Skeleton-based Action Recognition in Non-contextual, In-the-wild and Dense Joint Scenarios

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science
in
Computer Science and Engineering by Research

by

Neel Trivedi
20171015
neel.trivedi@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
September, 2022

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ‘Skeleton-based Action Recognition in Non-contextual, In-the-wild and Dense Joint Scenarios‘ by Neel Trivedi, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Ravi Kiran Sarvadevabhatla

To, my family and friends

Acknowledgement

Firstly, I would like to express my gratitude to my advisor, Prof. S Ravi Kiran whose constant guidance, support and motivation made my research journey possible. Be it his extreme emphasis on neat presentation skills, his insightful analysis and ideas about some problems or his spellbound enthusiasm for research, Prof. S Ravi Kiran made a significant and positive impact on me, both as a researcher and as a good human being. And for that I will be forever grateful to him.

I would like to express my extreme gratitude to my parents and my elder brother for supporting me in all my ventures be it academic or otherwise. Their constant support, motivation and unconditional love helped me successfully complete my research journey and will continue to help me in all my future endeavours. I would also like to extend my gratitude to my grandma, uncle, aunt, cousin and Dhwani, who have always been there to support me and my aspirations. I couldn't have become what I am today if it wasn't for all of them.

I would also like to thank my lab mates and co-authors, Pranay, Anirudh, Aditya, Shubh, Sai, Adithya and Debtanu for being my companions throughout my time at CVIT. Our extended discussions and brain storming sessions were a major source of knowledge and fun at the same time.

These acknowledgments would be incomplete without mentioning my amazing friends, Preet, Vaibhav, Anchit, Pulkit, Saraansh, Akshay, Anush, Manan and Priyank who made my five years long journey at IIIT nothing but memorable. Their friendship helped me keep a balance between academics and leisure. Also, I would like to extend this thanks to two of my best friends from back home, Khushi and Labdhi.

This thesis would not have been possible without the support of many people, and I would like to dedicate this work to all of them, and I hope to have their support in my future endeavours as well.

Abstract

Human action recognition, with its irrefutable and varied use cases across fields of surveillance, robotics, human object interaction analysis and many more, has gained critical importance and attention in the field of compute vision. Traditionally entirely based on RGB sequences, action recognition domain has shifted focus towards using skeleton sequences due to the easy availability of skeleton data capturing apparatus and the release of large scale datasets, in recent years. Skeleton based human action recognition, having superiority in terms of privacy, robustness and computational efficiency over traditional RGB based action recognition, is the primary focus of this thesis.

Ever since the release of large scale skeleton action datasets namely NTURGB+D and NTURGB+D 120, the community has solely focused on developing complex approaches, ranging from CNNs to complex GCNs and more recently transformers, to achieve the best classification accuracy for these datasets. However, in this rat race for state of the art performance, the community turned a blind eye to a major drawback at the data level which bottlenecks even the most sophisticated approaches. This drawback is where we start our explorations in this thesis.

The pose tree provided in the NTURGB+D datasets contains only 25 joints, out of which only 6 joints (3 for each hand) are finger joints. This is a major drawback since only 3 finger level joints are not sufficient enough to distinguish between action categories such as "Thumbs up" and "Thumbs down" or "Make ok sign" and "Make victory sign". To specifically address this bottleneck, we introduce two new pose based human action datasets - NTU60-X and NTU120-X. Our datasets extend the largest existing action recognition dataset, NTU-RGBD. In addition to the 25 body joints for each skeleton as in NTU-RGBD, NTU60-X and NTU120-X dataset include finger and facial joints, enabling a richer skeleton representation. We appropriately modify the state of the art approaches to enable training using the introduced datasets. Our results demonstrate the effectiveness of these NTU-X datasets in overcoming the aforementioned bottleneck and improving the state of the art performance, overall and on previously worst performing action categories.

Pose-based action recognition is predominantly tackled by approaches that treat the input skeleton in a monolithic fashion, i.e. joints in the pose tree are processed as a whole. However, such approaches ignore the fact that action categories are often characterized by localized action dynamics involving only small subsets of part joint groups involving hands (e.g. 'Thumbs up') or legs (e.g. 'Kicking'). Although part-grouping based approaches exist, each part group is not considered within the global pose frame, causing such methods to fall short. Further, conventional approaches employ independent

modality streams (e.g. joint, bone, joint velocity, bone velocity) and train their network multiple times on these streams, which massively increases the number of training parameters. To address these issues, we introduce PSUMNet, a novel approach for scalable and efficient pose-based action recognition. At the representation level, we propose a global frame based part stream approach as opposed to conventional modality based streams. Within each part stream, the associated data from multiple modalities is unified and consumed by the processing pipeline. Experimentally, PSUMNet achieves the state of the art performance on the widely used NTURGB+D 60/120 dataset and dense joint skeleton dataset NTU 60-X/120-X. PSUMNet is highly efficient and outperforms competing methods which use 100%-400% more parameters. PSUMNet also generalizes to the SHREC hand gesture dataset with competitive performance. Overall, PSUMNet's scalability, performance and efficiency make it an attractive choice for action recognition and for deployment on compute-restricted embedded and edge devices.

