Supplementary Material: Intention-based Long-Term Human Motion Anticipation

Julian Tanke*, Chintan Zaveri*, Juergen Gall
University of Bonn
{tanke|gall}@iai.uni-bonn.de

1. Implementation Details

The encoder e_l and decoder d_l in Figure 3 of the paper are both represented as single-layer GRUs with 16 hidden units. The decoder has an additional softmax layer to generate class probabilities. D_{label} is a feed-forward neural network with a single hidden layer with 32 units, a ReLU non-linearity, and a single unit sigmoid output. The input to D_{label} are the predicted output labels of d_l, stacked for 25 frames.

The pose encoder e_p and decoder d_p are represented as GRUs where e_p is a single-layer GRU with 512 hidden units and where d_p is a three-layer GRU with 512 hidden units and dropout rate 0.3. The encoder output h_p is passed as hidden state to the first decoder layer, while the remaining two layers are initialized with a zero hidden state. The noise state z is concatenated to all three hidden states before passing to the decoder. The pose discriminator D_{pose} is represented as one hidden layer feed-forward neural network with 512 hidden units, ReLU non-linearity and a single unit sigmoid output. The input to D_{pose} are the predicted output labels of d_p stacked for 25 frames. Each pose has dimension 54. The implementation of the NDMS metric and the source code for the approach are available at `https://github.com/jutanke/human_motion_ndms`.

2. Clustering

Figure 1 illustrates the difference between naive k-means clustering and the proposed clustering. While k-means generates small clusters, we greedily merge cycles of short clusters into larger clusters as shown in Figure 1. We do so by first segmenting the single poses using k-means and then detect cycles in the cluster ids for each sequence. We then greedily merge the most frequent occurring cycles. We start with 14 clusters and merge the clusters until we have only 8 clusters left.

The impact of the number of clusters is shown in Figure 2. The approach is not very sensitive to the number of clusters, but the NDMS score decreases when the number of clusters is too large. In this case, the clusters become too fine-grained.

*equal contribution

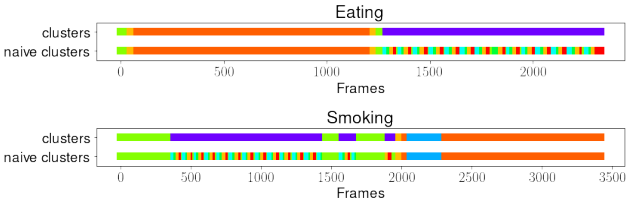

Figure 1: Comparing proposed clustering with naive k-means clustering on Human3.6M [9] for test actor S5. The cluster centers are obtained from the training set only.

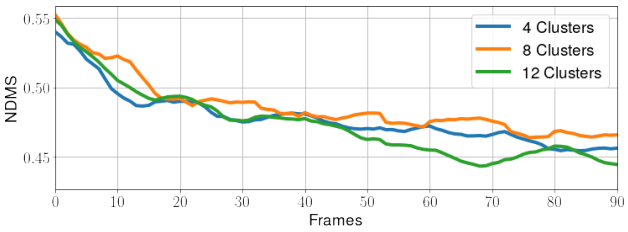

Figure 2: Impact of different numbers of clusters for human pose forecasting.

In Table 1, we evaluate the frame-wise forecast accuracy of the intention label. We evaluate on test actor S5 for all 15 actions.

3. Evaluation Score

3.1. NDMS vs. Euclidean Distance/Velocity

For our evaluation score, we propose NDMS since other measures like L2 distance or L2 velocity distance are insufficient. To show this, we take two sequences, one containing real ground-truth motion from the training set and one containing a static pose (zero velocity) (see Figure 3).

Action	Accuracy
Directions	0.789
Discussion	0.701
Eating	0.574
Greeting	0.429
Phoning	0.771
Posing	0.45
Purchases	0.592
Sitting	0.464
Sitting Down	0.326
Smoking	0.54
Taking Photo	0.157
Walking	1.0
Walking Dog	0.456
Waling Together	0.433
Average	0.528

Table 1: Frame-wise accuracy of label forecasting for 100 frames on all 15 actions on Human3.6M [9] for 8 intention labels.

