Simultaneous Human Action and Motion Prediction

Kourosh Darvish*, **Daniele Pucci**
Artificial and Mechanical Intelligence,
Center for Robotics and Intelligent Systems,
Istituto Italiano di Tecnologia (IIT), Genova, Italy.
(email: name.surname@iit.it)

Abstract

This paper presents a novel approach to solve simultaneously the problems of human whole-body motion prediction and action recognition for real-time applications. Starting from the dynamics of human motion and motor system theory, the notion of mixture of experts from deep learning literature has been extended to solve this problem. The work is accompanied by 66-DoFs human model experiments.

1 Introduction

This paper addresses the problem of human whole-body motion prediction and action recognition. Given an unfinished set of observed human motion, the prediction should fundamentally respond to two questions for a predefined time horizon in the future: what the human subject will do next in the short-term at the symbolic level, hence a classification problem; how the human subject will do that, i.e., motion prediction as a regression problem.

A primary application of prediction is in human-robot collaboration and joint action scenarios, where predictions of human forthcoming actions and motions allow the robot to plan and adapt itself to the human in a proactive fashion [1], hence enhancing the fluency, safety, and ergonomics. Other applications of this work include robot teleoperation in a remote environment under time delay [2], robot motion generation [3], and exoskeleton control [4].

Human action and motion prediction problems are very challenging due to intrinsic variations in human action execution, including both spatial and temporal variations. The challenges are further intensified when prediction should generalize over several human subjects, and when subjects are not restricted, i.e., at any moment can make a new decision and switch from an action to another. While many works address the problem of human action recognition, i.e., recognizing human action when its execution is over, in this work, we are interested in predicting human action for a specified time horizon in the future given an unfinished set of observations. The approaches to solve these two distinct problems are similar in the literature, thus we are using the two terms interchangeably in this manuscript.

In this paper, first, we describe human motor control policy for motion generation from a biological and dynamical system point of view. Later, we have extended the Mixture of Experts (MoE) deep neural network (DNN) approach to jointly learn the problems of human action recognition and motion prediction. For this purpose, each expert in MoE is enforced to learn a specific human motion generation policy related to a given action, hence expert outputs predict human motion, and a gating network output classifies human actions. Experiments follow.

*corresponding author. This work has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 731540 (An.Dy) and No 869855 (SoftManBot), and Italian National Institute for Insurance against Accidents (INAIL) ergoCub project.

2 Background & Problem Statement

To address human action and motion prediction problems, this section presents the underlying principles of human motion generation and action from dynamical system and human motor system perspectives. This study will support the formulation of human action and motion prediction problems with a holistic view, which in turn gives an idea of how to solve the problem in a generalized form. Following this investigation and indications, next section attempts at solving these problems.

We model a human as a Markov process expressed as a multi-body mechanical system with n joints, each with one degree of freedom, connecting $n+1$ links. The joint angle, velocity and torque vectors are denoted by $\boldsymbol{s} \in \mathbb{R}^n$, $\dot{\boldsymbol{s}} \in \mathbb{R}^n$, and $\boldsymbol{\tau} \in \mathbb{R}^n$ respectively, and the system state vector is denoted by $\boldsymbol{x} = (\boldsymbol{s}, \dot{\boldsymbol{s}}) \in \mathbb{R}^{2n}$.

