generate realistic interactions with performance comparable to models without attention and using significantly more computational resources.

1. Introduction

The human visual system operates efficiently by attending to the environment selectively in space and time, and combines information from fixations over time to build up an internal representation of the observation [25], guiding future eye movements and decision making. Inspired by the human visual system, we propose a predictive agent[1] model which observes its visual environment via a sequence of glimpses. The agent is implemented in software; its actions are limited to sampling the visual environment and its own body movements. The predictive agent actively makes inferences (predictive and causal), acts and learns by minimizing sensory prediction error in a perception-action loop.

[1] An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators [26]. There are many applications of such agent (e.g., [1, 2, 18, 19, 24]).

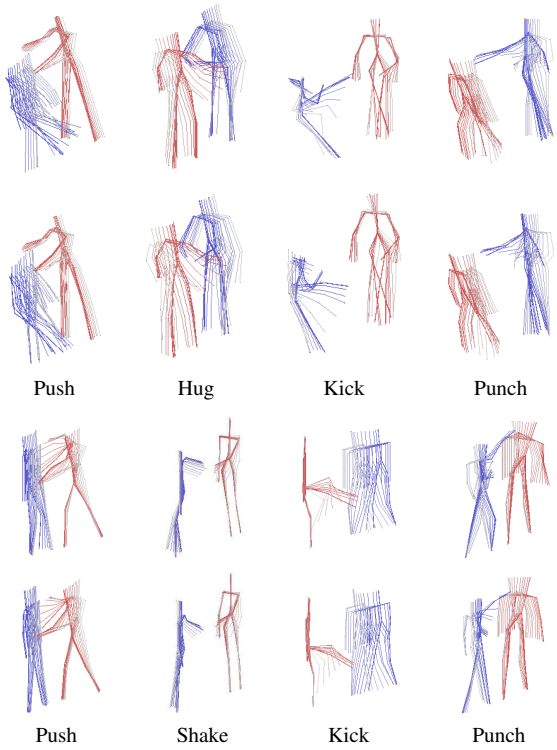


Figure 1: First and second rows show the actual and predicted data respectively for interactions push, hug, kick and punch from SBU Kinect interaction dataset. As the videos are short in length, continuous frames are shown. Third and fourth rows show the actual and predicted data respectively for interactions push, shake hands, kick and punch from K3HI interaction dataset. As the videos are longer in length, the frames are shown in intervals. Older frames are lighter in shade than more recent frames.

The model is unsupervised, and does not require reinforcement or utilities/values of states.

We apply the model for forecasting human interactions using 3D skeletal data. Interaction forecasting is a chal-

lenging problem as the model has to learn how the behavior of one person determines the behavior of the other. Spatiotemporal relations between different skeletal joints of a person as well as the two interacting persons have to be learned for accurate prediction. The ability to model dynamics of human interaction is useful for applications such as video surveillance, human-robot interaction, assistive robotics, and robotic surveillance. Though a large volume of work has been done on predicting actions using 3D skeletal data of a single person (e.g., [5, 6, 11, 12, 22, 32]) as well as predicting human motion in crowded scenes (e.g., [10, 15, 28, 29]), much less has been done on predicting interaction of two persons using 3D skeletal data.

In this paper, we model the environment from the perspective of one of the interacting persons; the other person constitutes his environment. The novelty of our approach is threefold: (1) the modeled person (agent) learns to sample (or attend to) the most informative (or salient[2]) locations of the other persons body using a saliency map at each glimpse; (2) taking into account the past observations and its learned knowledge, the agent completes the entire perceptual and proprioceptive patterns after each glimpse; and (3) the pattern completion component in our agent is a multimodal generative model where the prediction error in a perceptual modality provides the observation for the proprioceptive modality. Attending the environment selectively introduces sparsity in the agent's observations, leading to efficiency. To the best of our knowledge, the proposed agent is the first of its kind to interact with and learn end-to-end from two-person interaction environments, with performance comparable to models without attention that uses significantly less sparse observations.

2. Related Work

A taxonomy of the models used for generating actions with 3D skeletons is presented below.

- 3D skeletal data generation models
  - Single-person action generation
    - Non-attentional models [4--6, 11--13, 22, 23, 32, 34]
    - Attentional models [30]
  - Two-person interaction generation
    - Non-attentional models [17]
    - Attentional models [Our proposed model]

The model in [17] frames dual agent interaction as an optimal control problem by observing actions from one agent and predicting actions of the other agent. It does not model the observing agent's movement and predicts for short term only, unlike our proposed model. Work on predicting dual agent interactions using 3D skeletal data is limited. Most works report predicting motion of a single person using 3D skeletal data.

Few models have been proposed with attention mechanism for generating 3D skeletal data. The model in [30] predicts the 3D skeletal data of a person using a temporal attention layer which generates an attention parameter at each time step. In this model, attention is defined by internal parameters and is not a function of the model's sensory prediction error, making it difficult to interpret the model's behavior. It also requires a fixed length of the input sequence to be observed in order to calculate an attention value for each time step, which may not be realistic for online application. We propose a novel attention mechanism based on sensory prediction error, that can complete the observation from any time step, with an interpretable behavior.

3. Models and Methods

This section defines the problem and describes the proposed agent model.