Finally, we conclude this thesis by exploring new and more challenging frontiers under the umbrella of skeleton action recognition namely "in the wild" skeleton action recognition and "non-contextual" skeleton action recognition. We introduce Skeletics-152, a curated and 3D pose dataset derived from the RGB videos included in the larger Kinetics-700 dataset to explore in the wild skeleton action recognition. We further introduce, Skeleton-mimetics, a 3D pose dataset derived from recently introduced non-contextual action dataset-Mimetics. By benchmarking and analysing various approaches on these two new dataset we lay the ground for future exploration in these two challenging problems within skeleton action recognition.

Overall in this thesis, we draw attention to prevailing drawbacks in the existing skeleton action datasets and introduce extensions of these datasets to counter their shortcomings. We also introduce a novel, efficient and highly reliable skeleton action recognition approach dubbed PSUMNet. Finally, we explore more challenging tasks of in the wild and non-contextual action recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human action recognition has been a widely studied and explored domain in the field of computer vision. With its varied application from surveillance to autonomous cars and robots, human action recognition has come to the attention of computer vision researchers. Under this umbrella of human action recognition, the sub domain of skeleton based human action recognition has gained a lot of attention with the release of large scale datasets of skeleton sequences. Skeleton data, with its superiority to conventional RGB data in terms of storage, robustness and privacy, has attracted a plethora of approaches to work on this problem of skeleton based human action recognition. Hence, over the past decade, the methods to capture such skeleton datasets have also seen a paradigm shift. This, evolution of skeleton data capturing methods, is where we start our thesis.

1.1 Evolution of Skeleton Data capturing methods

One of the first attempts at the creation of a reliable pose based motion capture dataset was done by team of researchers at CMU in 2003. The introduced dataset dubbed CMU MoCap[22] contains several hours of motion data over various locomotion activities (such as "running", "walking",), Sports activities (such as "basketball", "soccer",...) and more. As shown in Fig. 1.1 (Left), 41 infra-red markers are tapped on the body of an actor and the action is captured via 12 Vicon infrared MX-40 cameras at a 120 Hz rate which generates the motion data. Following the wide academic interest in CMU MoCap dataset Muller et al.[28] introduced the HMD05 dataset in 2007 which included 1457 motion clips spanning across 100 different motion classes. As opposed to CMU MoCap, HMD05 focuses on a very limited number of motion classes. Similar to CMU MoCap, motion data in HMD05 is also created by using 40-50 retro-reflective sensors tapped on actors' bodies as shown in Fig. 1.1 (Right). The action is captured using six to twelve calibrated high-resolution cameras at a frame rate of up to 240 Hz.

However, it can be easily understood that the motion capturing method employed by these early datasets, where multiple sensors are tapped on actors' bodies, is not scalable nor is convenient. With the availability of frugal data capturing sensors and apparatus, such as Depth cameras, Microsoft Kinect and Open Sense, came the second generation of pose based action and motion capture datasets. These

