Comparison of Spatio-Temporal Models for Human Motion and Pose Forecasting in Face-to-Face Interaction Scenarios

Supplementary Material

German Barquero GBARQUGA9@ALUMNES.UB.EDU
Johnny Núñez JNUNEZCA11@ALUMNES.UB.EDU
Universitat de Barcelona and Computer Vision Center, Spain

Zhen Xu XUZHEN@4PARADIGM.COM
4Paradigm, Beijing, China

Sergio Escalera SERGIO@MAIA.UB.ES
Universitat de Barcelona and Computer Vision Center, Spain

Wei-Wei Tu TUWEIWEI@4PARADIGM.COM
4Paradigm, Beijing, China

Isabelle Guyon GUYON@CHALEARN.ORG
LISN (CNRS/INRIA) Université Paris-Saclay, France, and ChaLearn, USA

Cristina Palmero CRPALMEC7@ALUMNES.UB.EDU
Universitat de Barcelona and Computer Vision Center, Spain

1. Architecture details

In all architectures tested, a first embedding layer transformed the input coordinates and offsets from all landmarks into an intermediate representation vector of size 512 by means of a dense layer. For bimodal architectures, the metadata, audio, and transcriptions representations were also embedded through a dense layer into representations of sizes 16, 64, and 64, respectively. All non-linearities used after convolutional or dense layer were leaky ReLUs with a negative slope of 0.01.

1.1. Seq2Seq

Both the encoder and the decoder consisted of one-layer LSTM or GRU units with hidden and cell states of size 1024. The hidden states from the encoder were used to initialize the decoder's. Two dense layers of 1024 units transformed the output from the decoder to the predicted skeleton pose. A residual layer was added from the input to the output of the decoder, so that this predicts the offsets of movement between future frames.

1.2. TCN

Causal dilated temporal convolutions without padding transformed the embeddings of the sequence of N observed skeletons to a single vector whose receptive field was the whole observed window of skeletons. This was done by means of five sequential blocks of two temporal convolution: one with dilation factor D, and kernel size K (no padding), and another with dilation factor 1 and kernel size 1 (no padding). The K value was set to 2 for all blocks, and the D was set to 1, 3, 9, 27, and 59 for the five blocks, respectively. The decoder consisted of a one-layer GRU or LSTM unit followed by two dense layers with 1024 and 512 units, respectively. The same residual layer as the one used in the Seq2Seq method was implemented.

1.3. STGNN

A total of three blocks were used. Each block consisted of a temporal layer and spatial layer. In the temporal layer, TCN with inception-like kernel set to 2, 3, 9, 11 was used. In the spatial layer, a mixhop graph neural network of maximum neighbour order 2 was used. The graph structure was parametrized by an embedding of dimension 40. A residual connection was used. A final skip convolution connecting each output of TCN, raw input, output was implemented. The TCN channel, residual channel, graph channel, and skip channel were of sizes 32, 32, 32, 64, respectively. A final MLP converted the intermediate vector of size 128 channels into the final output.

1.4. Transformer

For the spatio-temporal model, the backbone from Zheng et al. (2021) was used. The depth of the transformer was set to 4, the number of heads to 8, and the dropout value for the stochastic depth was set to 0.2. No dropout was applied for the attention mechanism. The output of the learnt weighted averaging layer was used as initial hidden state of a one-layered GRU decoder unit. The details of the GRU decoder are equal to those from the TCN model. For the temporal model, the spatial attention was removed.

2. Unimodal results for face, body, and hands

In this section, we present the evaluation of the unimodal architectures with metrics computed for each part of the body, see Tables 1, 2 and 3. We can observe a significant higher accuracy for face and body behavior forecasting than for hands, which move with considerably faster motion (Δ=0.31/0.65/1.33 px/frame for body/face/hands, in average).

References

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021.