3.1. Problem Statement

Let $\mathbf{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}\}$ be a set of observable variables representing an environment in n modalities. The variable representing the i-th modality is a sequence: $\mathbf{X}^{(i)} = \langle X_1^{(i)}, X_2^{(i)}, \ldots, X_T^{(i)} \rangle$, where T is the sequence length. Let $\mathbf{x}_{\leq t} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ be a partial observation of $\mathbf{X}$ such that $\mathbf{x}^{(i)} = \langle x_1^{(i)}, \ldots, x_t^{(i)} \rangle$, $1 \leq t \leq T$. We define *pattern completion* as the problem of generating $\mathbf{X}$ as accurately as possible from its partial observation $\mathbf{x}_{\leq t}$. Given $\mathbf{x}_{\leq t}$ and a generative model p_θ with parameters θ and latent variables $z_{\leq t}$, the generative process of $\mathbf{X}$ is:

$$p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}) = \int p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t}) dz \qquad (1)$$

At any time t, the objective for pattern completion is to maximize the log-likelihood of $\mathbf{X}$, i.e. $\arg\max_\theta \int log(p_\theta(\mathbf{X}|\mathbf{x}_{\leq t}, z_{\leq t}; \theta) p_\theta(z_{\leq t})) dz$.

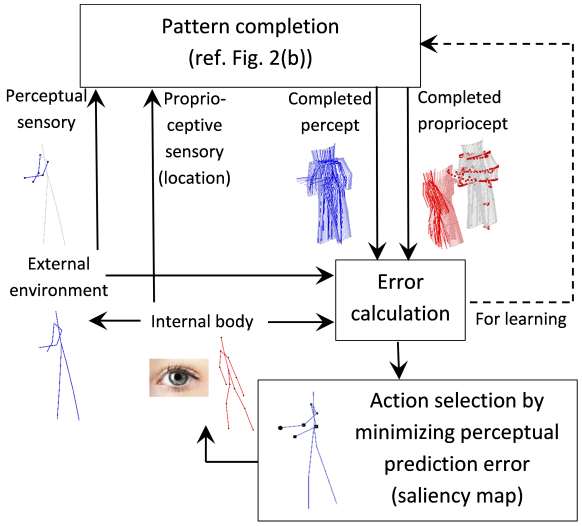
3.2. Agent Architecture

The proposed predictive agent architecture comprises of five components: environment, observation, pattern completion, action selection, and learning. See Fig. 2a.

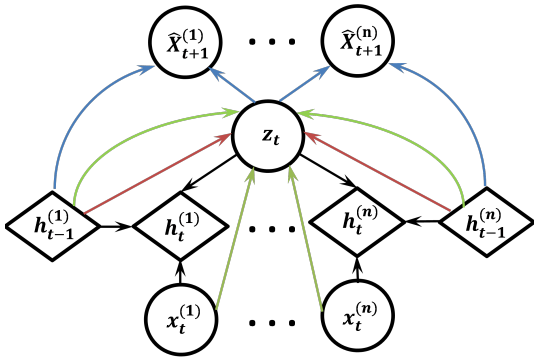
Environment. The environment is the source of sensory data and is dynamic (time-varying).

Observation. The agent interacts with the environment via a sequence of glimpses. The observations, sampled from the environment at each glimpse, are in two modalities: per-

[2]Saliency is a property of each location in a predictive agent's environment. The attention mechanism is a function of the agent's prediction error [3, 18, 19, 24, 27]. Other definitions of saliency (e.g., [8, 9]) are not relevant to this paper.



(a) Predictive agent architecture.



(b) Pattern completion model.

Figure 2: (a) Components of the proposed agent. The red skeleton is the agent's own body while the blue is that of the other person. (b) Graphical illustration of all operations of the multimodal VRNN used for pattern completion. Red arrows show computation of the conditional prior, blue arrows show the generation process, black arrows show the updating process of the RNN's hidden states, and green arrows show the inference of the approximated posterior.

ceptual[3] and proprioceptive[4]. In the context of interaction generation, we define perceptual and proprioceptive sensory observations for an interacting person as follows.

Perceptual sensory observation. Perceptual sensory reports the visual observation at some location or region in the environment. $\mathbf{x}^{(1)} = \langle x_1^{(1)}, \ldots, x_T^{(1)} \rangle$, where $x_t^{(1)} \in \mathbb{R}^{3\times N}$ denotes the other person's N 3D skeletal joints at time t.

[3]Perception is the mechanism that allows an agent to interpret sensory signals from the external environment [14].

[4]Proprioception is perception where the environment is the agent's own body. Proprioception allows an agent to internally perceive the location, movement and action of parts of its body [14].

Proprioceptive sensory observation. Proprioceptive sensory reports the activations of the agent's joint muscles due to body movement and oculomotor muscles due to fixation. The activations of joint muscles over time (or body proprioceipt sequence) is $\mathbf{x}^{(2)} = \langle x_1^{(2)}, \ldots, x_T^{(2)} \rangle$, where $x_t^{(2)} \in \mathbb{R}^{3\times N}$ denotes N 3D skeletal joints at time t. The activation of oculomotor muscles over time (or visual proprioceipt sequence) is represented by the sequence of fixation locations in the environment, denoted as $\mathbf{x}^{(3)} = \langle x_1^{(3)}, \ldots, x_T^{(3)} \rangle$, where $x_t^{(3)} \in \{0,1\}^M$ is the activation at time t of skeletal joints reduced to M fixated regions (see Fig. 3).