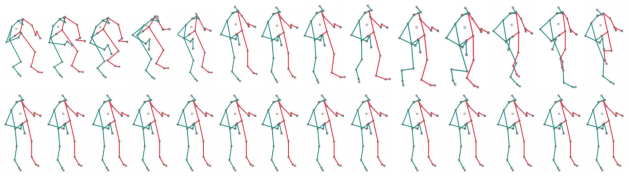

Figure 3: Real motion (top row) vs. zero velocity (bottom row) of around two seconds. Zero velocity is very unrealistic as it always produces the same pose as output [13].

We use the scores to measure the plausibility of a walking motion using the walking sequences of the test data as reference. While the real motion should have a high score or low distance, the zero velocity sequence should perform poorly since it is not a walking motion.

Our results are summarized in Figure 4. We observe that NDMS (a) scores the real motion high and the zero velocity very low, as it should be the case. Note that for the distances in (b) and (c) lower values are better, while for (a) higher values are better. For the Euclidean distance (b) and the mean squared error over velocity (c), the zero velocity performs better than the real motion. This shows that neither the L2 distance nor the mean squared error over velocity are useful metrics to measure the plausibility of a sequence.

3.2. NDMS vs. NPSS

The Pearson correlation coefficient with the user study and NDMS is 0.901. This shows that the proposed measure highly correlates with human perception. The correlation coefficient for NPSS [3] is -0.238. The negative correlation is due to the competitive NPSS of Grammar [14] although the generated motions are perceived as unrealistic by humans.

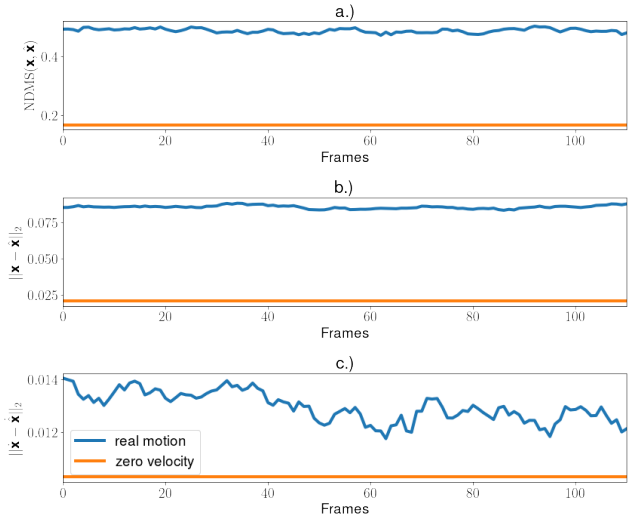

Figure 4: Comparing baseline distances with NDMS: a.), b.) and c.) show real motion (blue) and zero velocity prediction (orange) for NDMS (higher is better), L2 distance (lower is better) and L2 velocity distance (lower is better), respectively. While NDMS scores the real motion higher, the L2 and L2 velocity distance would rate the static pose as more plausible than the real motion.

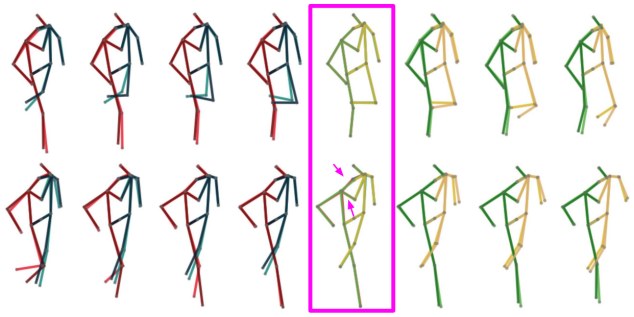

Figure 5: DLow [19] generates a motion discontinuity between the observed poses (blue-red skeletons) and forecast poses (green-yellow skeletons). For each frame, we show the current pose and the previous pose, with the current pose in darker shades and the previous one in lighter shades. At the pink box, the method transitions from the observed poses to forecast poses. We observe that the first predicted pose is very similar but not identical to the last input pose, which is a common motion artifact in human motion anticipation [13]. The motion of the left leg is still too small in the following frame.