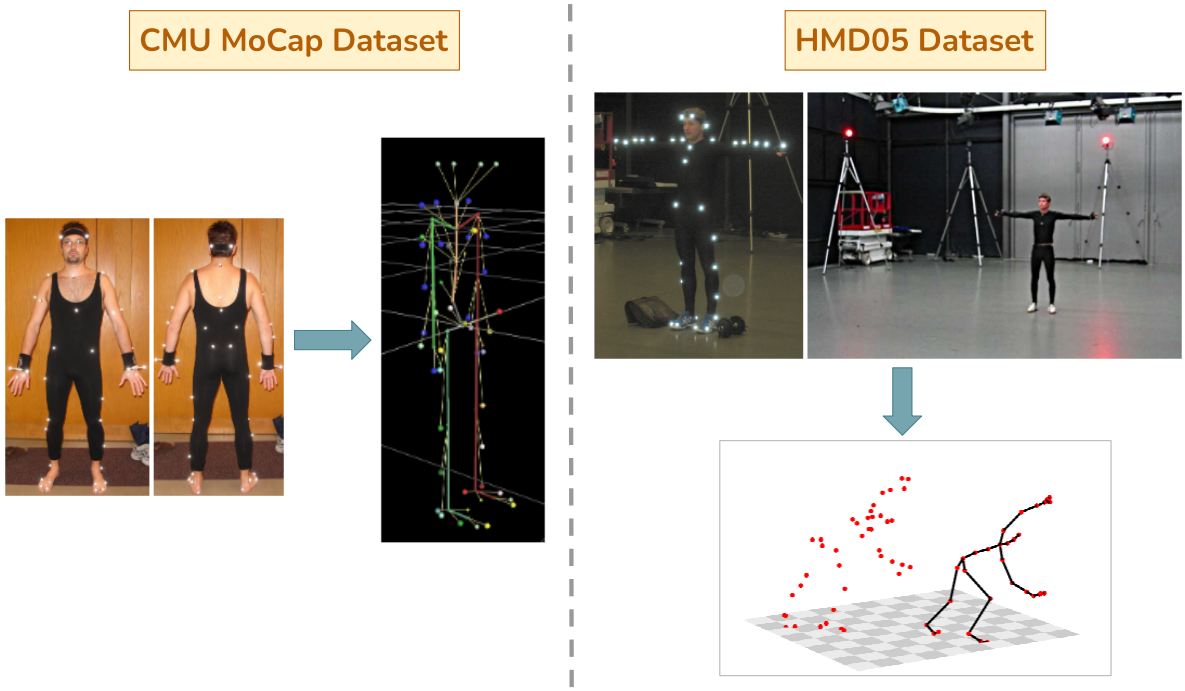


Figure 1.1 Generation one motion capture datasets: CMU MoCap (Left) and HMD05 (Right). These datasets were curated using multiple sensors tapped to actors body and then captured via several high resolution cameras yielding accurate pose estimations.

sensors capture multi modal data directly from visual feed which includes depth maps and 3d coordinates of the joints along with the RGB sequences eliminating the need of tapping multiple sensors to the body of the actors.

MSR Action3d dataset[24], was one of the first datasets which used depth cameras to obtain depth maps for each frame and sampled 3d skeleton point clouds using this depth information as shown in Fig. 1.2 (Left). Afterwards, with introduction of Microsoft Kinect sensor and its ability to capture multi modal data simultaneously (See Fig. 1.2 (Right)) showed a plethora of pose based action datasets, ranging from small scale datasets (MSRDailyActivity3D[43], UTD-MHAD [44], UWA3D Multiview II[32]) and large scale datasets (NTURGB+D[35], NTURGB+D 120[25]).

Even though the frugality and easy availability of Kinect like sensors has drawn huge attention to the aforementioned datasets, these datasets suffer from some inherent flaws,

- These Kinect (or some other) sensor based data capturing methods cannot handle the presence of a large number of subjects in the frame making the recognition models trained on such datasets unfit to identify group action activities involving multiple subjects.

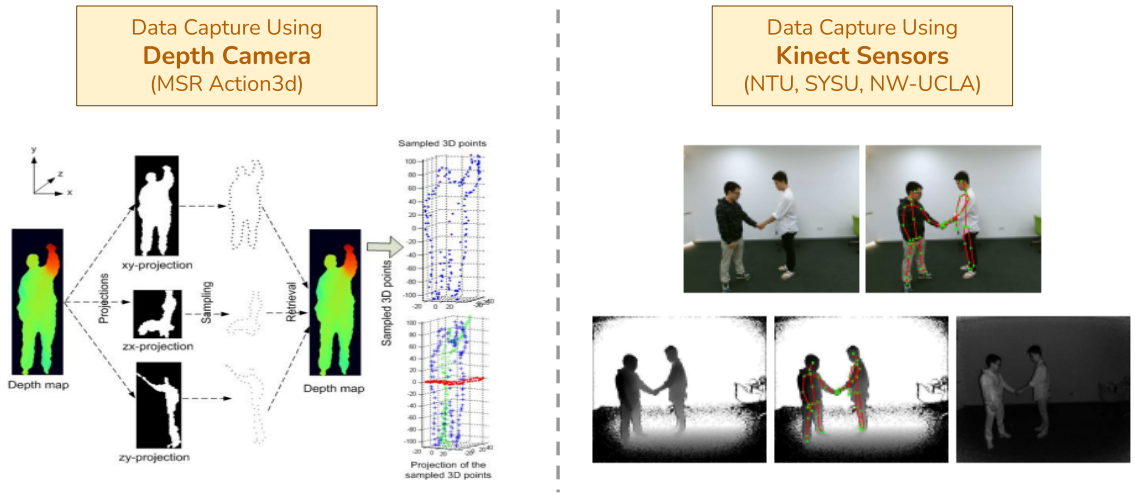


Figure 1.2 The left diagram shows the process of 3d skeleton joints estimation from depth maps as employed in creation of MSR Action3d dataset. Right diagram shows various data modalities, RGB, RGB + Joints, Depth, Depth + Joints, IR respectively, captured using Kinect sensors used in the creation of datasets such as NW-UCLA and NTURGB+D.