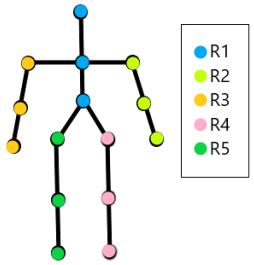


Figure 3: The M (=5) regions in the 3D human skeleton.

Pattern completion. A multimodal variational recurrent neural network (VRNN) for variable length sequences is used for completing the pattern for the three modalities (see Fig. 2b). The two processes involved in the operation of a VRNN are recognition and generation [7].

Recognition (Encoder). The recognition model, $q_\phi(z_t|\mathbf{x}_{\leq t}, z_{<t})$, is a probabilistic encoder [21]. Given the observations $\mathbf{x}_{\leq t}$, it produces a Gaussian distribution over the possible values of the code z_t from which the observations $\mathbf{x}_{\leq t}$ could have been generated. The recognition model consists of three RNNs, each with one layer of long-short term memory (LSTM) units. Each RNN generates the parameters for the approximate posterior distribution $(\mu_{z,t}^{(i)}, \sigma_{z,t}^{(i)})$ and the prior distribution $(\mu_{0,t}^{(i)}, \sigma_{0,t}^{(i)})$ for each modality i ($i = 1, 2, 3$), as in [7]. The parameters from each modality and for each distribution are combined using product of experts (PoE), as in [31], to generate the joint distribution parameters (see Fig. 2b) for both the prior $p_\theta(z_t|\mathbf{x}_{<t}, z_{<t})$ and the approximate posterior $q_\phi(z_t|\mathbf{x}_{\leq t}, z_{<t})$ given by $(\mu_{0,t}, \sigma_{0,t})$ and $(\mu_{z,t}, \sigma_{z,t})$ respectively. The recognition model can be formulated as:

$$[\mu_{0,t}^{(i)}\ \sigma_{0,t}^{(i)}] = \varphi_\tau^{prior}(h_{t-1}^{(i)}), \quad [\mu_{z,t}^{(i)}\ \sigma_{z,t}^{(i)}] = \varphi_\tau^{enc}(x_t^{(i)}, h_{t-1}^{(i)})$$

$$z_t \sim \mathcal{N}(\mu_{0,t}, \sigma_{0,t}), \quad z_t|x_t \sim \mathcal{N}(\mu_{z,t}, \sigma_{z,t})$$

$$\sigma_{0,t} = \Big(\sum_i \sigma_{0,t}^{(i)-2}\Big)^{-1}, \quad \sigma_{z,t} = \Big(\sum_i \sigma_{z,t}^{(i)-2}\Big)^{-1}$$

$$\mu_{0,t} = \Big(\sum_i \mu_{0,t}^{(i)} \sigma_{0,t}^{(i)-2}\Big)\sigma_{0,t}, \quad \mu_{z,t} = \Big(\sum_i \mu_{z,t}^{(i)} \sigma_{z,t}^{(i)-2}\Big)\sigma_{z,t}$$

where φ_τ^{prior} and φ_τ^{enc} are functions representing neural networks. It is assumed that the prior z_t and the approximated posterior $z_t|x_t$ are sampled from an isotropic multivariate Gaussian distribution.

Generation (Decoder). The generative model, $p_\theta(\mathbf{X}_{t+1}|z_{\leq t}, \mathbf{x}_{\leq t})$, generates the data from the latent variables, $z_{\leq t}$, at each time step. The generative model has three RNNs with one layer of hidden LSTM units. Each RNN generates the parameters of the distribution of the sensory data for a modality. The sensory data is sampled from this distribution which can be multivariate Gaussian or Bernoulli. In our model, $X_{t+1}^{(1)}|z_t, X_{t+1}^{(2)}|z_t$ are sampled from an isotropic multivariate Gaussian distribution and $X_{t+1}^{(3)}|z_t$ from a Bernoulli distribution. The generative model can be formulated as:

$$h_t^{(i)} = f_\theta(z_t, x_t^{(i)}, h_{t-1}^{(i)}), \quad [\mu_{x^{(i)},t}^{(i)} \; \sigma_{x^{(i)},t}^{(i)}] = \varphi_\tau^{dec}(z_t, h_t^{(i)}).$$

For Gaussian distribution, $X_{t+1}^{(i)}|z_t \sim \mathcal{N}(\mu_{x^{(i)},t}^{(i)}, \sigma_{x^{(i)},t}^{(i)})$. For Bernoulli distribution, $X_{t+1}^{(i)}|z_t = f_\sigma(h_t^{(i)})$. Here φ_τ^{dec}, f_θ are functions representing neural networks, and f_σ is a sigmoid function. The above equations facilitate one step ahead prediction. Beyond time t, for long term predictions or pattern completion, the input is the prediction from the previous time steps. Pattern completion is done at every time step.

Action selection. In the proposed agent model, action selection is to decide which location in the environment to sample from. The environment is a 3D skeleton of the interacting person. As the movement of a joint in the skeleton is dependent on its adjacent joints, we cluster the N skeletal joints into M regions (see Fig 3). Location refers to all the skeletal joints in top k_t salient regions, $1 \leq k_t \leq M$, and k_t is not fixed. At any time step, the agent selects k_t regions using a threshold. At any time, there are $\sum_{k_t=1}^{M-1} \binom{M}{k_t}$ possible actions to choose from.

