DeepHuMS: Deep Human Motion Signature for 3D Skeletal Sequences

Neeraj Battan, Abbhinav Venkat, and Avinash Sharma

International Institute of Information Technology, Hyderabad (IIIT-H), India.
{neeraj.battan, abbhinav.venkat}@research.iiit.ac.in,
{asharma}@iiit.ac.in

Abstract. 3D Human Motion Indexing and Retrieval is an interesting problem due to the rise of several data-driven applications aimed at analyzing and/or re-utilizing 3D human skeletal data, such as data-driven animation, analysis of sports bio-mechanics, human surveillance etc. Spatio-temporal articulations of humans, noisy/missing data, different speeds of the same motion etc. make it challenging and several of the existing state of the art methods use hand-craft features along with optimization based or histogram based comparison in order to perform retrieval. Further, they demonstrate it only for very small datasets and few classes. We make a case for using a learned representation that should recognize the motion as well as enforce a discriminative ranking. To that end, we propose, a 3D human motion descriptor learned using a deep network. Our learned embedding is generalizable and applicable to real-world data - addressing the aforementioned challenges and further enables sub-motion searching in its embedding space using another network. Our model exploits the inter-class similarity using trajectory cues, and performs far superior in a self-supervised setting. State of the art results on all these fronts is shown on two large scale 3D human motion datasets - NTU RGB+D and HDM05.

Keywords: 3D Human Motion Retrieval · Self-supervised Learning · 4D Indexing · MoCap Analysis

1 Introduction

3D Human Motion Retrieval is an emerging field of research due to several attractive applications such as data-driven animation, athletic training, analysis of sports bio-mechanics, human surveillance and tracking etc. Performing such analysis is challenging due to the high articulations of humans (spatially and temporally), noisy/missing data, different speeds of the same action etc. Recent research in pose estimation, reconstruction, as well as the advancement in motion

---

[1] Link for Code: https://github.com/neerajbattan/DeepHuMS.
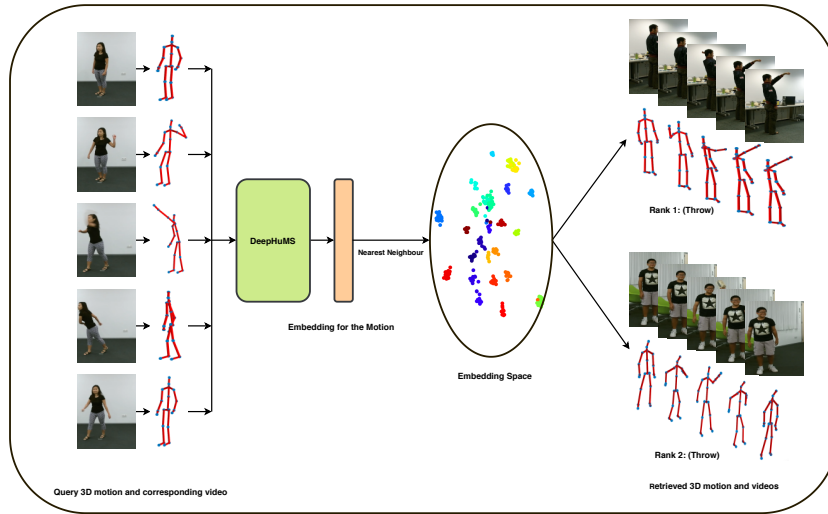[2] Link for Project Video: https://bit.ly/31B1XY2.

Fig. 1: Motivation of 3D Human Motion Retrieval. Given a query 3D Skeletal Motion Sequence, we retrieve the top-k most similar sequences. A major motivation is that the corresponding videos of the retrieved results are view, appearance and background invariant.

capture systems has now resulted in a large repository of human 3D data that requires processing. Moreover, since the procurement of new motion data is a time-consuming and expensive process, re-using the available data is of primary importance. To that end, we solve the problem of 3D Human Motion Retrieval and address several of the aforementioned challenges, using, a 3D human motion descriptor learned using a deep learning model.

While 3D human motion recognition is a commonly researched field, 3D human motion retrieval is much less explored. The task of human motion retrieval consists of two parts - building the feature representation and then the retrieval algorithm. Therefore, it requires recognizing the action as well as, importantly, enforcing a ranking i.e., a "low-dimensional" "recognition-robust" and "discriminative" feature embedding that is capable of fast retrieval is desirable.