3.3. NDMS vs. ADE, FDE, MMADE and MMFDE

DLow [19] proposes ADE, FDE, MMADE and MMFDE to evaluate the quality of multi-modal human motion pre-

Method	APD ↑	ADE ↓	FDE ↓	. MMADE ↓	MMFDE ↓	NDMS ↑
[15]	6.723	0.461	0.560	0.522	0.569	-
[17]	0.403	0.457	0.595	0.716	0.883	-
[4]	7.214	0.858	0.867	0.847	0.858	-
[5]	6.265	0.448	0.533	0.514	0.544	-
[6]	6.769	0.461	0.555	0.524	0.566	-
[8]	6.509	0.483	0.534	0.520	0.545	-
[18]	9.330	0.493	0.592	0.550	0.599	0.294
[19]	**11.741**	0.425	**0.518**	**0.495**	**0.531**	0.311
Ours	3.477	**0.413**	0.631	0.662	0.770	**0.366**
Baseline	*16.418*	*0.429*	*0.451*	*0.520*	*0.478*	0.166

Table 2: Evaluation over 2 seconds for multi-modal human motion anticipation on Human3.6M [9] as defined in DLow [19]. The motion is represented as 3D skeletons centered at the origin but with global rotation.

	walking					eating					smoking					discussion				
milliseconds:	80	160	320	400	560	80	160	320	400	560	80	160	320	400	560	80	160	320	400	560
Zero Velocity [13]	0.39	0.86	0.99	1.15	1.35	0.27	0.48	0.73	0.86	1.04	0.26	0.48	0.97	0.95	1.02	0.31	0.67	0.94	1.04	1.41
Seq2Seq [13]	0.28	0.49	0.72	0.81	0.93	0.23	0.39	0.62	0.76	0.95	0.33	0.61	1.05	1.15	1.25	0.31	0.68	1.01	1.09	1.43
AGED [7]	0.22	0.36	0.55	0.67	0.78	0.17	**0.28**	0.51	0.64	0.86	0.27	0.43	0.82	0.84	1.06	0.27	0.56	**0.76**	**0.83**	1.25
Imitation [16]	0.21	0.34	0.53	0.59	0.67	0.17	0.30	0.52	0.65	0.79	0.23	0.44	0.86	0.85	0.95	0.27	0.56	0.82	0.91	1.34
ConvSeq2Seq [10]	0.33	0.54	0.68	0.73	-	0.22	0.36	0.58	0.71	-	0.26	0.49	0.96	0.92	-	0.32	0.67	0.94	1.01	-
Trajectory [12]	**0.18**	**0.31**	**0.49**	**0.56**	0.65	**0.16**	0.29	0.50	0.62	0.76	0.22	**0.41**	0.86	0.80	0.87	**0.20**	0.51	0.77	0.85	1.33
Grammar [14]	0.26	0.44	0.67	0.77	0.84	0.20	0.34	0.54	0.68	0.85	0.27	0.50	0.92	0.90	1.00	0.30	0.65	0.92	1.00	1.37
Mix&Match [3]	0.33	0.48	0.56	0.58	**0.64**	0.23	0.34	**0.41**	**0.50**	**0.61**	0.23	0.42	0.79	**0.77**	**0.82**	0.25	0.60	0.83	0.89	**1.12**
DLow [19]*	0.31	0.42	0.53	0.75	0.83	0.24	0.32	0.44	0.55	0.77	**0.21**	0.43	0.80	0.79	0.97	0.31	0.55	0.80	0.88	1.15
Ours	0.23	0.42	0.73	0.83	0.89	0.17	0.33	0.66	0.85	1.02	0.29	0.50	**0.72**	0.78	0.83	0.27	**0.47**	0.82	1.07	1.31

Table 3: Mean Angular Error on Human3.6M [9]. *from [2]

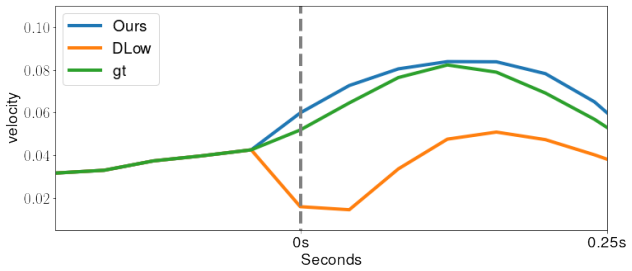

Figure 6: Frame-wise velocity of the left foot for DLow [19], our method and the ground-truth sequence for a walking sequence. The gray dotted vertical line represents the transition from input sequence to forecast sequence. Results for our method as well as for DLow are averaged over 50 samples. Our method clearly follows the ground-truth motion while DLow suffers from the discontinuity.

diction. In contrast to other state-of-the-art methods they utilize 3D skeletons rather than an angular representation. We train and evaluate our method using the data and evaluation code from [19]. The results in Table 2 show that DLow has a higher diversity while our approach has a lower ADE. For the other metrics FDE, MMADE, and MMFDE, DLow performs better. This contradicts the results from the user study where in particular walking sequences generated from DLow are considered as unrealistic. This is due to the motion discontinuity between the observed poses and the forecast poses as shown in Figure 5 and 6, but also due to the very high diversity of the generated sequences. The forecast sequences quickly generate motions that are very unlikely to occur after a walking motion. While these issues are not measured by ADE, FDE, MMADE and MMFDE, NDMS penalizes motion discontinuities. In Figure 7, we plot NDMS over time. As can be seen, NDMS drops for DLow very quickly and increases after 8 frames. This is due to the discontinuity between observed frames and forecast frames.

In order to show that the measures ADE, FDE, MMADE and MMFDE can be easily fooled, we construct a very unrealistic multi-modal baseline. In order to generate 50 samples for a single observation, we generate 50 sequences with static poses (zero velocity) as shown in Figure 3. To this end, we cluster the poses of the training data to obtain 48 clusters. For each cluster, we take the mean pose as static pose. Note that these 48 sequences are independent of the observation, but they generate a very high diversity (APD). For the remaining two sequences, we take the last pose of the observation and the mean pose of the observed sequence, respectively, as static pose. As shown in Table 2, this baseline performs very well for APD, ADE, FDE, MMADE and MMFDE although none of the 50 sequences contains any motion and all of them are highly unrealistic.

	Basketball					Basketball Signal					Directing Traffic					Jumping				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Seq2Seq [13]	0.50	0.80	1.27	1.45	1.78	0.41	0.76	1.32	1.54	2.15	0.33	0.59	0.93	1.10	2.05	0.56	0.88	1.77	2.02	2.4
convSeq2Seq [10]	0.37	0.62	1.07	1.18	1.95	0.32	0.59	1.04	1.24	1.96	0.25	0.56	0.89	1.00	2.04	0.39	0.6	1.36	1.56	2.01
Trajectory [12]	**0.33**	**0.52**	**0.89**	**1.06**	**1.71**	**0.11**	**0.20**	**0.41**	**0.53**	**1.00**	**0.15**	**0.32**	**0.52**	**0.60**	2.00	**0.31**	**0.49**	**1.23**	**1.39**	**1.80**
Ours	0.41	0.66	1.15	1.38	2.05	0.30	0.56	0.97	1.12	1.56	0.27	0.48	0.78	0.91	**1.50**	0.84	0.87	1.43	1.65	2.04
	Soccer					Walking					Washwindow					Average				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Seq2Seq [13]	0.29	0.51	0.88	0.99	1.72	0.35	0.47	0.60	0.65	0.88	0.30	0.46	0.72	0.91	1.36	0.38	0.62	1.02	1.18	1.67
convSeq2Seq [10]	0.26	0.44	0.75	0.87	1.56	0.35	0.44	**0.45**	**0.50**	0.78	0.30	0.47	0.80	1.01	1.39	0.32	0.52	0.86	0.99	1.55
Trajectory [12]	**0.18**	**0.29**	0.61	**0.71**	1.40	0.33	0.45	0.49	0.53	**0.61**	0.22	**0.33**	**0.57**	0.75	1.20	**0.25**	**0.39**	**0.68**	**0.79**	**1.33**
Ours	0.25	0.42	**0.60**	0.79	**1.06**	**0.26**	**0.41**	0.49	0.53	0.71	**0.21**	**0.33**	0.61	**0.74**	**1.18**	0.35	0.52	0.83	0.96	**1.33**

Table 4: Mean Angular Error on CMU Mocap [1].

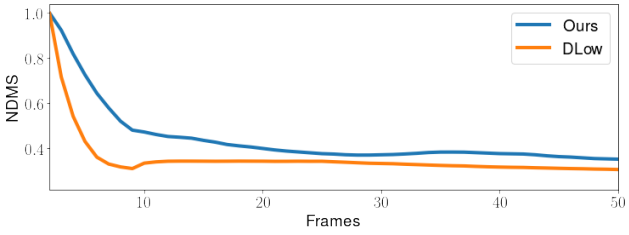

Figure 7: NDMS score for 50 samples per input sequence on Human3.6M [9] averaged over 15 actions. The first few frames exhibit very high scores as the motion words contain mostly poses from the observed sequence at the beginning, as described in Section 4 of our paper. We observe that DLow has a much sharper decline and a small dent due to the motion discontinuity between observed and forecast frames.

In contrast, the low NDMS score reliably indicates that the baseline does not generate plausible sequences.

3.4. NDMS vs. Inception Score

For evaluating long-term human motion forecasting we report the inception score (IS) as described in [3]. Since the original scoring model is not available, we followed the description [3] and re-trained a skeleton-based action classifier [11]. We pass then sequences with 16 observation frames and 60 prediction frames to the scoring model. For methods that forecast multiple sequences for one observation sequence, we generate 50 samples for each single input sequence and calculate the mean inception score as well as the standard deviation, following [3]. For the other methods, we compute the inception score of a single forecast sequence.

The results can be seen in Table 5. First, we validate that our newly trained *inception* network works properly by comparing our results of [3] with the results reported in their work. The reproduced result is even slightly better. While our approach outperforms all methods that generate multiple future sequences [17, 15, 4, 3], we observe

Method	IS
Seq2Seq [13]	7.5 ± 0
Trajectory [12]	9.2 ± 0
Grammar [14]	$\mathbf{10.3} \pm 0$
Yan et al. [17] $\star$	1.9 ± 0.4
Walker et al. [15] $\star$	1.8 ± 0.6
Barsoum et al. [4] $\star$	2.1 ± 1.3
Mix&Match [3] $\star$	7.3 ± 1.4
Mix&Match [3]	7.5 ± 1.1
Ours	$\underline{9.7} \pm 0.6$

Table 5: Inception score as described in [3] for Human3.6M. $\star$ denotes results reported in [3].

that Grammar [14] achieves a higher inception score. However, our user study shows that poses generated by Grammar are less realistic than the sequences that are generated by Mix&Match [3] or our approach. This indicates that the inception score is not a very reliable measure for the plausibility of the forecast human motion. The generative grammars [14] achieve a very high inception score, as can be seen in Table 5. The high score, however, is not supported by our user study and the qualitative results. This shows the weakness of the inception score, which does not occur for the proposed NDMS score.

4. Short-Term Forecasting

To evaluate short-term motion prediction, we follow the same protocol as described in [3]. Table 3 and Table 4 detail our short-term forecasting results on Human3.6M [9] and on CMU Mocap [1], respectively. Even though our main objective is long-term and not short-term human motion anticipation, our approach achieves competitive results.

Acknowledgment

The work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/8-1 and GA 1927/5-2 (FOR 2535 Anticipating Human Behavior).

References

[1] CMU. Carnegie-Mellon Mocap Database.

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. *arXiv preprint arXiv:1912.08521*, 2020.

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[4] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[5] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[6] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[7] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision*, 2018.

[8] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Pattern Analysis and Machine Intelligence*, 2014.

[10] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *International Joint Conferences on Artificial Intelligence*, 2018.

[12] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. *International Conference on Compuver Vision*, 2019.

[13] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[14] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S Ryoo. Adversarial generative grammars for human activity prediction. *European Conference on Computer Vision*, 2020.

[15] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision*, 2017.

[16] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *International Conference on Computer Vision*, 2019.

[17] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision*, 2018.

[18] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *International Conference on Learning Representations*, 2020.

[19] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, 2020.