A Neural Temporal Model for Human Motion Prediction - Supplementary

Anand Gopalakrishnan[1], Ankur Mali[1], Dan Kifer[1], C. Lee Giles[1], Alexander Ororbia[2]
Pennsylvania State University, University Park, PA, 16801[1]
Rochester Institute of Technology, Rochester, NY, 14623[2]
{aug440, aam35, duk17, clg20}@psu.edu[1], ago@cs.rit.edu[2]

1. Ablation Experiments

We conducted an ablation study to determine the value of each component of our overall neural architecture for long-term motion synthesis. The components investigated were: 1) the two-level processing mechanism, 2) the integration of the finite-difference motion derivative approximation features, and 3) the multi-objective cost function used to guide parameter optimization.

VGRU-d in Tables 2, 3 refers to our full two-level processing network (i.e. VTLN-RNN) with derivatives appended and trained using our multi-objective loss. Both the RNNs in the two-level system contain a single layer of 512 GRU units as described in section 4.1 (paragraph 2 of the main paper, which describes long-term motion model training). Dropout [4] with a probability of 0.3 was applied only to the Body-RNN (as shown in Figure 1). GRU-d in Tables 2, 3 refers to a regular two-layer network (512 GRU units/cells in each layer) trained with the proposed loss and derivatives appended. Dropout [4] with probability of 0.3 was applied to both layers of the *GRU-d* model.

VGRU-d + no-loss in Tables 2, 3 refers to a system that is identical to the VGRU-d system (described above) but trained without the proposed multi-objective loss. During the training phase, when predicting the data at time $t+1$, the ground-truth at time t was fed in as input to the model while at test-time, the model's own output at time t was used instead, i.e., standard Teacher Forcing. The rest of the training setup was identical to that described for the *VGRU-d* model. *GRU + no-d* in Tables 2, 3 refers to a system that was identical to that of *GRU-d* except that the finite difference approximations of the $\{1, 2, 3\}$-order derivatives, as described in section 3.3 of the paper, were not appended to the input vector.

Looking at the NPSS results in Table 3, we can see that the VGRU-d model, with all of the proposed components, achieves the lowest score on 2 out of 4 actions, e.g., walking and smoking, and with a score that is quite close for the act of eating. Discussion itself is a highly aperiodic and extremely difficult-to-model action, especially when only pure joint angle information is exclusively used, which was also noted in prior work [2]. The *GRU-d* achieves the overall second best performance across all 4 actions. Furthermore, when the approximate derivative features are dropped, the performance of the *GRU + no-d* drops significantly across all 4 actions. This indicates that approximate joint angle derivatives play a crucial role in guiding the model to producing smooth, realistic plausible (long-term) motion trajectories (with the added benefit that these finite-difference equations are parameter-free and thus readily/easily calculated). Lastly, observe that for the *VGRU-d + no-loss* in Table 3 there is a drastic drop in performance on periodic actions like walking and smoking when compared to aperiodic actions like discussion. Interestingly enough, for discussion, this model ablation achieves the best NPSS score. This possibly indicates that although progress has been made highly aperiodic actions such as discussion where there are no cyclic or obvious cues before the movement of a hand or a leg, purely using motion capture data alone is not a complete solution. Audio-visual information of the surroundings in such cases can give important information in such cases to help the model get a complete picture of the actor and his actions.

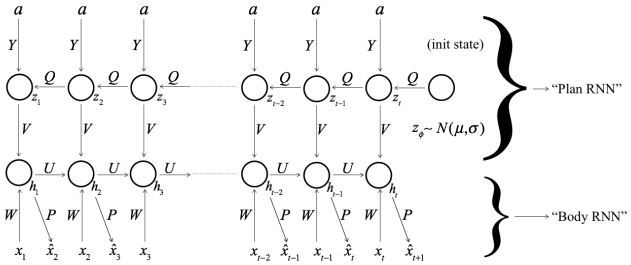

Figure 1. VTLN-RNN architecture

2. Additional User Study Analysis

We conducted further analysis of our user study results to further support our claim that the proposed NPSS evaluation

metric correlates more strongly with human judgment over MSE for specific timeslices $\{80, 160, 320, 400, 560, 1000\}$ milliseconds (used previously by [2], [1], [3]) as well as the sum of MSE scores over all timeslices. We have 80 pairs of generated sequences (and, for each pair, a ground truth sequence). For every comparison between two generated sequences A and B, we have 20 human judgements determining which one is closer to the ground truth sequence.

A good error measure should correlate well with human judgment as follows: if the vast majority of users prefer Sequence A over Sequence B, then the error of B should be much higher than that of A. On the other hand, if the preference is almost evenly split, then Sequences A and B should have similar error.

We test this kind of correlation as follows: For each A/B comparison, we define the *Disapproval* of A over B, denoted by *Disapproval*(A,B), to be the fraction of subjects who preferred B minus the fraction of subjects who preferred A. Thus if A is much worse than B, *Disapproval*(A,B) will be close to 1 and, if A and B are equally good, *Disapproval*(A,B) will be 0. If A is much better than B, then the *Disapproval*(A,B) will be close to -1.

For each A/B comparison, we can also compute

$$\text{NPSS(A)} - \text{NPSS(B)} \tag{1}$$

$$\text{MSE}_1\text{(A)} - \text{MSE}_1\text{(B)} \tag{2}$$

$$\text{MSE}_2\text{(A)} - \text{MSE}_2\text{(B)} \tag{3}$$

where, NPSS(A) is the NPSS error for sequence A with respect to the ground truth, $\text{MSE}_1(A)$ is the sum of MSE scores (with respect to ground truth) over timeslices $= \{80, 160, 320, 400, 560, 1000\}$ and $\text{MSE}_2(A)$ is the sum of MSE scores over all of the time slices. The reason for two MSE calculcations is that prior work only evaluated MSE at select time slices (so MSE_1 was also computed to ensure consistency with prior work).