Aiming at incorporating several of these properties, several hand crafted features from skeleton sequences have been developed [10, 7]. There has also been considerable research in the direction of improving the retrieval algorithm [3, 6] and having better similarity metrics for comparison [8, 12]. For retrieval purposes, one common method is to solve an optimization problem, which is however slow and susceptible to local minimas [14]. Alternatively, a few others perform a histogram/code-book matching. However, these methods are affected by noisy data, different length and variable frame rates of sequences etc. Moreover, they all demonstrate their retrieval accuracy over a very small number of sequences and classes. Hence, we would like to move towards learnable representations that

can account for several of these shortcomings, while still maintaining minimal supervision.

A closely related problem to retrieval in which learnable representations have been widely explored is 3D action/motion recognition. In the last few years, several deep learning model innovations have been made to better exploit the spatial and temporal information available in skeleton data [22, 21, 23]. While these models do a respectable job in recognition, they perform poorly in retrieval due to not having a discriminative enough embedding space. Further, several of them highly depend on the availability of class labels. The number of class labels available in existing datasets is fairly limited, and such supervised models are incapable of exploiting similar sub-actions amongst various classes. Hence, the requirement of a more generalized model is in order.

Therefore, in this paper, we would like to propose a discriminative learnable representation, DeepHuMS, for retrieval, that produces instantaneous retrieval with a simple nearest neighbour search in the repository. To summarize, our contributions are:

– We propose a novel deep learning model that makes use of trajectory cues, and optionally class labels, in order to build a discriminative and robust 3D human motion descriptor for retrieval.
– Further, we perform sub-motion search by learning a mapping from sub-sequences to longer sequences in the dataset by means of another network.
– Experiments are performed, both, with and without class label supervision. We demonstrate our model's ability to exploit the inter-class motion similarity better in the unsupervised setting, thus, resulting in a more generalized solution.
– Our model is learned on noisy/missing data as well as motions of different speeds and its robustness in such scenarios indicates its applicability to real world data.
– A comparison of our retrieval performance with the publicly available state of the art in 3D motion recognition as well as 3D motion retrieval on 2 large scale publicly available datasets is done to demonstrate the state-of-the-art results of the proposed model.

2   Related Work

Most approaches on the 3D human motion retrieval have focused on developing hand crafted features to represent the skeleton sequences[10, 14, 19]. In this section, we broadly categorize them by the method in which they engineer their descriptors.

While a few use an objective function[14], few others use codebook or histogram comparisons [20, 6] to do the same. The traditional frame based approaches extract out features for every frame. [12] proposed a geometric pose feature to encode pose similarity. [20] used joints' orientation angles and angles-forward differences as local features to create a codebook and generate a Bag of Visual Words to represent the action. [24, 9] suggest hand drawn sketch based

skeleton sequence retrieval methods. On the other end of the spectrum, sequence based motion features utilize global properties of the motion sequence [4, 13, **?**]. Muller et al [13] presented the motion template (MT) in which motions of the same class can be represented by an explicit interpretable matrix using a set of boolean geometric feature. To tolerate the temporal variance in the training process, dynamic time warping (DTW) was employed in their work. However, it still requires the training data in the same class to have the same number of repetitions and both the training as well as retrieval process of MT are time-consuming. [15] created a temporal motion model using HMM and a histogram of 3D joints descriptors after creating the dictionary. [11] applied a Gaussian Mixture Model to represent character poses, wherein the motion sequence is encoded, then they used DTW and a string matching to find similarities between two videos. Recently many graph based models have been proposed to exploit the geometric structure of the skeleton data [18, 16, 17]. The spatial features are represented by the edges connecting the body joints and temporal features are represented by the edges connecting the same body joint in adjacent frames.

For the task of retrieval, [29] proposed a simple auto-encoder that captures high-level features. However, their model doesn't explicitly use a temporal construct for motion data. Primarily, learnable representations from 3D motion data have been used for other tasks. [25, 21] are a few amongst many who used deep learning models for 3D motion recognition. Similarly, [28] adopts a unidirectional LSTM to encode the skeleton frames within the hidden network states and learn what subsequences of encoded frames belong to the specified action classes.

Broadly, the existing methods are affected by noisy data, the length and variable frame rates of sequences, and are slow at retrieval. Further, they lack a learned discriminative embedding which is capable of performing sub-sequence retrieval.

3   DeepHuMS: Our Method

In order to build a 3D human motion descriptor, we need to exploit the spatio-temporal features in the skeletal motion data. Briefly, we have three key components - (i) the input skeletal location and joint level motion trajectories to the next frame, (ii) an RNN to model this temporal data and (iii) a novel trajectory based similarity metric (explained below) to project similar content together using a Siamese architecture. We use two setups to train our model - (a) self-supervised, with a "contrastive loss" given by Equation 1 to train our Siamese model and (b) supervised setup, with a cross entropy on our embedding, in addition to the self-supervision. Refer to Figure 2 for a detailed architecture explanation.

$$L_{contrastive} = (1 - Y)\frac{1}{2}(D_w^2) + (Y)\frac{1}{2}\{max(0, m - D_w)\}^2 \qquad (1)$$
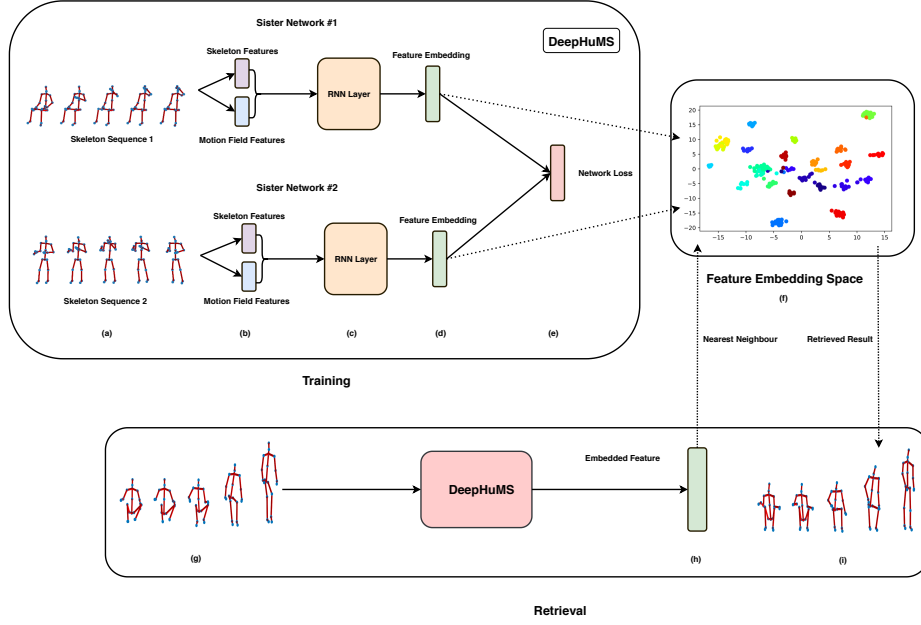
Fig. 2: Overview of our model - DeepHuMS. Given two skeleton sequences (a), we first extract the 3D joint locations and motion field between consecutive frames to represent the spatio-temporal data (b). The two are concatenated together and given to an RNN [21] to model the 4D data(c). The resulting embeddings (d) are compared based using (e) contrastive loss (and optionally classification loss) to make them "discriminative" and "recognition-robust". Similarity is enforced based on the full sequence's motion distance and motion field. At the time of retrieval, given a 3D sequence to the network (g), with the resultant embedding, a nearest neighbour search is done in the embedding space (f) generated from the training data.

In Eq. 1, Dw is the distance function (for e.g. "Euclidean distance"), m is the margin for similar and dissimilar samples and Y is if the label value (1 for similar samples and 0 for dissimilar).

$$L_{crossentropy} = \sum_{n=1}^{M} y_{o,c} log(p_{o,c}) \qquad (2)$$

In Eq. 2, y indicates (0 or 1) if class label c is the correctly classified, given o, the observation. M is the number of classes and p is the predicted probability, given an observation o of class c.

Similarity Metric. Two 3D human motion sequences are said to be similar if both the joint-wise "Motion Field" and joint-wise "Motion Distance" across

the entire sequence are similar. The motion field depicts the direction of motion as well as the importance of the different joints for that specific sequence. The motivation behind this is evident in Figure 3 in which the hand and elbow joints are more important for *waving*. However, the motion field can end up being zero as show in Figure 3. Therefore, we couple it with the joint-wise motion distance in order to build a more robust similarity metric. It is to be noted that having such a full video trajectory based similarity makes it difficult to directly retrieve sub-sequences of similar content. We handle this scenario in Section 4.5 using a second network.
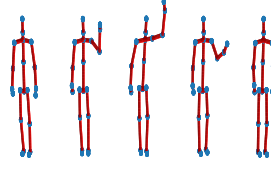

Fig. 3: Hand and elbow are important joints while waving, due to higher motion distance.

Equation 3 gives the motion field MF between two frames i and j. We use the difference between the first and last frame to depict the motion field of the sequence. Here $F[i]$ contains the 3D joints for ith frame in the skeleton sequence. Similarly, equation 4 gives the motion distance of the entire sequence. $MD[j]$ is the total distance covered by j^{th} joint and N is the number of frames in the 3D skeleton sequence.

$$MF[i, j] = F[i] - F[j] \tag{3}$$

$$MD[j] = \sum_{i=1}^{N-1} \|F[i+1][j] - F[i][j]\| \tag{4}$$

Different number of frames. In case of sequences that have different speeds of motion or sampling rate, but similar content, the information available at the input is different, but, the resulting motion field and motion distance across the entire sequence is the same. Hence, we augment our data and enforce such sequences to be projected together in our embedding space using the contrastive loss. In other words, we map sequences with less information to the same location to sequences with more information, in the embedding space (See section 4.5 for more on implementation details).
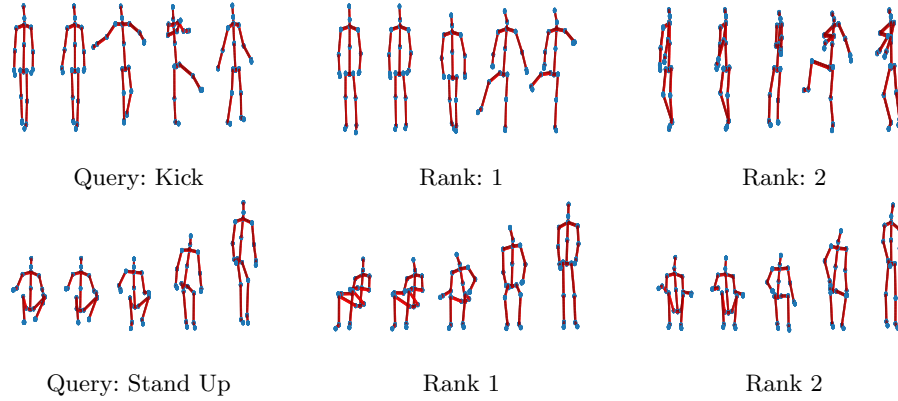
Fig. 4: Retrieval Results on NTU-RGB+D dataset [2] using our supervised setup. Given a query, we show the top-2 ranked results.

4   Experiments

4.1   Datasets

We use two commonly used large scale public MoCap datasets to evaluate our method for human 3D motion retrieval.

NTU RGB+D [2]: This dataset provides RGB, depth, infra red images and 3D locations of 25 Joints on the human body. It consists of around 56,000 sequences from 60 different classes acted by 40 performers. We use the given performer wise split for learning from this dataset.

HDM05 [1]: This dataset provides RGB images and 3D locations of 31 Joints in human body. There are around 2300 3D sequences of 130 different classes performed by 5 performers in this dataset. We follow [14] for evaluation and therefore combine similar classes (for e.g. *walk2StepsLstart* and *walk2StepsRstart*) to get a total of 25 classes. We follow a performer-wise split with the first 4 performers for training and the last one for testing.

4.2   Implementation Details

All of the trained models, code and data shall be made publicly available, along with a working demo. Please refer to our supplementary video for more results.

Data Pre-processing & Augmentation. In order to make it performer/character invariant, we normalized the 3D joint locations based on bone length of the performer. To diversify our datasets, for every 3D sequence, we create two more sequences - a faster and a slower one. The faster sequence is created by uniformly

sampling every other frame, and the slower sequence is created by interpolating between every pair of frames.