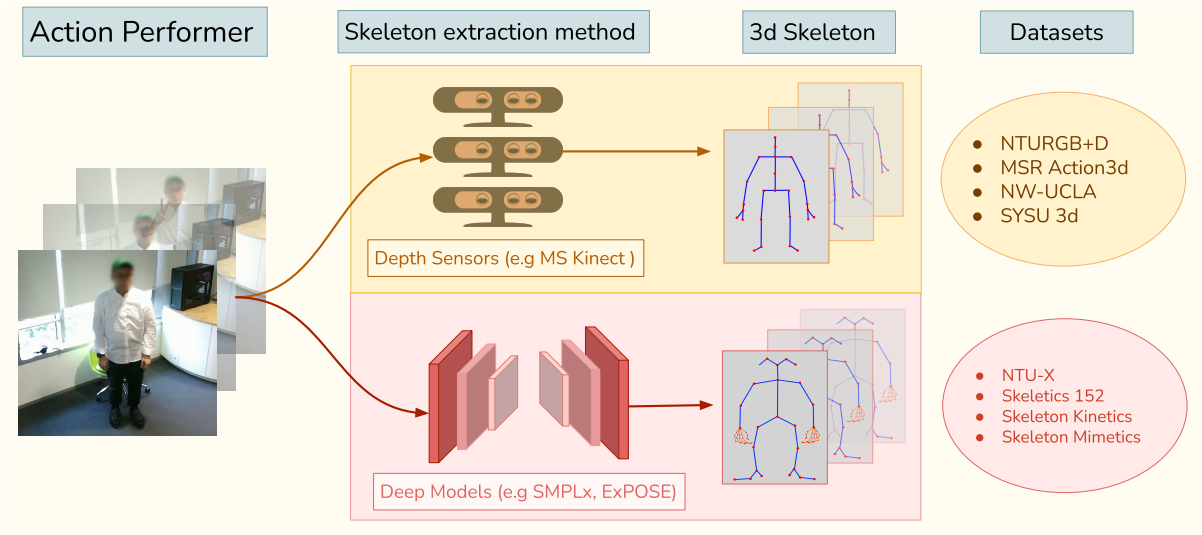


Figure 1.3 Depth Sensors based skeleton data capturing method (Above) against deep model based data capturing methods. Depth sensors captures the depth information along with 3d joints estimations. Deep models use RGB frames directly and estimates the 3d skeleton.

- Owing to its fixed calibration and capturing setups, these sensors can not capture fine 3d joints such as fingers or facial joints making the resulting pose tree sparse.

- These sensors are prone to noise caused by occlusion or partially visible subjects leading to their unreliability for data capturing for in the wild actions.
- Need of specialized sensors that can capture depth information restricts real time applications.

The aforementioned drawbacks in existing and widely used skeleton datasets naturally prompt attention in the direction of datasets created using deep learning methods (See Fig. 1.3). Deep pose estimation models with their high efficiency, generalizability and scalability can allow the creation of skeleton action datasets with include dense pose representation, multi person capture and immunity to occlusion induced noise. With the release of large scale annotated datasets such as Human3.6M[19], COCO[13], MPII[1] and highly reliable and generalizable pose capturing deep models such as OpenPose[2], SMPL-x[30], ExPose[9], a new generation of skeleton action datasets captured using such deep models are the future frontiers in this domain.

1.2 Our Contributions

- To begin our study, we thoroughly analyze the performance of various state of the art approaches for widely used skeleton action recognition datasets: NTURGB+D and NTURGB+D 120. (See Sec. 2.1)
- We experimentally draw attention to a major drawback in the existing skeleton action datasets, i.e. lack of finer finger level joints which bottlenecks recognition approaches to wrongly classify actions involving significant finger movements such as ”Make ok sign” or ”Thumbs Up”. (See Fig. 2.4)
- To overcome this major drawback, we introduce NTU-X, an extension of existing NTURGB+D datasets with a dense pose tree with a total of 118 joints involving 21 finger joints per hand, 51 facial joints and 25 body joints. (See Sec. 2.3)
- We benchmark existing state of the art approaches on our proposed NTU-X dataset and show the overall improvement in performance proving the superiority of our dataset over the conventional dataset. To Further justify our hypothesis, we show significant improvements in per class accuracy of action categories involving significant finger movements. (See Sec. 2.4)
- In the next chapter of our study, we propose a novel architecture, dubbed Part Stream Unified Modality Network (PSUMNet) which achieves the state of the art performance on NTU60-X, NTU120-X and NTURGB+D 120 datasets compared to existing competing methods which use 100%-400% more parameters. PSUMNet also generalizes to the SHREC hand gestures dataset with competitive performance. (Sec. 3.4.3)
- We introduce a unified modality processing approach as opposed to conventional independent modality approaches. This enables a significant reduction in the number of parameters. (Sec. 3.3.2)

- We propose a part based stream processing approach that enables richer and dedicated representations for actions involving a subset of joints (Sec. 3.3.1). The part stream approach also enables efficient generalization to dense joint (NTU-X [42]) and small joint (SHREC [10]) datasets.
- We perform extensive experiments and ablations to analyze and demonstrate the superiority of PSUMNet. (Sec. 3.4).
- We conclude our study by laying the ground for more challenging tasks within skeleton action recognition namely "In the wild" and "non-contextual" skeleton action recognition and introduce two new curated datasets: Skeletics-152 (See Sec. 4.2) and Skeleton-Mimetics (See Sec. 4.3) for these two problems respectively.

1.3 Thesis Layout

Chapter 2 draws attention to major shortcomings of existing widely used skeleton action datasets (NTURGB+D [35] and NTURGB+D 120 [25]) due to lack of finger level joint estimation. To remedy this, a new dense skeleton action dataset, NTU-X is introduced which provides a dense 118 joints pose tree. Extensive experiments and analysis are performed on this introduced dataset to convey its superiority over existing datasets.