According to the literature on biomechanics and motor system and human dynamics, we can write down the way a human generates new joint torques as a function of current $\boldsymbol{s}(t)$, $\dot{\boldsymbol{s}}(t)$, $\ddot{\boldsymbol{s}}(t)$, $\dddot{\boldsymbol{s}}(t)$ (joint jerks), $\boldsymbol{f}^c(t) \in \mathbb{R}^{6n_c}$ (n_c external forces), $\boldsymbol{\tau}(t)$, $\dot{\boldsymbol{\tau}}(t), \ddot{\boldsymbol{\tau}}(t)$ (the first and second derivative of joint torques resulted from muscle contractions), $\int \boldsymbol{\tau}^\mathsf{T}(t)\boldsymbol{\tau}(t)d(t)$ (joint efforts), $\int \dot{\boldsymbol{s}}^\mathsf{T}(t)\boldsymbol{\tau}(t)d(t)$ (kinetic energy of the joints), and $\boldsymbol{r}(t) \in \mathbb{R}^{n_r}$ (other n_r terms associated with the generation of joint torques) [5]. Some of the important terms that we can identify associated with $\boldsymbol{r}(t)$ are the human objective or the immediate task, the task space constraints such as obstacles, time constraints, and spatial constraints. In many works in robotics where human motion is predicted, $\boldsymbol{r}(t)$ is considered to be known implicitly. It is injected into the problem when a human should act in a structured environment or perform a given task sequence. However, in an unstructured environment or when human subjects are not provided with a description of the tasks to execute, $\boldsymbol{r}(t)$ can be considered as a hidden state in a Markov process and is required to be estimated given input data [6–8].

Remark 1 *Biomechanical studies tend to show that humans generate motion in order to minimize a cost function. This cost function combines mechanical energy expenditure (related to joint torques and velocities) and the motion smoothness (related to minimum jerk) while executing a reaching task [5, 9].*

Following Rem. 1, we can approximate the human policy for joint torque generation as an optimal control problem with an unknown cost function, i.e., an inverse optimal control problem. One can show the following optimal relationship between the next optimal state of the human $\boldsymbol{x}^*_{k+1}$ and the past human states $\boldsymbol{x}_{k-i}$, external forces $\boldsymbol{f}^c_{k-i}$, and hidden states $\boldsymbol{r}_{k-i}$ is hold:

$$\boldsymbol{x}^*_{k+1} = \mathcal{H}^*(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k-2}, \dots, \boldsymbol{x}_{k-N}, \boldsymbol{f}^c_k, \dots, \boldsymbol{f}^c_{k-N}, \boldsymbol{r}_k, \dots, \boldsymbol{r}_{k-N}). \tag{1}$$

By recursively applying (1), we can predict the future states of the human dynamical system for the time horizon T, i.e., $\boldsymbol{x}^*_{k+1}, \boldsymbol{x}^*_{k+2}, \dots, \boldsymbol{x}^*_{k+T}$. However, to estimate the future states of the human system in a recursive fashion, there are the following problems that needs to be addressed: i) the mapping $\mathcal{H}^*$ in (1) is not known; ii) external forces/torques acting on the human in the future $\boldsymbol{f}^c_{k+i}$ in (1) are not known; iii) the hidden states $\boldsymbol{r}_{k\pm i}$ in (1) are not known, neither in the past nor the future.

3 Proposed Solution

To address the challenges derived at the end of Sec. 2 for human motion prediction, we propose a learning-based approach, i.e., the mapping $\mathcal{H}^*$ in (1) is learned from human demonstrations. In the literature, approaches based on a single neural network have been proposed to learn the mapping $\mathcal{H}^*$. However, $\mathcal{H}^*$ is very complex, and yet no approach has resolved this problem effectively. Starting from (1), here first, we discuss different approaches to solve the difficulties that emerged previously and reformulate the action and motion prediction problem in a new form. Afterward, we adopt the Mixture of Experts (MoE) approach to solve the two problems simultaneously [10, 11].

In order to predict the external wrenches acting on the human in the future $\tilde{\boldsymbol{f}}_{k+i}$, one can come out with two approaches. First, given the predicted states of the human $\hat{\boldsymbol{x}}_{k+i}$, we can model the human and the world and perform simulations to predict the external forces acting on the human [12]. However, this solution is time-consuming and cumbersome to model the human and the world for different scenarios. Another approach is to learn a model of the world for relevant tasks from the human offline demonstrations and try to predict the interaction forces/torques acting on the human [13]. For this work, we have decided to go for the learning approach.