	MPJPE	ST	MT	LT	FDE	Δ	Δ_{ST}	Δ_{MT}	Δ_{LT}
Ground truth	-	-	-	-	-	0.65	0.62	0.66	0.67
Zero-velocity	13.29	4.35	11.99	17.65	18.80	0.00	0.00	0.00	0.00
LinearProp	25.41	4.67	18.48	37.86	48.22	0.54	0.54	0.54	0.54
RTOMean	13.63	4.68	12.41	17.95	19.29	0.16	0.16	0.16	0.16
RTOMean_L	13.58	4.67	12.37	17.86	19.15	0.17	0.17	0.17	0.17
Seq2Seq w/ LSTM	13.05	4.22	11.87	17.30	18.43	0.05	0.10	0.04	0.04
Seq2Seq w/ GRU	12.85	4.14	11.68	**17.04**	**18.19**	0.06	0.11	0.04	0.05
TCN + LSTM	13.14	3.85	11.63	17.76	19.78	0.12	0.15	0.07	0.15
TCN + GRU	13.14	3.84	11.52	17.82	20.15	0.13	0.18	0.07	0.15
TCN + MLP	13.05	4.04	11.76	17.42	18.75	0.06	0.12	0.05	0.05
STGNN	12.75	**3.65**	11.42	17.19	18.65	0.14	0.21	0.13	0.12
Transformer_T	12.77	3.74	**11.33**	17.25	19.08	0.12	0.21	0.09	0.10
Transformer_ST	**12.70**	3.68	11.35	17.12	18.58	0.10	0.20	0.09	0.07

Table 1: Performance of the short-term models for face behavior forecasting.

	MPJPE	ST	MT	LT	FDE	Δ	Δ_{ST}	Δ_{MT}	Δ_{LT}
Ground truth	-	-	-	-	-	0.31	0.29	0.32	0.31
Zero-velocity	5.82	2.07	5.20	7.69	8.42	0.00	0.00	0.00	0.00
LinearProp	10.73	2.82	8.27	15.37	19.00	0.18	0.18	0.18	0.18
RTOMean	6.28	2.22	5.52	8.37	9.32	0.04	0.04	0.04	0.04
RTOMean_L	6.41	2.30	5.61	8.54	9.52	0.08	0.08	0.08	0.08
Seq2Seq w/ LSTM	5.83	2.04	5.23	7.72	8.41	0.02	0.06	0.01	0.01
Seq2Seq w/ GRU	5.80	1.99	5.17	7.70	8.41	0.03	0.06	0.02	0.02
TCN + LSTM	6.01	1.99	5.12	8.16	9.61	0.06	0.07	0.03	0.08
TCN + GRU	6.01	**1.98**	5.14	8.15	9.60	0.07	0.08	0.03	0.09
TCN + MLP	5.81	2.01	5.17	7.71	8.45	0.03	0.06	0.02	0.02
STGNN	5.85	1.99	5.16	7.80	8.64	0.04	0.07	0.03	0.04
Transformer_T	5.80	2.02	**5.11**	7.72	8.55	0.04	0.09	0.03	0.03
Transformer_ST	**5.78**	1.99	5.13	**7.68**	**8.34**	0.03	0.09	0.02	0.01

Table 2: Performance of the short-term models for upper body behavior forecasting.

	MPJPE	ST	MT	LT	FDE	Δ	Δ_{ST}	Δ_{MT}	Δ_{LT}
Ground truth	-	-	-	-	-	1.33	1.28	1.31	1.36
Zero-velocity	25.87	9.04	20.37	31.98	34.51	0.00	0.00	0.00	0.00
LinearProp	55.00	12.20	38.94	77.91	97.87	0.95	0.95	0.95	0.95
RTOMean	28.81	9.78	22.49	36.29	39.32	0.26	0.26	0.26	0.26
RTOMean_L	28.81	9.81	22.51	36.28	39.37	0.29	0.29	0.29	0.29
Seq2Seq w/ LSTM	25.40	8.61	20.23	31.31	33.43	0.10	0.30	0.07	0.05
Seq2Seq w/ GRU	25.82	8.65	20.45	31.99	34.59	0.14	0.35	0.10	0.09
TCN + LSTM	25.83	8.45	19.97	32.36	36.32	0.18	0.37	0.08	0.18
TCN + GRU	26.31	8.28	19.95	33.36	38.21	0.24	0.38	0.13	0.27
TCN + MLP	25.69	8.53	20.18	31.95	34.60	0.13	0.32	0.08	0.09
STGNN	26.51	8.36	20.53	33.45	36.89	0.19	0.38	0.15	0.15
Transformer_T	**25.13**	8.37	**19.78**	**31.14**	33.43	0.15	0.45	0.09	0.08
Transformer_ST	25.15	**8.16**	19.81	31.24	**33.15**	0.12	0.40	0.08	0.04

Table 3: Performance of the short-term models for hands behavior forecasting.