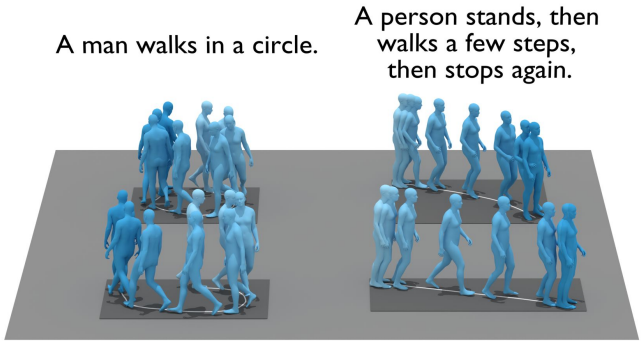
MVA Computer Vision Presentation Topic L: Text-Conditioned 3D human motion synthesis



Paul CHAUVIN Thomas DELLIAUX Thibaut LOISEAU

MATHÉMATIQUES VISION APPRENTISSAGE

école normale supérieure paris-saclay

université PARIS-SACLAY

A. Motivation and definition of the problem

The goal of the research is to **generate 3D human motion sequences from natural language descriptions.**

The problem is significant because it **has numerous applications** in both virtual (e.g., game industry) and real-world settings (e.g., controlling a robot with speech).

Previous work in this area has focused on **generating motions based on a single action label, not a full sentence.**

The proposed method allows for sampling from a distribution of human motions based on natural language descriptions and uses **transformer models for both language and motion encoding**.

Evaluation of generated motions is challenging and this work uses **multiple quantitative and qualitative measures** to assess performances.

B. Research question

How can we generate 3D human motion sequences from natural language descriptions using transformer models ?

C. Literature review

Previous work on human motion synthesis **has employed Generative Adversarial Network (GANs)** [1,2], **Variational Autoencoders (VAEs)** [3,4], and **normalizing flows** [5,6].

Motion synthesis methods can be divided into **unconstrained generation** [7,8,9], which models the **entire space of possible motions**, and **conditioned synthesis** [10,11,12], which allows for **controllability through conditions** such as music, speech, action, or text.

Text-conditioned motion generation has been approached as a **machine translation problem** [1, 13,14] or through joint **cross-modal embeddings mapping text and motion to the same space** [15, 16, 17].

Many state-of-the-art text-conditioned motion **generation models are deterministic** [16, 17] and use **cross-modal latent spaces,** while others use **impoverished body motion representations** [13,15] or only **model upper body motion** [1].

The paper we are mainly studying here [17] builds on Language2Pose [18] and integrates a variational approach for sampling a diverse set of motions from a single text input, using **Transformers** [19] to **encode motion sequences into a single embedding** and without hand-crafted separation of upper and lower body.

D.1. Methodology - Data used

	KIT Dataset	Human ML 3D Dataset
Number of motion sequences	3 911	14 616
Number of sequence description	6 353	44 970
Average number of word per sequence	9.5	12
Other	Can be **converted into SMPL body format** using correspondences with the AMASS MoCap collection, resulting in a subset of 2888 annotated motion sequences	**More isolated cases** (and challenging cases) than in KIT. If successful, can lead to better simulation in a **much wider context**

D.2. Methodology - Implementation

KIT: reproducing TEMOS paper: same seed, same data, only a difference in the hardware used (T4 vs V100)

HUMAN ML 3D: 1/ Creation of a script cleaning data of csv to the right format. (*load_annotation function*). 2/ Modification of the Data Loader (*load_hml3D_keyid function*). 3/ Creation of a config file.

```python
def load_annotation(keyid, datapath):
    annpath = datapath/ 'texts'/ (keyid + ".txt")
    anndata=[]
    with open(annpath, "r") as f:
        line = f.readline()

        while line !='':
            lines = line.split("#")
            anndata.append(lines[0])
            line = f.readline()


    if len(anndata) == 0:
        logger.error(f"{keyid} has no annotations")
        return None, False

    return anndata, True
```

```python
def load_hml3d_keyid(keyid, hml3d_path, *, index):
    keyid = int(keyid)
    smpl_datapath = os.path.join(hml3d_path , index.loc[keyid]["source_path"])
    try:
        smpl_data = np.load(smpl_datapath)
    except FileNotFoundError:
        return None, False

    smpl_data = {x: smpl_data[x] for x in smpl_data.files}
    return smpl_data, True
```

D.3. Methodology - Experiments

1/ Reproduction of TEMOS

2/ Change of certain parameters of TEMOS to check their impacts on the results (change in batch size, in number of epochs)

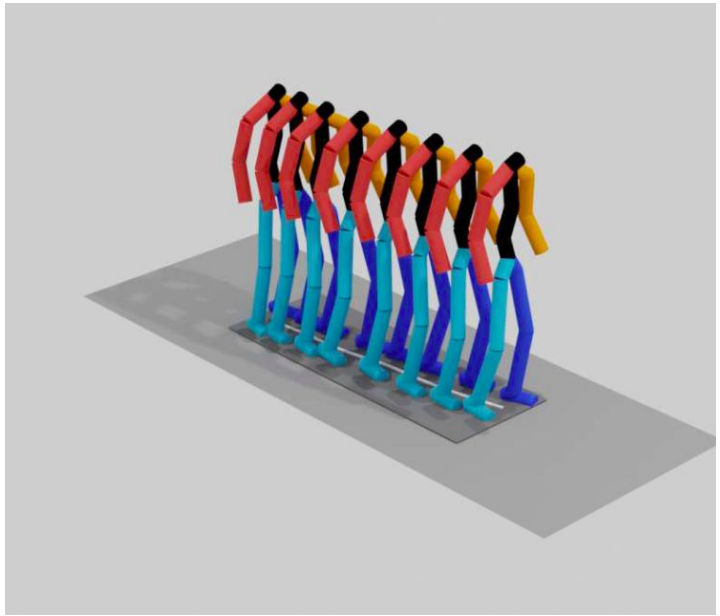
3/ Use of HUMAN ML 3D dataset to train and test the model on more complex cases than with KIT dataset