Network Training. We use Nvidias GTX 1080Ti, with 11GB of VRAM to train our models. A batch size of 128 is used for NTU RGB+D dataset, and a batch size of 8 is used for training the HDM05 dataset. We use the ADAM optimizer with an initial learning rate of 10^{-3}, to get optimal performance on our setup. The training time for NTU RGB+D dataset is 6 hours and HDM05 is 1 hour. Each dataset is trained individually from scratch.

4.3   Evaluation Metrics

Retrieval Accuracy.   This is a class-specific retrieval metric. In "top-n" retrieval accuracy, we find out how many of the "n" retrieved results belong to the same class as the query motion.

Dynamic Time Warping (DTW) distance. Inspired from [12], we use Dynamic Time Warping as a quantitative metric to find out the similarity between two sequences based on distance. Two actions with different labels can be very similar, for example, drinking and eating. Likewise, the same class of actions performed by two actors can have very different motion. Hence using only the class-wise retrieval accuracy as metric doesn't provide the complete picture, and therefore, we use DTW as well.

4.4   Comparison with State of the Art

 Since all of the existing state of the art methods use supervision, we compare our supervised setup with them in two ways - (a) with existing 3D Human Motion Retrieval models and (b) with 3D Human Motion Recognition embeddings. Class-wise retrieval accuracy of the top-1 and top-10 results are reported for the same.

3D Human Motion Retrieval. Most of the existing retrieval methods [26, ?,27] show results on only upto 10 classes, and on very small datasets. [14] use the same number of class labels as us, and we therefore compare with them in Figure 5. As shown in Figure 5, the area under the PR curve is far larger for our method, and we have learned a much more robust 3D human motion descriptor.

3D Human Motion Recognition.   We compare with learned representations from 3D Motion recognition. The results for the recognition models in Table 1 and Figure 6 are computed using their embeddings trained on our datasets.
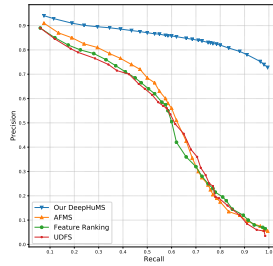
Fig. 5: A comparison of our method with 3D Human Motion Retrieval Methods via a Precision Recall curves for HDM05 dataset

Method	Top 1 Ret. Acc.	Top 10 Ret. Acc.
HCN [25]	0.61	0.56
IndRNN [21]	0.69	0.62
DeepHuMS (Ours)	**0.78**	**0.753**

Table 1: A Comparison of Retrieval accuracy of our method with state of the art Motion Recognition embedding
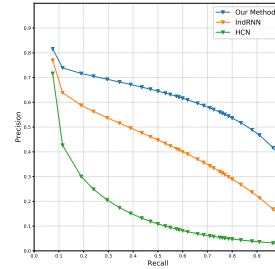

Fig. 6: A comparison of 3D Motion Recognition methods for retrieval with ours via Precision Recall curves for NTU RGB+D dataset

Retrieval v/s Recognition. Figure 7 shows how our model produces a more clustered and therefore, disciminative space, suitable for retrieval, in comparison with the embedding space of [21], a state of the art 3D motion recognition algorithm. Recognition algorithms only focus on learning a hyperplane that enables them to identify limited motion classes. Adding a generalized similarity metric enforces an implicit margin in the embedding space - a motion trajectory based clustering.

4.5   Discussion

The results and inferences reported below are consistent for all datasets. For more detailed results, please see our supplementary video.
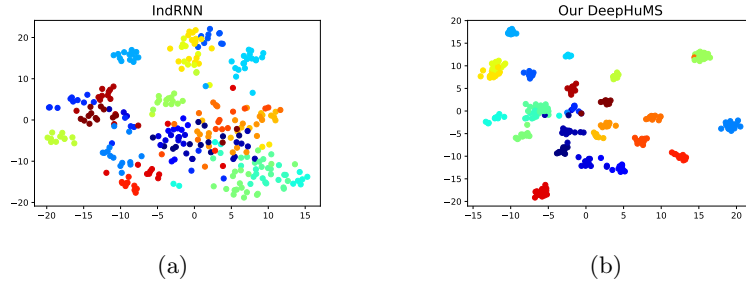
(a)   (b)