An action at time t is generated as a function of the saliency map. We denote the saliency map at time t as $S_t \in \mathbb{R}^N$ and the value of the saliency map at region ℓ as $S_t^{(\ell)}$. The saliency map is a function of the prediction error computed as $S_t = \|X_t^{(1)} - \hat{X}_t^{(1)}\|_1$, where $X_t^{(1)}, \hat{X}_t^{(1)} \in \mathbb{R}^{3\times N}$ are the true and predicted perceptual data (skeleton joint coordinates) respectively, and $\|.\|_1$ denotes L_1 norm. We consider saliency over $M = 5$ regions in the skeleton. This region-based saliency map, $S_t^\ell \in \mathbb{R}^M$, is obtained by averaging the saliencies over the joints in each region. The region ℓ is considered salient if $S_t^\ell \geq \frac{1}{M}\sum_{r=1}^{M} S_t^r$. Thus, at any time, at least one region will be salient. A variable number of salient regions at each time step is more effective. Setting the number of salient regions to a constant value might occasionally lead to selection of regions with low saliency or discard regions with high saliency as saliency, S_t, is a function of time, the agent's observations and its predictive model. In the proposed model, for the salient joints, the observation is sampled from the environment; for the non-salient joints, the observation is predicted from the last time step.

The salient regions at any time t is the proprioceptive observation $x_{t+1}^{(3)}$ for time $t+1$. Therefore, the salient regions at $t = 0, 1, 2, \ldots, T-1$ constitutes the proprioceptive pattern $\mathbf{X}^{(3)}$. Hence, prediction error (saliency) guides the sampling of the observations in our model. Unlike typical multimodal models, the modalities in our model interact at the observation level as the perceptual prediction error provides the observation for the proprioceptive modality.

The agent learns a policy to generate the proprioceptive pattern or the sequence of expected salient locations by minimizing the proprioceptive prediction error (first term in Eq. 2 for $i = 3$). This error, at any time, is a function of the difference between predicted fixation location from the learned policy and the most salient location in the scene.

The most salient location is the most informative location in the environment. These are the locations where the agent's prediction error is the highest given all the past observations. The agent attends to these locations to update its internal model.

Learning. The recognition and generative model parameters are jointly learned by maximizing the ELBO for the multimodal VRNN. This objective function, obtained by modifying the objective for multimodal VAE (Eq. 2 in [31]) with VRNN (Eq. 1 in [7]), is as follows:

$$\mathbb{E}_{q_\phi(z_{\leq T}|\mathbf{x}_{\leq T})}\Big[\sum_{t=1}^{T-1}\Big[\sum_{i=1}^{n}\lambda_i \log p_\theta(X_{t+1}^{(i)}|z_{\leq t}, x_{\leq t}^{(i)}) - \beta \mathrm{KL}[q_\phi(z_t|\mathbf{x}_{\leq t}, z_{<t}), p_\theta(z_t|\mathbf{x}_{<t}, z_{<t})]\Big]\Big] \quad (2)$$

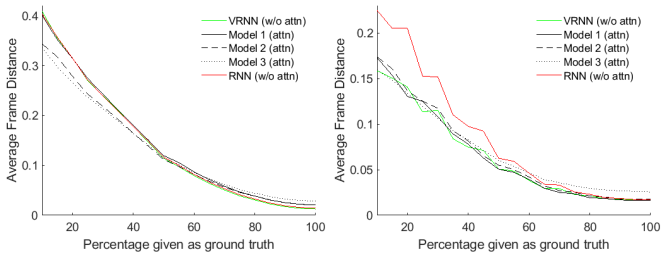
where n is the number of modalities, the first term for $i = 1, 2, 3$ is the expected negative prediction error for the three modalities. The KL-divergence is a regularizer to prevent overfitting during training.

The negative of the ELBO is also referred to as negative log-likelihood (NLL). In this paper, we refer to the negative of the first term in Eq. 2 for $i = 1$ and $i = 2, 3$ as perceptual NLL and proprioceptive NLLs respectively.

4. Experimental Results

4.1. Datasets

SBU Kinect Interaction Dataset [33] is a two-person interaction dataset comprising of eight interactions: approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. The data is recorded from 7 participants forming a total of 21 sets such that each



(a) SBU Kinect Interaction Dataset (b) K3HI Interaction Dataset

Figure 4: AFD averaged over all actions and each dataset for different percentage of ground truth given as input. For any percentage p, $p\%$ of the actual data is given as input and the prediction is considered as input for the rest of the time steps.

set consists of a unique pair of participants performing all actions. The dataset has approximately 300 interactions of duration 9 to 46 frames. The dataset is divided into five distinct train test split.

K3HI: Kinect-based 3D Human Interaction Dataset [16] is a two-person interaction dataset comprising of eight interactions: approaching, departing, kicking, punching, pointing, pushing, exchanging an object, and shaking hands. The data is recorded from 15 volunteers. Each pair of participants performs all the actions. The dataset has approximately 320 interactions of duration 20 to 104 frames. The dataset is divided into three distinct train test split.

4.2. Experimental setup

Each dataset consists of interactions where one person initiates an action and the other person reacts to it. In our experiments, we model one interacting person irrespective of its initiating or reacting nature. We consider 15 skeletal joints from each person for each dataset. Each skeletal joint is normalized before training.

Each modality in the agent architecture (ref. Fig. 2b) has a recurrent hidden layer of 256 hidden units and a latent layer of 10 latent variables.