4/ Representation of the results to assess the model qualitatively

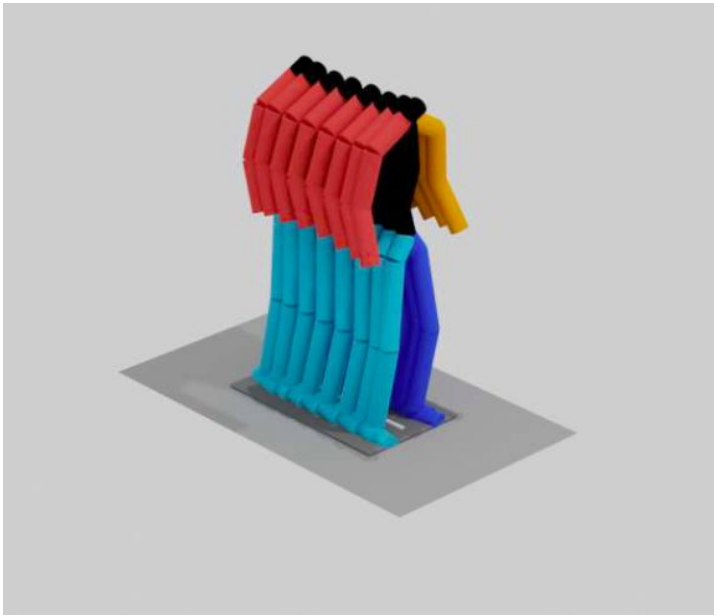
E.1. Quantitative results and comparison to state-of-the-art

	Average Positional Error				Average Variance Error			
	root joint	**global traj.**	**mean local**	**mean global**	**root joint**	**global traj.**	**mean local**	**mean global**
TEMOS original paper with batch size 32 (KIT dataset)	0.963	0.955	0.104	0.976	0.445	0.445	0.005	0.448
TEMOS reproduction from ourselves with batch size 32	1.03450	1.02527	0.10629	1.04687	0.47438	0.47378	0.00495	0.47716
TEMOS with batch size 24	1.04985	1.04007	0.10487	1.06355	0.45413	0.45331	0.00506	0.45688
HUMAN ML 3D dataset	0.8333	0.79365	0.08739	0.81707	0.23603	0.23560	0.00582	0.24166

E.2. Qualitative results - Visualization



A person is walking but paranoid someone is following them



A person is jogging on the spot

F. Difficulties

- Complex code architecture
- Use of Google Virtual Machine which crashes quite often
- Difficult and unusual data types (MMM and SMPL manipulations)

G. Conclusion

- Almost same results when reproduction of TEMOS paper
- Performances decrease when changing epochs number and batch size
- Able to obtain some results on HUMAN ML 3D dataset with good quantitative results
- However, qualitative results are not sufficient

H. References (1/2)

[1] Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2Action: Generative adversarial synthesis from language to action. In: International Conference on Robotics and Automation (ICRA) (2018)

[2] Lin, X., Amer, M.: Human motion modeling using DVGANs. arXiv preprint arXiv:1804.10652 (2018)

[3] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2Motion: Conditioned generation of 3D human motions. In: ACM International Conference on Multimedia (ACMMM) (2020)

[4] Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Visio (ICCV) (2021)

[5] Henter,G.E.,Alexanderson,S.,Beskow,J.:MoGlow:Probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics (TOG) (2020)

[6] Zanfir, A., Bazavan, E.G., Xu, H., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In: European Conference on Computer Vision (ECCV) (2020)

[7] Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: International Conference on Computer Vision (ICCV) (2019)

[8] Zhao, R., Su, H., Ji, Q.: Bayesian adversarial human motion synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2020)

[9] Zhang, Y., Black, M.J., Tang, S.: Perpetual motion: Generating unbounded human motion. arXiv preprint arXiv:2007.13886 (2020)

H. References (1/2)

[10] Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. In: Neural Information Processing Systems (NeurIPS) (2019)

[11] Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to generate diverse dance motions with transformer. arXiv preprint arXiv:2008.08171 (2020)

[12] Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI choreographer: Music conditioned 3D dance generation with AIST++. In: International Conference on Computer Vision (ICCV) (2021)

[13] Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics Auton. Syst. (2018)

[14] Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. Visually Grounded Interaction and Language (ViGIL) NeurIPS Workshop (2018)

[15]Yamada, T., Matsunaga, H., Ogata, T.: Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. Robotics and Automation Letters (2018)

[16] Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: International Conference on Computer Vision (ICCV) (2021)

[17] Mathis Petrovich, Michael J Black, and Gul Varol. Temos: Generating diverse human motions from textual descriptions. arXiv preprint arXiv:2204.14109, 2022.

[18] Ahuja, C., Morency, L.P.: Language2Pose: Natural language grounded pose forecasting. In: International Conference on 3D Vision (3DV) (2019)

[19] Vaswani,A.,Shazeer,N.,Parmar,N.,Uszkoreit,J.,Jones,L.,Gomez,A.N.,Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing