Multitask Variational Autoencoding of Human-to-Human Object Handover

Haziq Razali and Yiannis Demiris[1]

***Abstract*— Assistive robots that operate alongside humans require the ability to understand and replicate human behaviours during a handover. A handover is defined as a joint action between two participants in which a giver hands an object over to the receiver. In this paper, we present a method for learning human-to-human handovers observed from motion capture data. Given the giver and receiver pose from a single timestep, and the object label in the form of a word embedding, our Multitask Variational Autoencoder jointly forecasts their pose as well as the orientation of the object held by the giver at handover. Our method is in large contrast to existing works for human pose forecasting that employ deep autoregressive models requiring a sequence of inputs. Furthermore, our method is novel in that it learns both the human pose and object orientation in a joint manner. Experimental results on the publicly available Handover Orientation and Motion Capture Dataset show that our proposed method outperforms the autoregressive baselines for handover pose forecasting by approximately 20% while being on-par for object orientation prediction with a runtime that is 5x faster.** [a]

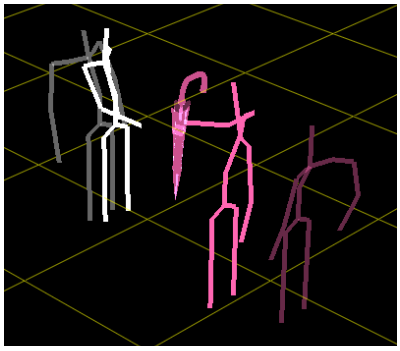


Fig. 1: Given the giver and receiver input poses in transparent pink and white respectively, and the object label, our method learns to forecast both their poses and an appropriate orientation of the object (umbrella) for handover, shown via the brighter colors.

I. Introduction

Understanding how humans interact with one another is essential for the deployment of social or assistive robots that will share the same ecosystem as us. One important task in particular, is the act of handover which is defined as a collaborative action between two participants in which the giver hands an object to the receiver [1]. However, performing a handover is not straightforward due to a number of common sense rules and social conventions that apply during the action. For instance, both the giver and receiver respect personal space during the handover. They avoid standing chest-to-chest and instead, perform the handover at or near arm's length, and at the giver-receiver midpoint [2], [3], [4]. The giver also understands that some objects are more restricted in how they are held or oriented. For instance, a mug should be oriented with its base as parallel to the ground plane as possible in order to avoid spilling its contents. A book in contrast, is far less restricted in how it should be oriented. In order for a robot to operate in the same environment as humans without making the interaction awkward, it needs to model these behaviours and use them to perform handovers.

In this paper, we propose a Multitask Variational Autoencoder for learning human-to-human object handovers using motion capture data. Specifically, our architecture takes as input the giver and receiver pose that we denote together as $P_i \in \mathbb{R}^{Nx3}$ where N indicates the total number of joints, the object label $O_l \in \mathbb{R}^M$ with M classes, and forecasts both their poses P_h and the orientation of the object $O_r \in \mathbb{R}^3$ centered on the giver's grasping hand at handover (Figure I). In short, we seek to learn a model that maximizes $p(P_h, O_r | P_i, O_l)$. To the best of our knowledge, our method is the first that directly forecasts the handover pose and object orientation in a joint manner.

II. Related Works

A. Object Handover

The learning of human-to-human object handovers is a problem that has been extensively studied in the literature [1]. Stralaba et al. [5] studied non-verbal cues such as eye-gaze and body pose leading to the handover. Pan et al. [6] used support vector machines on kinematic features of the human joints to detect the intent to hand over an object. Carfi et al. [7] released a dataset of two humans performing a handover with different objects that varied in weight and size. Chan et al. [8] studied grasp object orientations and later proposed an affordance axes [9] to learn handover configurations by observing two participants performing the handover.