For each Equation (Eq. 1, 2, or 3), strong positive correlation means they strongly agree with human judgment while a correlation close to 0 means they do not appear to be related to human judgment. We use Spearman's Rank Correlation for this task. We report both the correlation coefficient and the p-value. The p-value is designed to test the null hypothesis that the correlation is 0. A low p-value indicates that there is evidence against the null hypothesis, or, in other words, a low p-value indicates that there is a correlation and that it is statistically significant. Typically, statistical significance is claimed when the p-value is less than 0.01. We show the results of our correlation test in Table 1.

The analysis we conducted shows that our proposed NPSS has a reasonably large, positive correlation and a very small p-value, meaning that it is strongly correlated with human judgment and the correlation found is statistically significant. Meanwhile MSE_1 empirically shows a small negative correlation and has a relatively high p-value which means that it is still possible for it to be completely unrelated to human judgment (which would be consistent with the observations made in prior work). When the MSE is computed across all time slices (i.e. MSE_2) the empirical correlation is even closer to 0.

	MSE_1 (Eq. 2)	MSE_2 (Eq. 3)	NPSS (Eq. 1)
Correlation	-0.143	-0.0638	0.5635
p-value	0.2049	0.5738	5.23×10^{-8}

Table 1: Spearman Correlation Results.

Overall, especially when the analysis of this supplementary material is included, our user study strongly suggests that the proposed NPSS metric is a much more suitable quantitative metric for evaluating motion sequences generated by statistical motion-synthesis models.

	Walking						Eating						Smoking						Discussion					
models	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
GRU + no-d	1.410	1.436	1.412	1.419	1.471	1.541	1.318	1.366	1.459	1.531	1.627	1.771	2.108	2.215	2.327	2.382	2.452	2.614	1.847	2.095	2.083	1.989	2.186	2.144
VGRU-d + no-loss	1.210	1.294	1.408	1.424	1.477	1.550	1.139	1.230	1.346	1.427	1.503	1.635	1.689	1.930	2.273	2.350	2.433	2.533	1.499	1.837	1.974	1.970	2.327	2.507
	± 1e-5	± 1e-5	± 2e-5	± 5e-5	± 5e-5	± 6e-5	± 1e-5	± 1e-5	± 1e-5	± 2e-5	± 4e-5	± 4e-5	± 1e-4	± 1e-4	± 3e-4	± 3e-4	± 4e-4	± 2e-4	± 1e-5	± 1e-5	± 1e-5	± 1e-5	± 1e-5	± 1e-5
GRU-d	1.311	1.333	1.369	1.364	1.350	1.370	1.275	1.305	1.386	1.466	1.530	1.702	1.943	2.062	2.201	2.255	2.342	2.486	1.744	1.980	2.026	1.994	2.214	2.172
VGRU-d	1.108	1.146	1.211	1.200	1.220	1.280	1.090	1.160	1.240	1.330	1.370	1.560	1.670	1.800	1.940	1.980	2.060	2.320	1.749	2.037	2.011	1.868	2.088	2.318
	± 1e-4	± 1e-4	± 2e-4	± 2e-4	± 3e-4	± 2e-4	± 2e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4	± 1e-4

Table 2: Ablation study on long-term motion synthesis models. The MSE of euler angles on test set sequences is shown.

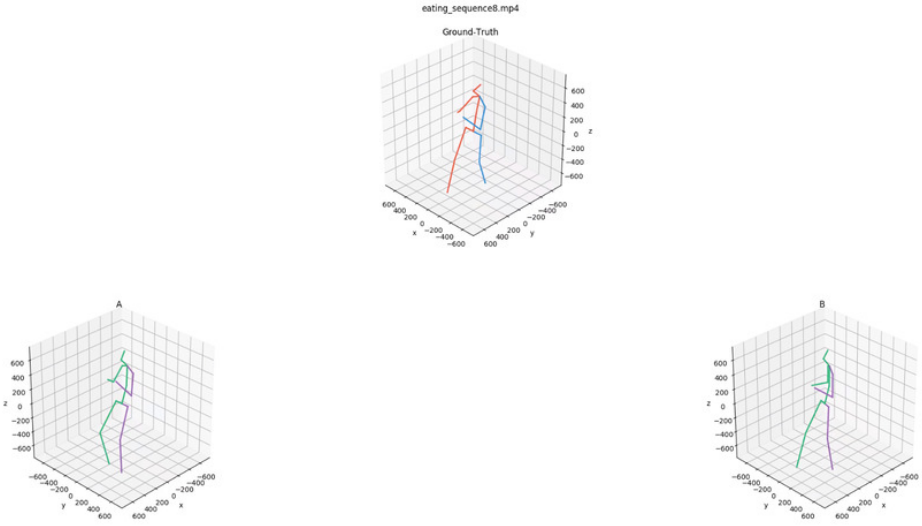

Figure 2. A sample screenshot taken from the user survey.

Models	Walking	Eating	Smoking	Discussion
GRU + no-d	1.138	1.147	1.443	1.812
VGRU-d + no-loss	1.541	0.911	1.474	**1.621**
GRU-d	0.931	**0.836**	1.274	1.688
VGRU-d	**0.887**	0.846	**1.235**	1.777

Table 3: Test-set NPSS scores for ablation study models (lower is better).

References

[1] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[2] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[3] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683. IEEE, 2017.

[4] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014.