We use Adam optimizer with a learning rate of 0.001, and default hyper-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [20]. A minibatch size of 100 is used and number of training iterations is fixed at 25,000 (SBU Kinect) and 10,000 (K3HI). To avoid overfitting, we use a dropout probability of 0.8 at the generation layer (final layer). All the hyperparameters are determined experimentally.

For evaluation, we consider three variants of our model:

1. **Model 1: VRNN with 2 modalities**. Perceptual and body proprioceptive are the two modalities. Here, $i = 1, 2$ in Eq. 2.

2. **Model 2: VRNN with 3 modalities**. Ref. Section 2.

3. **Model 3: VRNN with 3 modalities and perceptual input sampled from predicted visual proprioception**. This is a special case of Model 2. Here the perceptual input is sampled from the prediction, $\hat{X}_t^{(3)}$, instead of the true saliency map, S_t, at all time steps.

The difference between Model 1 and the other two models is the addition of the third modality $\mathbf{x}^{(3)}$ to the model. This difference will show the effect of adding modalities in the model. The difference between Model 3 and the other two models is the data from which the perceptual input is sampled. This difference will show how well the model learns to predict the salient joints.

We evaluate the model by comparing it with models without attention. The perceptual observation is sampled from the ground truth, $X_t^{(1)}$, at all time steps.

1. **RNN (without attention).** We use a standard LSTM encoder-decoder model and to generate data for two modalities: perceptual and body proprioceptive. The two modalities interact at the latent layer, where the latent variables are concatenated; it thus has a total of 20 latent variables.

2. **VRNN (without attention).** We use a variational LSTM autoencoder model to generate data for two modalities: perceptual and body proprioceptive. The two modalities interact at the latent layer, where the latent variables are combined using PoE.

For fair comparison, the number of layers and number of neurons are kept consistent for both models with respect to the proposed models.

We evaluate results from the perceptual modality ($i = 1$) and body proprioceptive modality ($i = 2$) using average frame distance (AFD), as in [17]: $\frac{1}{T-1}\sum_{t=2}^{T} \|X_t^{(i)} - \hat{X}_t^{(i)}\|^2$, where $X_t^{(i)}$ and $\hat{X}_t^{(i)}$ are the true and predicted skeletal joint coordinates respectively at time t, and T is the sequence length.

We use percentage measure to evaluate the two proprioceptive modalities in Model 2. The measure reflects how good the learned policy (generated sequence of salient regions) is when compared to the true policy (true sequence of salient regions). At each time step, the true policy can generate multiple salient regions. We define the average percentage as:

$$\frac{1}{T-1}\sum_{t}\frac{No.\ of\ correctly\ predicted\ salient\ regions}{Total\ no.\ of\ salient\ regions}$$

4.3. Evaluation Results

Fig. 1 shows one time step ahead prediction of the two skeletons (perception and body proprioception) for four

Table 1: Performance comparison for different versions of the proposed model and other models for different interactions on the SBU Kinect Interaction dataset for one step ahead prediction. The reported AFD is the average of the perceptual (Perc.) AFD and proprioceptive (Prop.) AFD averaged over all examples in the test set and all the train-test splits. The visual proprioceptive performance is shown in the last column.

Interaction	Perc. error and Body Prop. error (AFD)						Visual Prop. (%)
	[17]	RNN (w/o attn)	VRNN (w/o attn)	Model 1 (VRNN, attn)	Model 2 (VRNN, attn+ true policy)	Model 3 (VRNN, attn+ pred. policy)	Model 2 (VRNN, attn+ true policy)
Approaching	-	0.0097	0.0082	0.0128	0.0138	0.0189	61.08
Departing	-	0.0117	0.0098	0.0140	0.0150	0.0199	61.41
Kicking	0.660	0.0210	0.0192	0.0358	0.0360	0.0411	61.77
Pushing	0.413	0.0142	0.0125	0.0212	0.0215	0.0267	64.79
Shaking	0.389	0.0094	0.0079	0.0130	0.0130	0.0276	62.25
Hugging	0.504	0.0197	0.0181	0.0273	0.0272	0.0412	63.80
Exchanging	0.574	0.0111	0.0095	0.0136	0.0145	0.0195	65.16
Punching	0.510	0.0175	0.0159	0.0252	0.0258	0.0326	63.19
Average	0.508	0.0143	0.0126	0.0204	0.0208	0.0284	62.93

Table 2: Performance comparison for different versions of the proposed model and other models for different interactions on the K3HI dataset for one step ahead prediction. The reported AFD is the average of the perceptual (Perc.) AFD and proprioceptive (Prop.) AFD averaged over all examples in the test set and all the train-test splits. The visual proprioceptive performance is shown in the last column.

Interaction	Perc. error and Body Prop. error (AFD)					Visual Prop. (%)
	RNN (w/o attn)	VRNN (w/o attn)	Model 1 (VRNN, attn)	Model 2 (VRNN, attn+ true policy)	Model 3 (VRNN, attn+ pred. policy)	Model 2 (VRNN, attn+ true policy)
Approaching	0.0844	0.0912	0.0714	0.0735	0.0796	72.68
Departing	0.0065	0.0067	0.0097	0.0102	0.0130	69.44
Exchanging	0.0022	0.0024	0.0036	0.0038	0.0072	73.54
Kicking	0.0047	0.0049	0.0078	0.0078	0.0117	69.43
Pointing	0.0025	0.0028	0.0048	0.0048	0.0089	69.77
Punching	0.0038	0.0040	0.0064	0.0067	0.0101	70.35
Pushing	0.0035	0.0038	0.0062	0.0065	0.0090	66.99
Shaking	0.0019	0.0021	0.0031	0.0034	0.0065	67.86
Average	0.0137	0.0147	0.0141	0.0146	0.0182	70.00

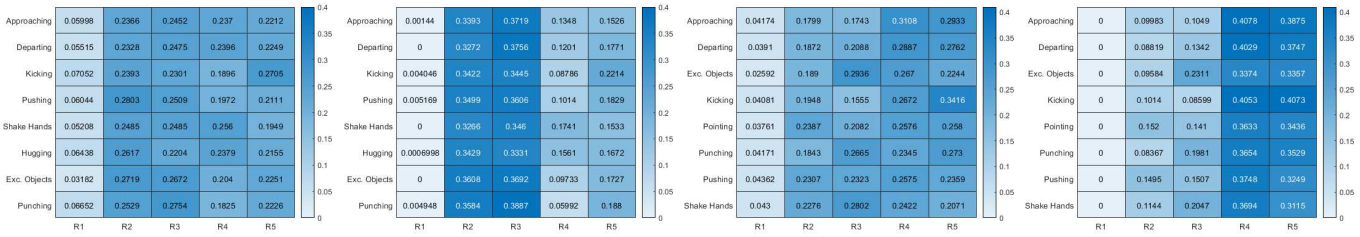
kinds of interactions from each dataset. The prediction over space and time looks quite realistic for all the cases.

For long term predictions, the prediction improves exponentially with the percentage of data given as ground truth (see Fig. 4). For SBU Kinect dataset, the performance of RNN (without attention), VRNN (without attention) and Model 1 (with attention) is slightly poorer than the proposed Model 2 (with attention) and Model 3 (with attention) until around 50% of the ground truth is given as the input. For ground truth $\geq 50\%$, the AFD for non-attention models and proposed Model 1 slightly improves compared to proposed Models 2 and 3. For K3HI dataset, the performance of RNN is poorer than Model 1, Model 2, Model 3, and VRNN until around 75% of the ground truth is given as the input. For ground truth $\geq 75\%$, the AFD for all the models except Model 3 are close. Thus, overall performance of attention and non-attention models are comparable.

RNN and VRNN without attention are more prone to error propagation as the predicted data is fed as the input for consecutive prediction whereas during training, the ground truth is fed as input to the model. Our model is more robust to noise as during training, for the non-salient joints, the predicted data is fed as the input for consecutive prediction. VRNNs though in general are more robust to noise, adding the proposed attention mechanism with sparsity can help in combating error propagation and improve long-term predictions. Detailed AFD for all the interactions for one time step ahead prediction are shown in Tables 1 and 2. The AFD for all the interactions is lower than the results reported in [17] for SBU Kinect dataset. This shows that our model is able

Table 3: Average percentage of saliency joints for both SBU Kinect Interaction and K3HI dataset for all cases in an interaction. This percentage is also the proportion of joints that are sampled from the observation (ground truth).

SBU	Approaching	Departing	Kicking	Pushing	Shaking	Exchanging	Punch	Hugging	Average
	46.21	46.99	44.87	47.54	47.23	47.36	47.80	47.45	46.93
K3HI	Approaching	Departing	Kicking	Pushing	Shaking	Exchanging	Punching	Pointing	Average
	45.19	46.27	43.79	48.06	48.11	50.00	47.57	47.70	47.09



(a) SBU true dist. (b) SBU predicted dist. (c) K3HI true dist. (d) K3HI predicted dist.

Figure 5: Salient region distribution (dist.) over all interactions shown in (a–d) averaged over all the examples in an interaction. True salient distribution obtained from the saliency map for SBU Kinect Interaction dataset is shown in (a) and for K3HI dataset is shown in (c). The predicted salient distribution obtained from the predicted joints as in proposed Model 2 are shown for SBU Kinect Interaction dataset and K3HI dataset in (b) and (d) respectively.

to learn better representation of the underlying dynamics of interaction. For K3HI dataset, we are the first to report the AFD. Among the three variants of the proposed model, for one step ahead prediction, Model 1 performs the best. No significant difference in long-term prediction performance is observed among the three variants (see Fig. 4). However, from a practical standpoint, Model 2 and Model 3 can be more useful as they can learn the policy and automatically determine salient regions for future time steps. So the agent can decide what action to take much earlier than the actual event occurs.

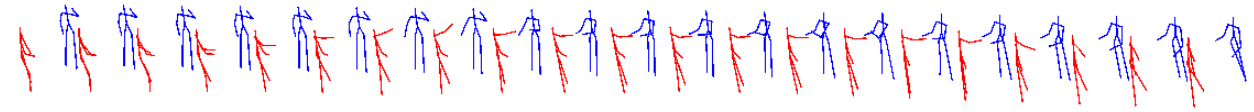
Predictions closer to the current time step are better, as observed from Figs. 6, 7. There is continuity and the two predicted skeletons are well synchronized. The agent's predicted action or reaction at each time step also complies with the actual interaction.