B. Human Pose Forecasting

Advances in deep neural networks for human pose forecasting can also be used for learning human-to-human object handovers. These methods deploy an autoregressive model such as a recurrent [10], [11] or a convolutional sequence-to-sequence [12] model where the output is conditioned on the

[1]The authors are with the Personal Robotics Laboratory, Dept of EEE, Imperial College London, SW7 2AZ, UK {h.bin-razali20, y.demiris}@imperial.ac.uk YD is supported by a Royal Academy of Engineering Chair In Emerging Technologies.

[a]The code is available at www.imperial.ac.uk/personal-robotics/software

history of inputs and predicted poses thus far. These methods however, have focused on modelling the spatiotemporal relation of the human joints without contextual information. Corona et al. [13] later proposed a model that utilized additional information such as object label and location. However, learning handovers via autoregressive models become redundant when there is no need to learn the spatiotemporal relationships of the body joints. In our work, we are mainly interested in learning the posture two individuals take as they perform the handover. Our method is thus most closely related to the above-mentioned works on pose forecasting except that we do not forecast the pose one step at a time. Rather, we directly forecast the giver and receiver pose, and the object orientation when the handover takes place. Furthermore, our method is the first that attempts to jointly model human pose and object orientation at handovers in a multitask data-driven fashion by observing two interacting agents.

C. Multitask learning

Multitask learning aims to increase generalization power by using the domain information contained in the training signals of related tasks as an inductive bias [14]. Multitask models typically have a number of shared layers termed as a base network followed by several task-specific layers that are also known as head networks. Their arrangement is both problem dependent and empirical, with some using a single shared representation as input to the head networks [15], [16], and others, opting for a hierarchical approach in which increasingly dependent tasks are predicted at successively deeper layers [17]. Despite its potential, a caveat that comes with it is the increased training difficulty due to the number of tunable task weights, particularly in the variants with the parallel task heads [15], [16].

III. Method

The goal of our method is to learn a model that maximizes $\log p(P_h, O_r|P_i, O_l)$ where $P_i \in \mathbb{R}^{Nx3}$ denotes the giver and receiver input pose with a total of N joints, $O_l \in \mathbb{R}^M$ the object label with M classes, P_h the giver and receiver pose at handover, and $O_r \in \mathbb{R}^3$ the object orientation at handover. The most widely used method to solve for any multi-output expression resembling the form above is to train a multi-task architecture with multiple head networks to optimize for each corresponding output. However, directly optimizing the above expression results in a deterministic model that is not able to account for the variability of how different individuals perform the handover and in how they handle the same object differently.

A. Multi-Task Variational Autoencoder

Variational Autoencoders (VAEs) [18] provide a solution to this problem by assuming that the data is generated by a low dimensional latent random variable z. This latent variable tends to be a random gaussian noise that when decoded during the forward pass, results in a unique output even if provided the same input. As a result, the model is converted into a stochastic one that allows it to capture the variability in the dataset. We extend the VAE framework into a multi-task setting, having it jointly learn the variability shared between the giver and receiver input poses, and the object label and rotation. Specifically, our goal is to learn a model that maximizes the following:

$$\log p(P_h, O_r, z|P_i, O_l) \tag{1}$$
$$= \log p(P_h, O_r|z, P_i, O_l) + \log p(z|P_i, O_l) \tag{2}$$

that when reformulated as a maximization of the evidence lower bound, results in the following multi-task objective:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{p}} \log p(P_h|z, P_i, O_l) \\ & + \lambda_{\text{o}} \log p(O_r|z, P_i, O_l) \\ & - \lambda_{\text{KL}}\ \text{KL}(q(z|P_h, O_r, P_i, O_l)||N(0,1))\end{aligned} \tag{3}$$

where KL denotes the Kullback-Leibler divergence and the lambdas are used to balance the losses. Intuitively, the above expression tells us that the model is trained to map the set $\{P_h, O_r, P_i, O_l\}$ to a distribution that is likely to produce them such that at test time, the sampled z, together with the set $\{P_i, O_l\}$ can be decoded back to the set $\{P_h, O_r\}$. Note that we chose to use a VAE since it is more thematic with future forecasting as it is able to emulate the fact that there can be multiple handover poses and object orientations at handover. The component used for forecasting can thus be replaced by any deterministic ANN such as an Autoencoder which can be optimized using equation 3 but without the KL term.