Fig. 7: A comparison of the t-SNE representation of Motion Recognition [21] with our method on NTU-RGB+D dataset [2]

Results of Self-supervision. Going beyond class labels, we are able to exploit inter-class information when trained with only self-supervision; therefore, the resulting retrieved motions are more closer to the query motion than the supervised setup, in terms of per frame error after *DTW - 34mm of supervised v/s 31mm of unsupervised*. This is a promising result, particularly because existing datasets have very limited number of labels and it enables us to exploit 3D sub-sequence similarity and perform retrieval in a label-invariant manner.

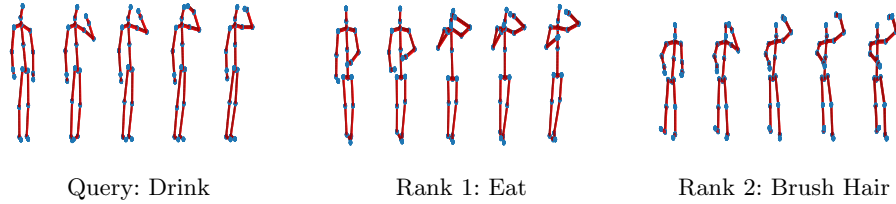Query: Drink   Rank 1: Eat   Rank 2: Brush Hair

Fig. 8: Retrieval Results for Self-supervised setup, which shows that we exploit inter-class similarity. The retrieved videos do not belong to the same class but belong to similar class.

Noisy/Missing data. To prove the robustness of our method towards noisy data, we trained and tested out model with missing data - random 20% of joints missing from all frames of each sequence. This scenario simulates sensor noise or occlusions while detection of 3D skeletons. As shown below, we still achieve an impressive retrieval accuracy in scenarios where optimization based state of the art methods would struggle.

Sub Motion Retrieval. Sub Motion retrieval becomes important when we would like to search for a smaller action/motion in longer sequences. But it

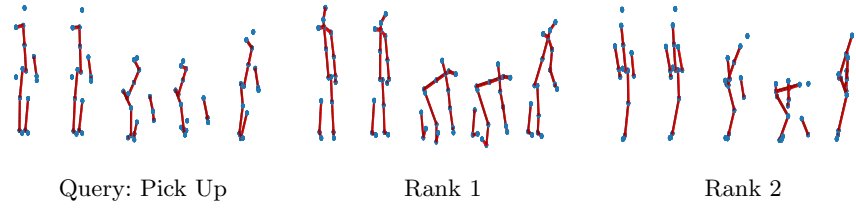Query: Pick Up   Rank 1   Rank 2

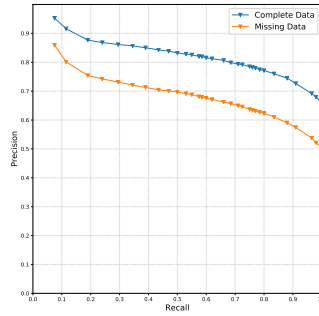Fig. 9: Retrieval Results for Noisy Data

Fig. 10: A comparison of the effect of Noisy Data on our retrieval performance.

is very challenging task due to the variations in length and actions in sub se-
quences. Moreover, our similarity metrics, in their current form can't account for
sub sequences directly. To address this, we follow the model shown in Figure 11.
Using this simple model, we retrieve the whole sequence it is a part of. This is a
good starting point for the community and we believe that better solutions can
be developed that directly incorporate sub-sequence information in the motion
descriptor.

Sequences of different speeds. Irrespective of the sampling rate/speed of
motion, the motion field and distance would be the same. So, we take care of
motions performed at different speeds by minimizing all to the same embedding.
We do this by simulating a sequence that is twice as slow and twice as short
by interpolation and uniform sampling respectively, and training a siamese over
them. It shows that more the number of frames, more amount of information is
given to the network, and therefore, better the results. We handle short to very
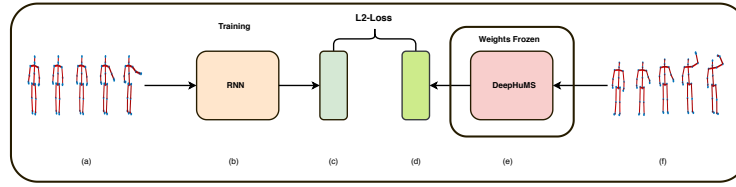long sequences ranging in length from 15 to 600 frames.