In regard to $\boldsymbol{r}_{k\pm i}$, when the human subject is not asked to do a given task, the problem becomes even more complex and depends on many variables. For example, for daily-life activities, to estimate what a human will do and how will do them, we should know the hidden internal objective (state) of the human in his mind. Using other sensory modalities like cameras, we may infer the human action, e.g., reaching an object, and human motion and trajectory, e.g., depending on the object's location and obstacles. However, this is out of the scope of this work, and we are only considering the human dynamical states and interaction forces. Moreover, depending on the type of $\boldsymbol{r}_{k\pm i}$, we can consider $\boldsymbol{r}_{k\pm i}$ as the solution of a classification or a regression problem. In this work, as a simplifying assumption, we only consider human symbolic actions as the hidden state, and will estimate it as a classification problem. In the offline phase, human actions are annotated by experts, while in the online phase, given the input data human next action is estimated, i.e., $\mathcal{P}(\boldsymbol{a}_{k+1}|\boldsymbol{x}_k, \ldots, \boldsymbol{x}_{k-N}, \boldsymbol{f}^c_k, \ldots, \boldsymbol{f}^c_{k-N})$. Noticeably, in (1), $\boldsymbol{r}_k, \ldots, \boldsymbol{r}_{k-N}$ are compacted and approximated as $\tilde{\boldsymbol{a}}_{k+1}$. Hence, equation (1) can be revised as the following set of equations:

$$\tilde{\boldsymbol{a}}_{k+1} = \mathcal{D}^*_1(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k-2}, \ldots, \boldsymbol{x}_{k-N}, \boldsymbol{f}^c_k, \ldots, \boldsymbol{f}^c_{k-N}), \tag{2a}$$

$$\tilde{\boldsymbol{x}}_{k+1}, \tilde{\boldsymbol{f}^c}_{k+1} = \mathcal{D}^*_2(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k-2}, \ldots, \boldsymbol{x}_{k-N}, \boldsymbol{f}^c_k, \ldots, \boldsymbol{f}^c_{k-N}, \tilde{\boldsymbol{a}}_{k+1}), \tag{2b}$$

where $\mathcal{D}^*_1$ and $\mathcal{D}^*_2$ are two optimal mappings to learn. As presented, the original complex problem of motion prediction introduced in (1) is transformed into action recognition in (2a) and motion prediction in (2b) problems.

In order to solve the problem of human action recognition (i.e., learn $\mathcal{D}^*_1$ in (2a)) and motion prediction (i.e., learn $\mathcal{D}^*_2$ in (2b)) jointly, we have elaborated on the idea of MoE. We consider the outputs of both the gating and expert networks as the two sets of outputs of a single and large MoE network. The gating network tries to solve the human action recognition as a classification problem. Each expert tries to predict human motion associated with an action as a regression problem. In this case, the total loss function L for MoE can be written as a linear combination of the two output losses L_1 and L_2 with the weights b_1 and b_2 that are set by the user. Here, L_1 is set as a categorical cross-entropy loss and L_2 is the mean squared error; in other problems, the user may choose different loss functions. In our case, we define the total loss as:

$$\begin{gathered} L = b_1\, L_1 + b_2\, L_2 = -\tfrac{b_1}{2M} \sum_{j=1}^{M} \sum_{i=1}^{N} a_i^j log(\tilde{a}_i^j) + \tfrac{b_2}{2M} \sum_{j=1}^{M} \|\tilde{\boldsymbol{y}}^j - \boldsymbol{y}^j\|_2, \\ s.t.\ \ \tilde{\boldsymbol{y}}^j = \sum_{i=1}^{N} \tilde{a}_i^j \tilde{\boldsymbol{y}}_i^j, \end{gathered} \tag{3}$$