B. GloVe Embedded Object Labels

As our model requires as input the object label, a natural question that arises is how we encode these labels into a representation that makes it usable by a neural network. A quick and easy way of doing so is to generate a one-hot encoding that maps object labels to a probability distribution over a discrete set of the n labels. However, this simplicity leads to the curse of dimensionality as it requires a dimension for each new class. Moreover, a one-hot encoding does not inform the model semantics of the object as the vectorization of each unique class results in an orthogonal representation that is equidistant to all others. As such, a model learning to map an object label to a corresponding set of orientation ignores the relationship among labels such as "fork" and "spoon" that can in fact, be exploited during training.

Label embeddings address the above-mentioned limitations by mapping every word in a training corpus into a dense vector with an arbitrary but fixed number of dimensions. The Global Vectors for Word Representation (GloVe) [19] in specific, incorporates co-occurrence statistics of words that frequently appear together within the document. Intuitively, the co-occurrence statistics encode meaning since semantically similar words such as "fork" and "spoon" occur together more frequently than semantically dissimilar words such as "fork" and "camera." The training objective is to then

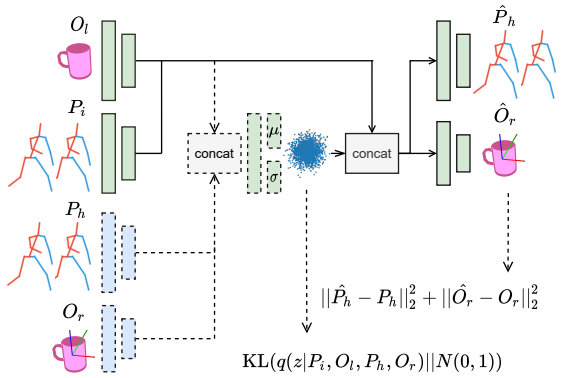


Fig. 2: Overview of our architecture. The training inputs to the network include the coordinates of the giver and receiver input and handover poses P_i and P_h respectively, the object label O_l, and its orientation O_r. The VAE then encodes these set of information while the decoder predicts the set $\{P_h, O_r\}$ given a random gaussian noise and the output from the green encoder blocks. The blocks and arrows with a dotted outline are removed during testing. The colored blocks represent multilayer perceptrons as explained in section IV. The loss functions used for the corresponding components are also shown.

learn word vectors such that their dot product equals the co-occurrence probability of these two words. In our work, we use the pretrained GloVe vectors available at [20].

IV. Architecture

Our architecture is shown in Figure 2. Its inputs are the absolute coordinates of the giver and receiver pose, the object label in the form of a GloVe embedding, and the orientation of the object expressed as rotations about the X, Y, and Z axes. The approximate posterior q is made up of 4 sets of multilayer perceptrons (MLPs) to process the set $\{P_h, O_r, P_i, O_l\}$ where both P_i and P_h are the concatenation of the giver and receiver input and output poses respectively. Their outputs are then concatenated before being sent through another MLP with two heads to produce the parameters of the gaussian distribution. We then sample a random gaussian noise vector and concatenate it to the features representing the set $\{P_i, O_l\}$ before feeding them to 2 parallel head networks to predict $\{P_h, O_r\}$. Note that the blocks and arrows with a dotted outline are removed during testing.