Fig. 11: Given a subsequence (a) as the input, we use another RNN (b) to learn an embedding (c). This is minimized with L2 loss w.r.t the ground truth (d) generated from DeepHuMS (e) trained on long sequences (f)
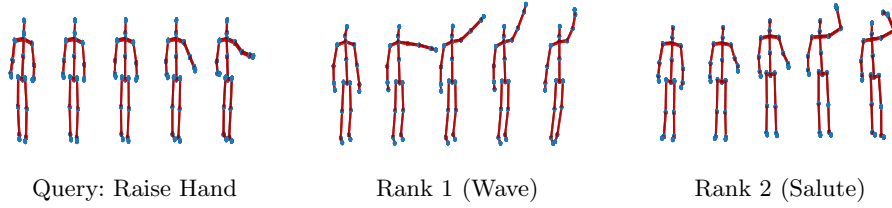

Query: Raise Hand   Rank 1 (Wave)   Rank 2 (Salute)

Fig. 12: Retrieval Results for Sub sequence Retrieval
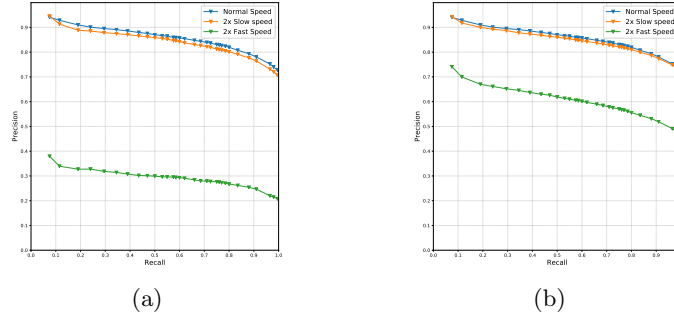

(a)   (b)

Fig. 13: A Comparison of our Precision Recall curves before and after training for different speeds on HDM05 dataset. **(a)** Retrieval Performance before training for different speeds. **(b)** Retrieval Performance after training for different speeds.

Retrieval Time.   We have a fairly low dimensional embedding of size 512, and perform a simple nearest neighbour search throughout the training dataset. This yields an average retrieval time for the test set to be 18ms for NTU RGB+D and 0.8ms for HDM05 dataset. The retrieval time is proportional to the dataset size, and one could use more advanced algorithms such as tree based searching.

4.6   Limitations

Although our current model demonstrates impressive results, there exists some shortcomings in terms of generalisability and design. Firstly, indexing the sub-motion in the full sequence isn't trivial. Secondly, sequences with repetitive actions would be sub-optimal to handle with our full video-DeepHuMS descriptor. Both of these are because of different motion fields/distances as well as the lack of explicit temporal indexing of individual key-frames in the learned 3D human motion descriptor. In other words, we need either better utilization of the "semantic context" injected by the existing similarity metrics as well as need additional constructs to incorporate better semantic context. This extends to a larger discussion about how to design models to learn in a unsupervised manner.

5   Conclusion

In this paper, we make a case for using a learned representation for 3D Human Motion retrieval by means of the first deep learning based model. Our model uses trajectory cues in a self-supervised manner to learn a generalisable, robust and discriminative desciptor. We overcome several of the limitations of current hand-crafted 4D motion descriptors such as their inability to handle noisy/missing data, different speeds of the same motion, generalise to a large number of sequences and classes etc, thus making our model applicable to real world data. Lastly, we provide a initial model in the direction of for 3D sub-motion retrieval, using the learned sequence descriptor as the ground truth. We compare with state-of-the-art 3D motion recognition as well as 3D motion retrieval methods on two large scale datasets - NTU RGB+D and HDM05 and demonstrate far superior performance on all fronts - class-wise retrieval accuracy, time and frame level distance.