It is also observed that the number of actual salient regions may change at each time step depending on the prediction error (highlighted with markers in the joints of the skeletons). This change may occur quite randomly depending on what the model has learned. Therefore, learning to predict the salient regions is a challenging task. However, our model is able to predict them correctly most of the time (ref. Tables 1, 2, and Fig. 5). Ideally, the joint trajectories that occur rarely are more difficult to learn, and hence more salient. Thus, the actual salient regions for punching, exchange objects, push, handshake and hug are mostly the hands while for action kicking its the legs. In our case, as the modeled agent can be the reacting or interacting agent, the salient region distribution is similar but not exactly the same as the ideal case.

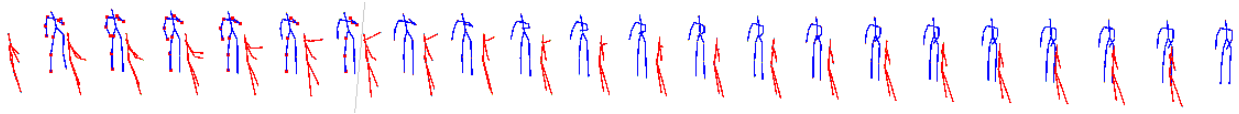
We compute the average (over all the cases for an interaction) of the percentage of number of salient joints chosen by the model at each time step (ref. Table 3). On average, for all the interactions, our approach considers less than 50% of the joints in a skeleton as observation to the model. For both datasets, the highest sparsity is for kicking, the lowest is for punching and shaking hands for SBU Kinect and K3HI datasets respectively. Selectively attending to fewer joints makes our model more efficient without compromising its accuracy.

5. Conclusions

A multimodal predictive agent with perceptual and proprioceptive pathways is proposed. It completes the observed pattern for perceptual and proprioceptive modalities after each glimpse. The perceptual prediction error provides the observation for the proprioceptive modality. Experimental results using our agent for two-person interaction forecasting are comparable to non-attentional models even though our agent's observations have higher than 50% sparsity. The agent model is learned end-to-end in an unsupervised manner, without any reinforcement signal or utilities/values of states. This is the first work on an attention-based agent that actively samples its environment guided by prediction error and generates realistic 3D human skeleton interactions.

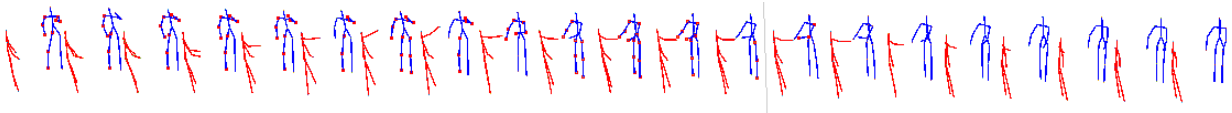


(a) Actual



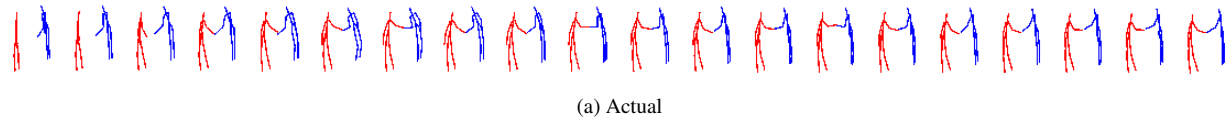
(b) Predicted (30% ground truth given)

(c) Predicted (50% ground truth given)

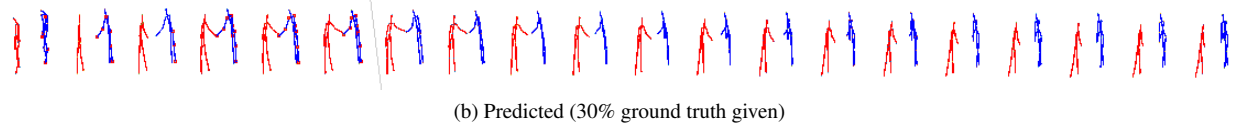


(d) Predicted (70% ground truth given)

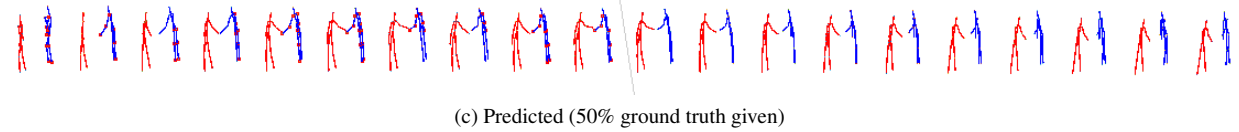
Figure 6: The top row represents true skeletal data for the prediction at alternate time steps for SBU Kinect Intersection data for exchanging object. Each skeleton in rows 2, 3 and 4 shows one step ahead prediction until 30%, 50% and 70% of the ground truth is given (highlighted by the grey line) respectively. Beyond that, the model uses its own prediction as input for completing the patterns until the final time step is reached. The salient joints are marked red.



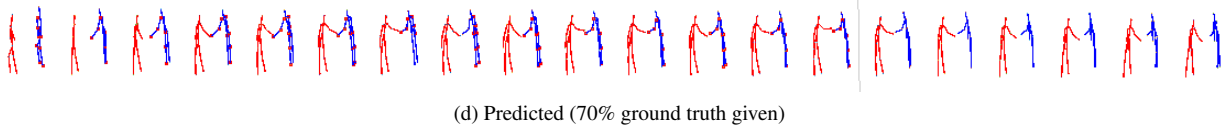
(a) Actual



(b) Predicted (30% ground truth given)



(c) Predicted (50% ground truth given)



(d) Predicted (70% ground truth given)

Figure 7: The top row represents true skeletal data for the prediction at every third instant for K3HI Intersection data for shaking hands. Each skeleton in rows 2, 3 and 4 shows one step ahead prediction until 30%, 50% and 70% of the ground truth is given (highlighted by the grey line) respectively. Beyond that, the model uses its own prediction as input for completing the patterns until the final time step is reached. The salient joints are marked red.