V. Experiments

A. Dataset

We evaluate our method on the publicly available Handover Orientation and Motion Capture Dataset [9] that contains motion capture data recorded during object handovers between humans, as well as the extracted handover orientations for 20 daily objects. The system tracks 21 upper body joints per person but excludes the fingers. The giver selects an item from the table behind and hands it over to the receiver. The giver and receiver moves a maximum of 1.2m and 0.7m respectively and were asked to use only their right hand for the handover. Each video lasts an average of 6 seconds. After cleaning the data and removing sequences that contained missing joints, we are left with approximately 600 recordings with a total of 600k frames from 10 pairs of participants. We train on the first 7 pair of participants and test on the last 3 pair, resulting in a train:test ratio of approximately 70:30.

B. Setup

We compare our approach to three recurrent baselines that we adapt for human pose and object orientation forecasting.

- **Martinez et al.** [10]: This method takes as input a sequence of the concatenated giver and receiver pose, and the object orientations and uses LSTMs to forecast the pose and orientation velocities.
- **Chung et al.** [11]: This variational recurrent model incorporates a VAE into [10] at every timestep.
- **Li et al.** [12]: This method forgoes LSTMs in favour of convolutions. It performs a series of convolutions over the near and distant past of the concatenated data to predict their velocities.

We train the baselines to forecast the next 10 frames given the first 10 as input, each of which is spaced 0.1 seconds apart. At test time, we have these models forecast the maximum length of the sequence and select the handover pose to be at the timestep that has the lowest error between the predicted and ground truth handover pose. Models that contain the VAE are evaluated by randomly sampling from the VAE 256 times for each input that are then averaged at the output. We scale all models such that their number of parameters are on par at approximately 100k. We found this especially important to prevent the autoregressive models from overfitting. Finally, we also evaluate our model without the latent component i.e. a simple ANN / Autoencoder.

All models, including ours, are trained for 50k iterations (approximately 12 hours) using PyTorch's ADAM optimizer [21] with default hyperparameters, a learning rate of 1e-3 and batch size of 64. We then select the epoch with the lowest validation error. We set λ_{KL} to 1 and used grid search to select the optimal set of weights (λ_o, λ_p) for each model where $\lambda_o + \lambda_p = 1$. We preprocess the object orientation by computing it with respect to a reference vector whose X axis points from the receiver's chest to the giver's chest, a Z axis that points upwards from the ground plane, and a Y axis that completes a right-handed frame. We represent the object orientation as a sequence of rotations about the X,Y then the Z axis. We do not preprocess the human joints nor do we augment the data in any way.

C. Quantitative Results

We first present quantitative results for handover pose and object orientation forecasting in Table I where we compute the L1 losses for each of the human joints in metres and the object's rotation about the X, Y and Z axes in radians. We sum and tabulate the errors for the giver and receiver based on their location on the body i.e. the spinal column,

the left half and the right half of the body. Both the VAE and AE versions of our method outperform the baselines in handover pose forecasting by approximately 20% from a summed total of 11.15m to approximately 8.85m over 42 joints, or an average single joint error of 0.2m. They show a definitive improvement in predicting the giver's pose with very comparable results on receiver pose and object orientation forecasting. The results are explainable if we recall from Section V-A that the giver moves a maximum distance of 1.2m while the receiver a maximum distance of 0.7m. Because the errors coming from an autoregressive model are known to accumulate the further predictions are made into the future, we get the characteristic observed in Table 1 with the errors coming from the giver being considerably higher than the receiver. Next, we also observe that the errors for the human poses are also reflective of the dataset's characteristics, specifically in that the methods all obtain lower errors for the right half of the body, the side that was used to perform the handover. Naturally, the errors are lower since the joints on the right half performing the handover are more restricted in their positioning. This shows that our single-timestep input method is able to learn the handover characteristics better than the autoregressive models and is thus a strong alternative if there is no need to learn the motion of the human joints from input to handover.