where scalar value a_i^j and vector $\boldsymbol{y}_i^j$ are human action and motion (e.g., joint values) ground truth related to the i'th action and j'th data, and $\tilde{\cdot}$ indicates estimated values that are stochastically found. M is the total number of data, and N is the number of experts or modeled actions. When designing the network, b_1 and b_2 are positive numbers chosen manually as hyperparameters such that both classification (action recognition) and regression (motion prediction) problems converge while training. For this purpose, a suggested approach is first to tune b_1 such that the classification problem converges, and later accordingly, set the parameter b_2. In this way, we are ensuring each expert is learning the motion associated with an action. Moreover, l_1 and l_2 regularization terms are used to penalize the weight values and avoid overfitting, however they are not reported in the loss function in (3). Looking at (3), during back-propagation while training, we can observe that the gate weights rely on both L_1 and L_2 losses, while the expert weights only depend on L_2. This shows an important feature of the proposed approach, not only does human action affects how human moves, but also the way human moves affects the recognized action. Moreover, when human subject is performing an action a_i (assuming the optimization problem is converged), a_k goes close to zero $\forall k \neq i$, hence expert i is enforced to learn the human motion associated with i'th action. Finally, in the transient phase when the subject alters from an action to another, the two associated experts try together to reduce the error on the motion prediction output, proportional to the gate outputs.

4 Experiments, Results, and Discussions

For experiments, human data are collected using a wearable motion capture system equipped with inertial measurement units (IMUs) and a pair of shoes equipped with force/torque sensors. Human model has 66 DoFs, and an inverse kinematic implementation computes the joint values and velocities [14]. Data are sampled at 25 Hz. The programs run on a 64 bit i7 2.8 GHz workstation, equipped with 32 GB RAM, Ubuntu 20.04 LTS, and Intel(R) Iris(R) Xe Graphics. In the experimental process, the human subject was asked to walk naturally inside the room, and in total less than $8\ mins$ of data

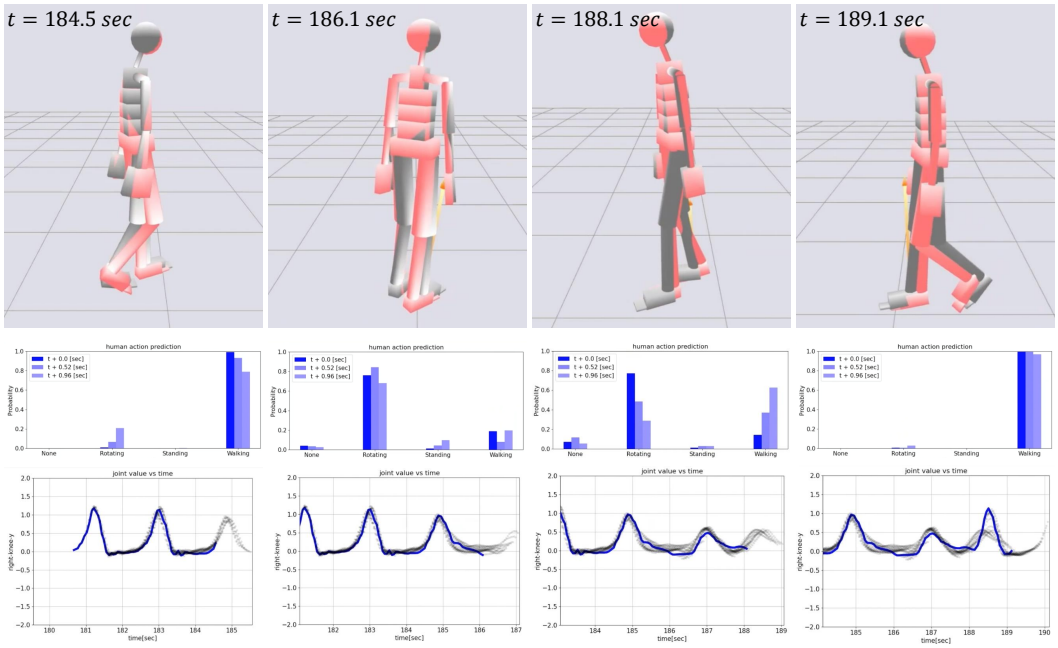

Figure 1: Snapshots of human motion prediction (top), action recognition (middle), and right knee joint angle (in radian, bottom) taken from accompanied video at https://youtu.be/XK4vmD6pJ9Q.

have been collected and carefully annotated. 70% of data are considered as the training data, 20% validation data, and the last 10% as the test dataset. The human subject was doing the following actions: *Walking*, *Rotating*, *Standing*, and other irrelevant actions were labeled as *None*. To implement MoE (using TensorFlow2), four similar experts associated with the number of human actions and one gate network have been considered. In the implementation, LSTM, Dense, Softmax, Dropout, Batch Normalization layers have been used. Moreover, l_2 regularization terms have been added to the layers to avoid overfitting. The inputs to the network are joint values and velocities, and ground reaction forces/torques with $N = 5$ in (2). Since LSTMs are inherently recursive, we predict the future of human motion directly (without autoregressive implementation) for the future time horizon of $1\ sec$, i.e., $T = 25$ steps. The training, validation, and test set results are shown in the accompanying video.