Chapter 3 introduces a novel architecture - PSUMNet which proposes a Part streams unified modalities approach as opposed to conventional multi modality approaches. PSUMNet achieves outperforms other methods, which use $100\% - 400\%$ more parameters, across various skeleton action datasets.

Chapter 4 focuses on more challenging tasks of in the wild and non-contextual skeleton action recognition. Further, two new datasets Skeletics-152 and Skeleton-mimetics are introduced for these tasks respectively.

Chapter 5 concludes this thesis by providing closing remarks and motivation for future works.

Chapter 2

Dense pose based action recognition: A new direction

2.1 Introduction

The recent adoption of graph neural networks which process the skeleton action sequence as a spatio-temporal graph has enabled a steady rise in average accuracy for skeleton action recognition [49, 26, 8, 40, 38]. However, an analysis of sorted per-class accuracies reveals that the actions with lowest accuracy involve the usage of fingers (see Table 2.3,2.4). The underlying reason is that hand joints in Kinect-based skeletons provided in the original dataset are represented by just two finger joints (Figure 2.1-d). As a result, actions involving subtle finger movements (e.g. 'eating', 'writing', 'make ok sign', 'make victory sign') often fail to be recognized correctly. Sometimes, even the non-hand, main body joints are localized poorly by the Kinect-based capture system, as Figure 2.2 shows. These shortcomings at the raw data level cannot be addressed at the architecture level, i.e. by proposing novel architectures.

To address the mentioned data-level issues, we introduce NTU60-X and NTU120-X, curated and extended versions of the existing NTU dataset. Obtained from RGB videos present along with NTU skeleton data, the pose representations in the new dataset include 42 finger joints (21 for each hand), 51 facial keypoint joints and 25 body joints similar to those present in Kinect-based NTU-60 and NTU-120, for a total of 118 joints per skeleton (see Figure 2.1). We also modify state of the art approaches to enable experimental evaluation and benchmark the modified variants on NTU60-X and NTU120-X. As a result, we set the new state of the art benchmark on NTU-60 and NTU-120. Our results also demonstrate the benefit of the newly introduced datasets for overcoming the performance bottleneck mentioned earlier and enabling recognition of subtle human actions involving hand-based joints.

2.2 Related Work

Before the creation of NTU RGB-D dataset, a number of datasets enabled progress for skeleton-based human action recognition. MSR-Action3d [24] was one of the first action recognition datasets which provided depth and skeleton joint modalities, albeit from a single viewpoint. However, it only covered a limited set of gaming actions (e.g. forward punching, side boxing). The Northwestern-UCLA

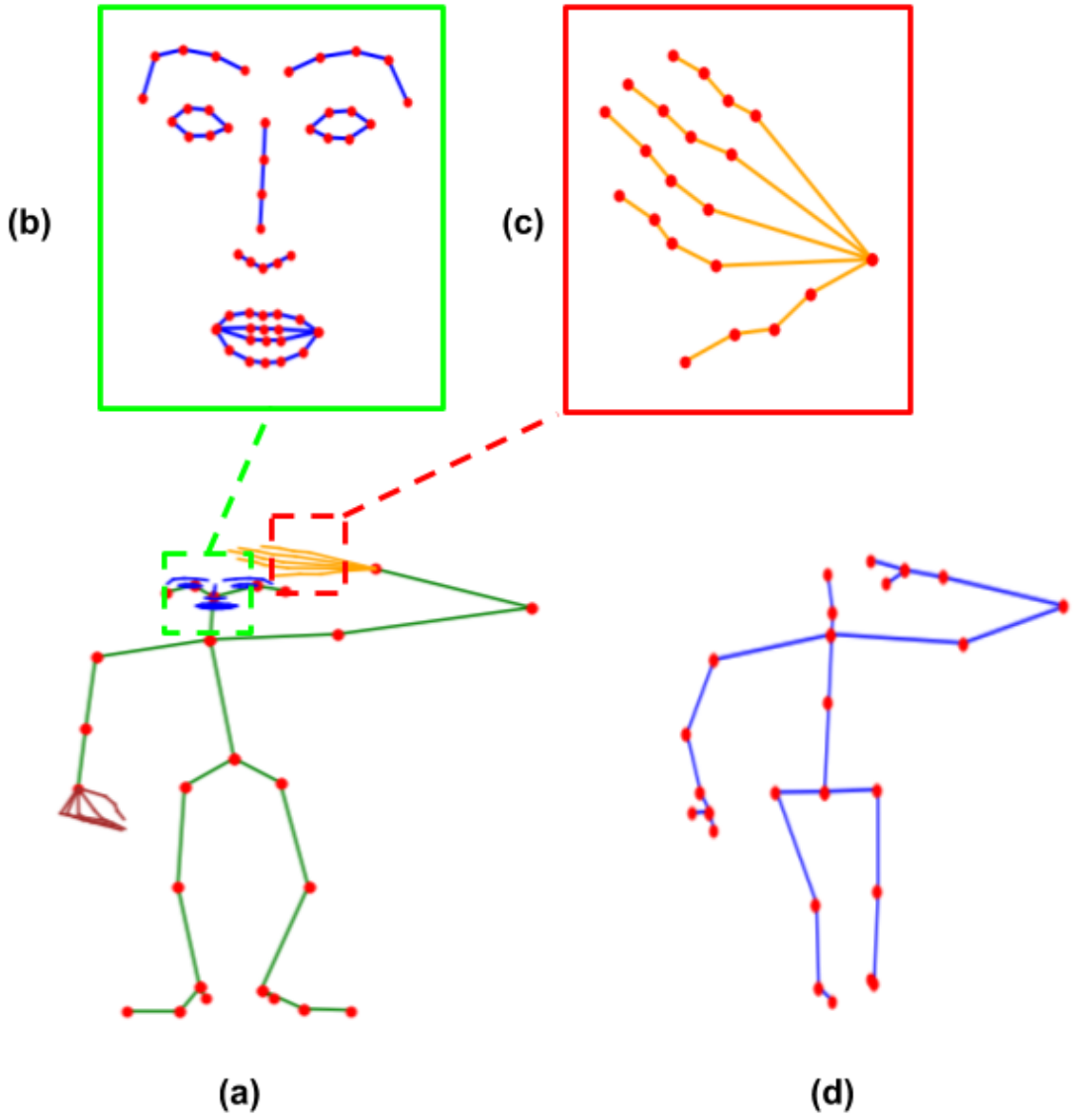


Figure 2.1 (a) The 118 joint skeleton introduced in the new NTU-X datasets. The 25 body joints are indicated by red dots.(b) 51 facial joints (c) 21 finger joints (d) 25 body joints present in original NTU datasets .

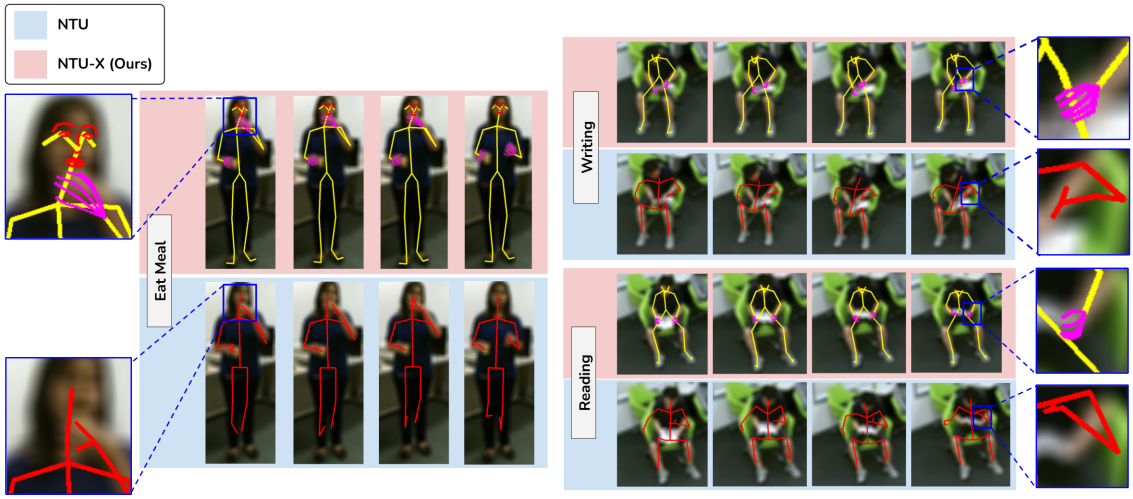


Figure 2.2 Sample skeletons from original NTU Kinect dataset (blue background) and proposed NTU-X dataset (pink background). Note that blurred RGB frame is included only for reference and is not part of skeleton data. The three classes mentioned - 'eat meal', 'writing' and 'reading' are a few of the most confused classes for the NTU dataset (see Table 2.3). As the zoomed insets illustrate, the quality of joints captured by NTU-X dataset is better compared to the original NTU dataset.

dataset [45] scaled up the diversity to include videos from multiple views and with actions performed by 10 different actors. The NTU RGB-D 60 dataset [35] comprises of 60 action categories, performed by 40 subjects. Its extension NTURGB-D 120 [25] is one of the largest and most diverse skeleton action dataset comprising of 120 actions performed by 106 subjects from 155 viewpoints.

Varying-view RGB-D Action Dataset (VAD) [20] comprises view-varying Kinect captured sequences covering the entire 360 view angles, containing 40 actions that are performed by 118 distinct performers. Unfortunately, the full dataset is not publicly available (as of current). Notably, the datasets mentioned above do not provide fine-grained joints for hands and faces which limits their utility for certain actions as mentioned previously.

An alternative approach for skeleton estimation infers the joints from RGB video frames without requiring specialized capture equipment. In the Kinetics-skeleton dataset [49], the 2D skeleton joint coordinates predicted from RGB frames are combined with the joint estimation confidence to obtain a pseudo 3D skeleton representation on videos from Kinetics-400 action dataset [4]. However, the resulting skeleton dataset contains many invalid sequences [14].