We next present the rate of inference of the various methods in Table II in Hz. As indicated, our method has a runtime that is at least 5x faster than the baselines. More importantly, because our method does not rely on a sequence of inputs, its runtime remains independent of the total distance taken by the participants to perform the handover. This is in contrast to the autoregressive models whose rate of inference drops drastically in setups that require long-term or finer forecasting. The results here further highlight the efficacy of our method as it allows existing pose estimators to additionally use our system with very little overhead.

Finally, we trained 3 models but with different object label embedding strategies: (i) One-Hot encoding (ii) GloVe embedding and (iii) no object information i.e. the model does not take O_l as input. The results are presented in Table III. It can firstly be seen that all 3 models share the same pose error-profiles in that the errors coming from the giver are greater than the receiver. The numbers also suggest that object information, or lack thereof has no bearing on the model's ability to accurately forecast the giver's pose. For rotation forecasting however, we noted that nearly all the classes have very large variances about their Z-axis evidenced by the error scores. This can be qualitatively observed from the dataset [9] but to provide an example, a person passing a plate tends to orient the flat side parallel to the ground without needing to consider how else it is oriented about the axis that is perpendicular to the ground plane i.e. about its Z-axis. As such, the very low correlation that exists between the object and its Z rotations make it very difficult for learning. It can still be noted however that models trained with object information tend to achieve lower errors. And although the One-Hot encoding performs as well as the GloVe embeddings, we still promote GloVe since it alleviates the curse of dimensionality especially with increasing number of objects.

D. Qualitative Results

We show some qualitative results in Figure 5 where the receiver is shown in white and the giver in pink, in transparent colors for the input and in brighter colors for the predicted handover poses. The ground truth is shown in purple. Our method is able to generate poses that look realistic and that are typically observed during a handover. The predicted poses also obey social norms in that they are not standing too close to each other. Rather, both participants reach out their arms at midpoint to perform the exchange. Note that the results are not exclusive to our model. Rather, it shows that there is no need to use an autoregressive model if the desired output is the pose at handover since it is achievable using only a single input that decreases runtime as evidenced in Table II.

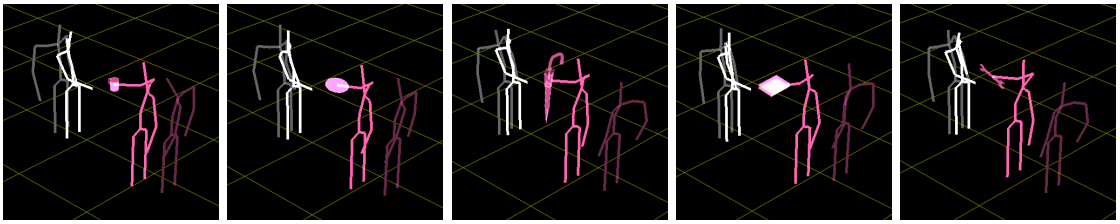
In Figure 6, we get a zoomed in view of the object orientation at handover where the red, green and blue lines denote the object's X, Y and Z axes respectively, and where the thick lines represent the object's true axes for the given object and the thin lines its predictions sampled through multiple runs. We observe some errors although they occur mainly about the Z axis and are not incorrect since the objects are still being handed over appropriately. These figures let us understand the relatively large rotation errors in Tables I and III. They also show the benefit of using a VAE over a simple ANN (AE) i.e. the model learns that there are multiple correct ways of orienting the object although the variation learnt is ultimately limited by how clean the dataset is and the weighting between the KL and L2 loss in equation 3.

E. Importance of Multitask Learning

We study the effect different task weights have on the overall performance of the model in Figure 7 where we varied the pose weight λ_p from 0 to 1 and set the object rotation weight λ_o to $1 - \lambda_p$. The figure illustrates the advantages of multi-task learning as the model shows an improved performance on both tasks at the optimal weight of $\lambda_p = 0.8$ and $\lambda_o = 0.2$, outperforming even their single-task counterparts. It can also be observed that the object rotation errors tend to worsen as we reduce the value of λ_p. This is principally due to the fact that the object orientations were defined with respect to a reference axes that points from the receiver's chest to the giver's chest (Section V-A). As such, a weight combination that deemphasizes the pose errors will result in poorer pose predictions which ultimately results in rotation estimates that are off.