References

[1] B. Banerjee and B. Chandrasekaran. A constraint satisfaction framework for executing perceptions and actions in diagrammatic reasoning. *J. Artif. Intell. Res.*, pages 373–427, 2010. 1

[2] B. Banerjee and B. Chandrasekaran. A spatial search framework for executing perceptions and actions in diagrammatic reasoning. In *Diagrammatic Representation and Inference, LNAI*, volume 6170, pages 144–159. Springer, Heidelberg, 2010. 1

[3] B. Banerjee and J. K. Dutta. SELP: A general-purpose framework for learning the norms from saliencies in spatiotemporal data. *Neurocomputing*, 138:41–60, 2014. 2

[4] E. Barsoum, J. Kender, and Z. Liu. HP-GAN: Probabilistic 3D human motion prediction via GAN. In *CVPR Workshops*, pages 1418–1427, 2018. 2

[5] J. Bütepage, H. Kjellström, and D. Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *ICRA*, pages 1–9. IEEE, 2018. 2

[6] H. Chiu, E. Adeli, B. Wang, D. Huang, and J. Niebles. Action-agnostic human pose forecasting. In *WACV*, pages 1423–1432. IEEE, 2019. 2

[7] J. Chung et al. A recurrent latent variable model for sequential data. In *NIPS*, pages 2980–2988, 2015. 3, 4

[8] J. K. Dutta and B. Banerjee. Online detection of abnormal events using incremental coding length. In *AAAI*, pages 3755–3761, 2015. 2

[9] J. K. Dutta, B. Banerjee, and C. K. Reddy. RODS: Rarity based outlier detection in a sparse coding framework. *IEEE Trans. Knowl. Data Eng.*, 28(2):483–495, 2016. 2

[10] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+ hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection. *Neural Netw.*, 108:466–478, 2018. 2

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 2

[12] P. Ghosh, J. Song, E. Aksan, and O. Hilliges. Learning human motion models for long-term predictions. In *Intl. Conf. 3D Vision*, pages 458–466. IEEE, 2017. 2

[13] L. Gui, Y. Wang, X. Liang, and J. Moura. Adversarial geometry-aware human motion prediction. In *ECCV*, pages 786–803, 2018. 2

[14] J. Han, G. Waddington, R. Adams, J. Anson, and Y. Liu. Assessing proprioception: A critical review of methods. *J. Sport Health Sci.*, 5(1):80–90, 2016. 3

[15] Y. Hoshen. VAIN: Attentional multi-agent predictive modeling. In *NIPS*, pages 2701–2711, 2017. 2

[16] T. Hu, X. Zhu, W. Guo, and K. Su. Efficient interaction recognition through positive action representation. *Math. Probl. Eng.*, 2013, 2013. 5

[17] D. Huang and K. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*, pages 489–504. Springer, 2014. 2, 5, 6

[18] M. H. Kapourchali and B. Banerjee. State estimation via communication for monitoring. *IEEE Trans. Emerg. Topics Comput. Intell.*, 2019. 1, 2

[19] M. H. Kapourchali and B. Banerjee. EPOC: Efficient perception via optimal communication. In *AAAI*, 2020. 1, 2

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] D. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[22] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee. Convolutional sequence to sequence model for human dynamics. In *CVPR*, pages 5226–5234, 2018. 2

[23] X. Lin and M. Amer. Human motion modeling using DV-GANs. *arXiv preprint arXiv:1804.10652*, 2018. 2

[24] S. Najnin and B. Banerjee. A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Commun.*, 92:24–41, 2017. 1, 2

[25] R. A. Rensink. The dynamic representation of scenes. *Vis. Cogn.*, 7(1-3):17–42, 2000. 1

[26] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2002. 1

[27] M. Spratling. Predictive coding as a model of the V1 saliency map hypothesis. *Neural Netw.*, 26:7–28, 2012. 2

[28] D. Varshneya and G. Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. *arXiv preprint arXiv:1705.09436*, 2017. 2

[29] A. Vemula, K. Muelling, and J. Oh. Social attention: Modeling attention in human crowds. In *ICRA*, pages 1–7. IEEE, 2018. 2

[30] P. Vinayavekhin, S. Chaudhury, A. Munawar, D. Agravante, G. Magistris, D. Kimura, and R. Tachibana. Focusing on what is relevant: Time-series learning and understanding using attention. In *ICPR*, pages 2624–2629. IEEE, 2018. 2

[31] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *arXiv:1802.05335*, 2018. 3, 4

[32] Y. T. Xu, Y. Li, and D. Meger. Human motion prediction via pattern completion in latent representation space. In *Conf. Computer and Robot Vision*, pages 57–64. IEEE, 2019. 2

[33] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPR Workshops*, pages 28–35. IEEE, 2012. 4

[34] Y. Zhou, Z. Li, S. Xiao, C. He, Z. Huang, and H. Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018. 2