We summarize the salient aspects of these datasets and our proposed NTU60-X and NTU120-X in Table 2.1.

Dataset	Body	Face	Fingers	Sequences	Classes	Joints
MSR-Action3D[24]	✓			567	20	20
Northwestern-UCLA[45]	✓			1,475	10	21
VAD[20]	✓			25,600	40	25
NTU RGB+D[35]	✓			56,880	60	25
NTU RGB+D 120[25]	✓			114,035	120	25
NTU60-X (Ours)	✓	✓	✓	56,148	60	118
NTU120-X (Ours)	✓	✓	✓	113,821	120	118

Table 2.1 Comparison between NTU-X and some of the other publicly available skeleton-action recognition datasets. We are one of the first datasets to include body, face and hands joints in 3D for multi-person and occlusion case as well.

2.3 NTU-X

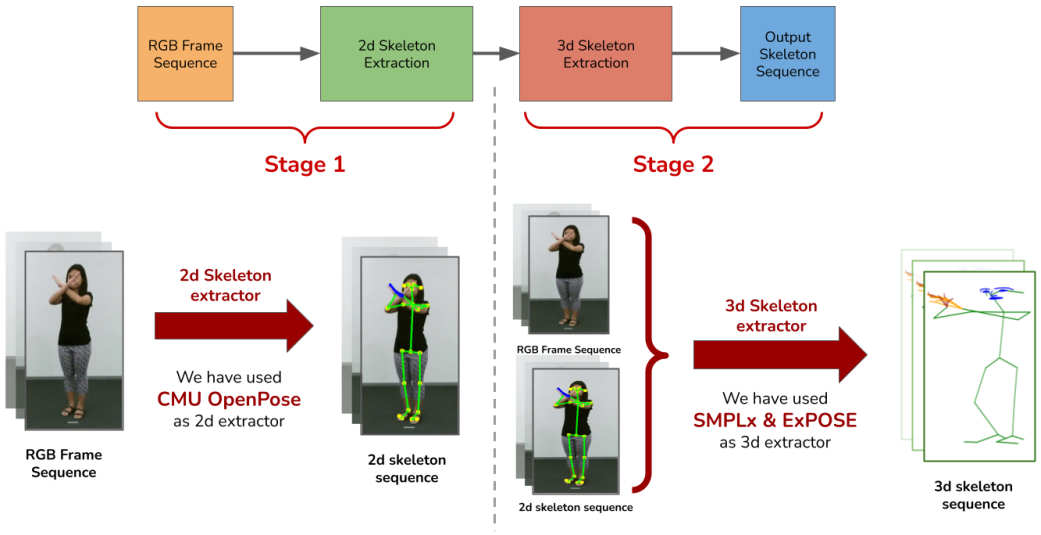


Figure 2.3 This figure gives the overview of the pipeline employed for the creation of proposed NTU-X dataset. In the stage 1 of the pipeline, using OpenPose[2], 2d skeleton in extracted from RGB frames. In stage 2 of the pipeline, using SMPL-X[30] or ExPose[9], 3d skeleton is extracted from RGB frames and 2d skeleton acquired in stage 1.

The NTU RGB+D dataset [35] provides RGB videos along with 3D Kinect skeleton data. We first extract the RGB frames from the videos at the frame rate of 30 FPS. We estimate 3D poses from RGB frames using SMPL-X [30]. SMPL-X uses strong 2D pose priors estimated using Openpose [2] on each RGB frame. However, SMPL-X based pose estimation is rather slow and is reliant on optimization heuristics. It also fails on blurred images and in the presence of light occlusion. To compensate for these issues, we use ExPose [9]. ExPose uses a part-wise attention-based model that feeds high resolution patches of the corresponding body parts to their dedicated refinement module. Unlike SMPL-X, ExPose estimates the full 3D pose (body, finger and face joints) from the RGB image without relying on 2D pose prior and is much faster compared to SMPL-X. The entire pipeline of NTU-X creation is illustrated in Fig. 2.3

To ensure good dataset quality, we remove corrupted videos from the original dataset, using a procedure similar to one adopted for the original dataset [35]. We also omit videos in which people are completely absent. Additionally, for some samples OpenPose provides poor estimates and hence we discard instances of such videos as well.

2.4 Experiments

To evaluate the impact of NTU60-X and NTU120-X on overall performance, we benchmarked models with state-of-the-art performance on NTU60 and NTU120. We selected DSTA-Net[38], 4s-ShiftGCN [8], MS-G3D [26] and PA-ResGCN [40] as the models to benchmark the newly introduced datasets. For the models DSTA-Net[38], MS-G3D [26] and 4s-ShiftGCN [8], we updated the graph structure of the skeletons to incorporate the newly introduced joints. Figure 2.1(d) shows the skeleton topology for the original kinect data. We changed the input graph topology for these models according to our new skeleton structure as shown in Figure 2.1(a).