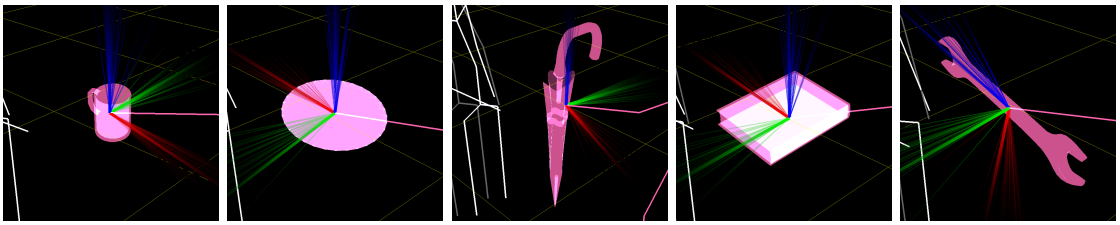
VI. Conclusion

We have presented the first method that jointly models the human pose and object orientation at handover. Our method is an attractive alternative for learning the handover problem when there is no need to model the spatiotemporal



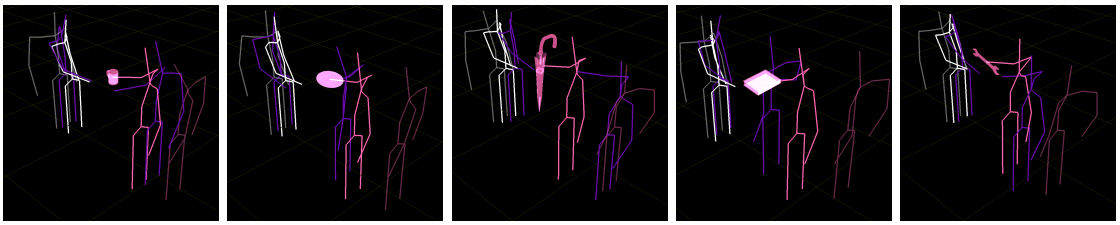
(a) mug (b) plate (c) umbrella (d) book (e) wrench

Fig. 3



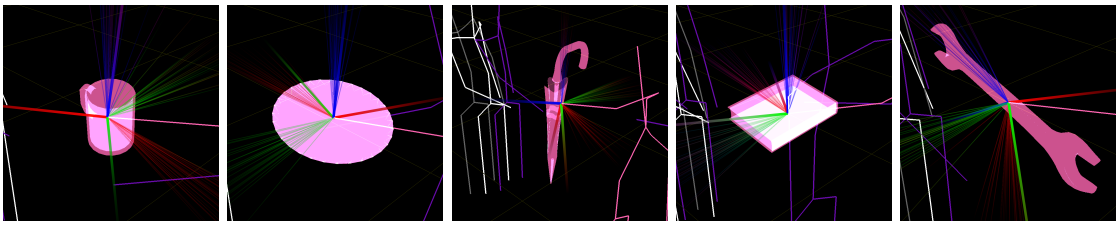
(a) mug (b) plate (c) umbrella (d) book (e) wrench

Fig. 4



(a) mug (b) plate (c) umbrella (d) book (e) wrench

Fig. 5: The giver and receiver input poses are shown in transparent pink and white respectively, their predicted handover poses shown via the brighter colors and the ground truth in purple. Our model has learnt to generate the typical human pose that obeys social norms when giving an object to a receiver. Note that the lower half of the body from the hip downwards have been manually drawn since the dataset only provides the upper half.