Figure 1 shows the results of the human action recognition and motion prediction at different moments. Online inference takes $26ms$ on average at each time step. On the top, it shows the snapshots of the human motion in gray color and the results of the prediction for $0.2sec$ in the future in the red color. Notice that, currently, the future base pose is not estimated, hence the two avatar bases coincide. In the middle, the results of the human action recognition are shown for three 1, 12, 24 steps in the future. Finally, at the bottom, it shows the results of the prediction of the human right knee joint angle in radians (in small gray circles) for the prediction time horizon at each step, and the blue line shows the actual value of the knee joint angle. Notice when human starts to rotate at $t = 185sec$, the probability of the human *rotating* action at $t + 0.96sec$ is higher than the one at t, and reversely for the *walking*. Later, at $t = 186sec$, it predicts the subject will rotate for the next T time horizon. Finally, at $t = 188sec$ (the fourth column on the right side), the prediction results show a trend from *rotating* to *walking* action. For $t \in [188, 189]$, first, knee joint angle is predicted with a *rotating* pattern, while later this has been transformed to a *walking* pattern as the human starts to walk.

5 Conclusions

In this paper, we proposed a novel approach for simultaneous human action and motion prediction for the short time horizon in the future. The mixture of experts (MoE) notion has been adopted to solve the two problems together, and the results show the effectiveness of the proposed solution for real-time applications. In the future, the proposed approach will be scaled to predict the actions and motion of several subjects, and use the hierarchical version of MoE in order to generalize the solution to a wider range of human actions.

References

[1] K. Darvish, E. Simetti, F. Mastrogiovanni, and G. Casalino, "A hierarchical architecture for human–robot cooperation processes," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 567–586, 2021.

[2] P. Farajiparvar, H. Ying, and A. Pandya, "A brief survey of telerobotic time delay mitigation," *Frontiers in Robotics and AI*, vol. 7, 2020.

[3] P. M. Viceconte, C. Raffaello, G. Romualdi, D. Ferigo, S. Dafarra, S. Traversaro, G. Oriolo, L. Rosasco, and D. Pucci, "Adherent: Learning human-like trajectory generatorsfor whole-body control of humanoid robots," *arXiv preprint*, 2021.

[4] S. Qiu, W. Guo, D. Caldwell, and F. Chen, "Exoskeleton online learning and estimation of human walking intention based on dynamical movement primitives," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 1, pp. 67–79, 2020.

[5] B. Berret, E. Chiovetto, F. Nori, and T. Pozzo, "Evidence for composite cost functions in arm movement planning: an inverse optimal control approach," *PLoS Computational Biology*, vol. 7, no. 10, p. e1002183, 2011.

[6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd ed., 2010.

[7] Z. Moghaddam and M. Piccardi, "Training initialization of hidden markov models in human action recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 394–408, 2013.

[8] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, "Understanding human intentions via hidden markov models in autonomous mobile robots," in *Proceedings of the 2008 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Amsterdam, Netherlands), March 2008.

[9] N. Hogan, "An organizing principle for a class of voluntary movements," *Journal of Neuroscience*, vol. 4, no. 11, pp. 2745–2754, 1984.

[10] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[12] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Sendai, Japan), October 2004.

[13] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[14] L. Rapetti, Y. Tirupachuri, K. Darvish, S. Dafarra, G. Nava, C. Latella, and D. Pucci, "Model-based real-time motion tracking using dynamical inverse kinematics," *Algorithms*, vol. 13, no. 10, p. 266, 2020.