References

1. Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B. & Weber, A. Documentation mocap database hdm05. (Citesee,2007
2. Shahroudy, A., Liu, J., Ng, T. & Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. 2016.
3. Li, C. & Prabhakaran, B. Indexing of Motion Capture Data for Efficient and Fast Similarity Search.. *Jcp*. **1**, 35–42 2006.
4. Junejo, I., Dexter, E., Laptev, I. & Perez, P. View-independent action recognition from temporal self-similarities. *Ieee Transactions On Pattern Analysis And Machine Intelligence*. **33**, 172–185 2010.
5. Müller, M., Baak, A. & Seidel, H. Efficient and robust annotation of motion capture data, 2009.
6. Liu, X., He, G., Peng, S., Cheung, Y. & Tang, Y. Efficient human motion retrieval via temporal adjacent bag of words and discriminative neighborhood preserving dictionary learning. *Ieee Transactions On Human-machine Systems*. **47**, 763–776, 2017.

7. Feng, T., Gunawardane, P., Davis, J. & Jiang, B. Motion capture data retrieval using an artists doll, 2008.
8. Wang, M., Hua, X., Tang, J. & Hong, R. Beyond distance measurement: constructing neighborhood similarity for video annotation. *Ieee Transactions On Multimedia.* **11**, 465–476 2009.
9. Choi, M., Yang, K., Igarashi, T., Mitani, J. & Lee, J. Retrieval and visualization of human motion data via stick figures, 2012.
10. Xiao, Q., Li, J., Wang, Y., Li, Z. & Wang, H. Motion retrieval using probability graph model, 2013.
11. Qi, T., Xiao, J., Zhuang, Y., Zhang, H., Yang, X., Zhang, J. & Feng, Y. Real-time Motion Data Annotation via Action String. *Comput. Animat. Virtual Worlds.* **25**, 293–302, 2014.
12. Chen, C., Zhuang, Y., Nie, F., Yang, Y., Wu, F. & Xiao, J. Learning a 3D human pose distance metric from geometric pose descriptor. *Ieee Transactions On Visualization And Computer Graphics.* **17**, 1676–1689, 2010.
13. Müller, M. & Röder, T. Motion templates for automatic classification and retrieval of motion capture data, 2006.
14. Wang, Z., Feng, Y., Qi, T., Yang, X. & Zhang, J. Adaptive multi-view feature selection for human motion retrieval. *Signal Processing.* **120** pp. 691–701, 2016.
15. Xia, L., Chen, C. & Aggarwal, J. View invariant human action recognition using histograms of 3d joints, 2012.
16. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y. & Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition, 2019.
17. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Skeleton-Based Action Recognition with Directed Graph Neural Networks, 2019.
18. Sijie, Y., Yuanjun, X. & Dahua, L. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, 2018.
19. Wang, J., Liu, Z., Wu, Y. & Yuan, J. Learning actionlet ensemble for 3D human action recognition. *Ieee Transactions On Pattern Analysis And Machine Intelligence.* **36**, 914–927, 2013.
20. Kapsouras, I. & Nikolaidis, N. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal Of Visual Communication And Image Representation.* **25**, 1432–1445, 2014.
21. Li, S., Li, W., Cook, C., Zhu, C. & Gao, Y. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. *2018 Ieee/cvf Conference On Computer Vision And Pattern Recognition*, 2018.
22. Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y. & Luo, J. Action Recognition by Learning Deep Multi-Granular Spatio-Temporal Video Representation, 2016.
23. Tang, Y., Tian, Y., Lu, J., Li, P. & Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition, 2018.
24. Chao, M., Lin, C., Assa, J. & Lee, T. Human motion retrieval from hand-drawn sketch. *Ieee Transactions On Visualization And Computer Graphics.* **18**, 729–740, 2011.
25. Li, C., Zhong, Q., Xie, D. & Pu, S. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. *Proceedings Of The Twenty-seventh International Joint Conference On Artificial Intelligence*, 2018.
26. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R. & Bajcsy, R. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal Of Visual Communication And Image Representation.* **25**, 24–38, 2014.

27. Gowayyed, M., Torki, M., Hussein, M. & El-saban, M. Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition.

28. Carrara, F., Elias, P., Sedmidubsky, J. & Zezula, P. LSTM-based real-time action detection and prediction in human motion streams. *Multimedia Tools And Applications.* pp. 1–23 (2019)

29. Wang, Y. & Neff, M. Deep signatures for indexing and retrieval in large motion databases.