(a) mug (b) plate (c) umbrella (d) book (e) wrench

Fig. 6: The object's X, Y and Z true axes are indicated by the thick red, green and blue lines respectively, and the thin lines their predictions sampled through multiple runs. Our method can learn a diverse set of rotation vectors and is important as it lets the learner replicate these behaviours without being too inflexible. We observe some errors although they are not incorrect since the object is still appropriately oriented.

TABLE I: L1 errors in metres for the upper body joints and in radians for the object orientation. Our methods outperform the autoregressive baselines for handover pose forecasting by approximately 20% from a summed total of 11.15m to 8.85m over 42 joints or an average single joint error of 0.2m, with comparable results for orientation forecasting.

	Receiver			Giver			Object		
	Left Half	Spinal Column	Right Half	Left Half	Spinal Column	Right Half	R_x	R_y	R_z
Mean Predictor	2.321	0.811	2.637	3.812	1.342	2.808	1.081	0.431	1.931
Martinez et al. [10]	1.592	0.469	1.540	3.616	1.235	2.924	0.976	0.447	1.757
Chung et al. [11]	1.559	0.452	1.527	3.580	1.200	2.838	0.994	**0.430**	1.798
Li et al. [12]	**1.549**	0.475	**1.513**	4.303	1.391	3.072	1.001	0.457	**1.724**
Ours (VAE)	1.562	**0.445**	1.558	**2.494**	**0.738**	2.054	0.961	0.452	1.749
Ours (AE)	1.554	0.450	1.551	2.495	0.743	**2.045**	**0.960**	0.460	1.781

TABLE II: Inference rates of the various methods in Hz for a batch size of 64. Note that the runtimes for [10], [11], [12] were recorded for a sequence of length 20 that will increase with longer prediction lengths. Our method in contrast, has an inference rate that is fixed.

Martinez et al. [10]	66
Chung et al. [11]	66
Li et al. [12]	100
Ours (VAE / AE)	**500**

TABLE III: L1 errors in metres for the predicted giver and receiver poses and in radians for the object rotation with different types of object information. The results show the slight boost in performance when including object information for learning the appropriate rotations.

	Receiver	Giver	R_x	R_y	R_z
Mean Predictor	5.770	7.962	1.081	0.431	1.931
None	3.581	5.292	1.097	0.499	2.041
One-Hot	**3.559**	**5.271**	0.982	0.484	1.761
GloVe [19]	3.565	5.286	**0.961**	**0.452**	**1.749**

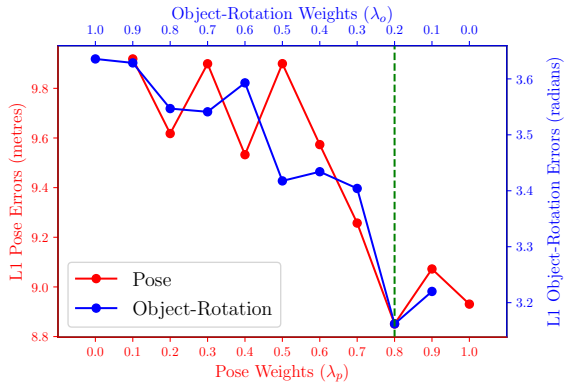


Fig. 7: Performance of the models at various task weights. The green dashed line indicates the optimal weight combination of $\lambda_p = 0.8$ and $\lambda_o = 0.2$ and shows that the joint learning of multiple tasks improves the model's individual task performance.

joint information. Because our method is general in that it learns the direct mapping from the input pose and object label to the output pose and orientations, the architecture selected for forecasting can thus be replaced any existing model. Finally, because the core of our method is a data-driven, stochastic mapping from an observed variable to a distribution over the output, it can thus be used on datasets where either the receiver or the giver stands still without requiring any modification. The former would be more useful in a robot-to-human paradigm [22], [23], [24] where the giver, for example, carries a heavy object over to the receiver or to a person with some form of mobility impairment while the latter in a human-to-robot paradigm [25], [26] where the receiver is for example, a rehabilitative robot that assists the practice of tasks relevant to daily life such as object handover. The robot can then learn how a receiver behaves by observing patient-physiotherapist mocap data, how the physiotherapist orients his hand if the patient struggles to pass the object in an appropriate orientation.

A limitation of our method is that because it is data-driven, it cannot generalize to samples where the giver and receiver are standing at distances far greater or where the object classes are completely different than what is observed in the dataset. Furthermore, the large variance in object rotations and small number of classes also make it difficult for our model to generalize to unseen classes. As such, in future works, it would be beneficial to create a dataset - even a synthetic one - where the participants are instructed on the appropriate handover orientations in order to minimize variance and ease learning. It would also be beneficial to have the full body pose including the positions of the finger joints as it would allow our model to learn how objects are grasped in order to make the handover as natural as possible.

REFERENCES

[1] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object handovers: a review for robotics," *IEEE Transactions on Robotics*, 2021.

[2] M. Huber, A. Knoll, T. Brandt, and S. Glasauer, "Handing over a cube," *Annals of the New York Academy of Sciences*, vol. 1164, no. 1, pp. 380–382, 2009.

[3] P. Basili, M. Huber, T. Brandt, S. Hirche, and S. Glasauer, "Investigating human-human approach and hand-over," in *Human centered robot systems*. Springer, 2009, pp. 151–160.

[4] C. Hansen, P. Arambel, K. Ben Mansour, V. Perdereau, and F. Marin, "Human–human handover tasks and how distance and object mass matter," *Perceptual and motor skills*, vol. 124, no. 1, pp. 182–199, 2017.

[5] K. Strabala, M. K. Lee, A. Dragan, J. Forlizzi, S. S. Srinivasa, M. Cakmak, and V. Micelli, "Toward seamless human-robot handovers," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 112–132, 2013.

[6] M. K. Pan, V. Skjervøy, W. P. Chan, M. Inaba, and E. A. Croft, "Automated detection of handovers using kinematic features," *The International Journal of Robotics Research*, vol. 36, no. 5-7, pp. 721–738, 2017.

[7] A. Carfì, F. Foglino, B. Bruno, and F. Mastrogiovanni, "A multi-sensor dataset of human-human handover," *Data in brief*, vol. 22, pp. 109–117, 2019.

[8] W. P. Chan, M. K. Pan, E. A. Croft, and M. Inaba, "Characterization of handover orientations used by humans for efficient robot to human handovers," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1–6.

[9] W. P. Chan, M. K. Pan, E. A. Croft, and M. v. Inaba, "An affordance and distance minimization based method for computing object orientations for robot human handovers," *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[10] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[11] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2980–2988, 2015.

[12] C. Li, Z. Zhang, W. Sun Lee, and G. Hee Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5226–5234.

[13] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-aware human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6992–7001.

[14] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[16] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 977–11 986.

[17] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 270–287.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] Pretrained glove vectors. (Date last accessed 04-July-2021). [Online]. Available: https://nlp.stanford.edu/projects/glove/

[21] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[22] J. R. Medina, F. Duvallet, M. Karnam, and A. Billard, "A human-inspired controller for fluid human-robot handovers," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 324–331.

[23] A. Bestick, R. Bajcsy, and A. D. Dragan, "Implicitly assisting humans to choose good grasps in robot to human handovers," in *International Symposium on Experimental Robotics*. Springer, 2016, pp. 341–354.

[24] L. Peternel, W. Kim, J. Babič, and A. Ajoudani, "Towards ergonomic control of human-robot co-manipulation and handover," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 55–60.

[25] K. Yamane, M. Revfi, and T. Asfour, "Synthesizing object receiving motions of humanoid robots with human motion database," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1629–1636.

[26] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan, V. Ortenzi, P. Corke, and M. Grafinger, "Object-independent human-to-robot handovers using real time robotic vision," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 17–23, 2020.