PA-ResGCN [40], being a semantic part-based model, required more significant modification. Along with changes in input skeleton graph structure as done for the other two models, we defined new parts to incorporate the newly introduced joints and thus enable richer feature extraction. Since this model learns attentive weights for each of the input skeleton joints by dividing the skeleton into different parts, the definition of parts was also changed based on the NTU-X skeleton. In case of NTU-RGBD skeleton, PA-ResGCN defines total 5 parts: torso, left arm, right arm, left leg and right leg. In the new NTU-X skeleton, 3 additional parts were defined for 67 joints (body + fingers) skeleton: left fingers, right fingers and head, resulting in a total of 8 parts. For 118 joint (body + fingers + face) skeleton along with these 3 additional parts, one more part of face was added resulting in a total of 9 parts.

2.4.1 Results

The results of training the four selected models on the new NTU60-X and NTU120-X datasets, with finger joints included for Cross Subject protocol, are shown in Table 2.2. Clearly, our modi-

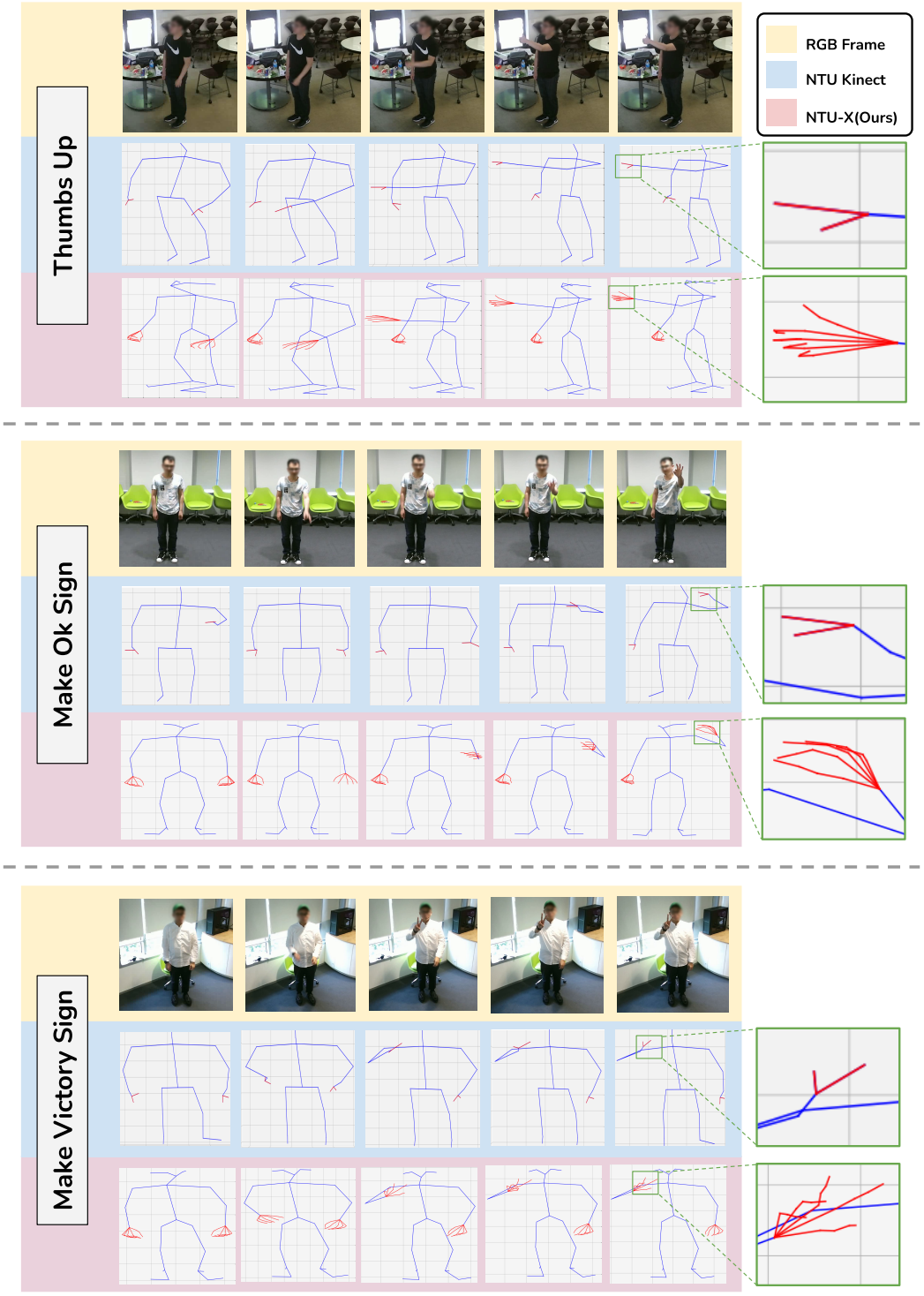


Figure 2.4 Samples showing 3d plot of the original NTU kinect skeletons and newly proposed NTU-X skeletons with corresponding RGB frames. The zoomed insets show the finger joints estimated in both NTU-Kinect and NTU-X and it clearly shows that NTU-X represents the action much more comprehensively than original NTU-Kinect data.

model	NTU60	NTU60-X (Ours)	NTU120	NTU120-X (Ours)
DSTA-Net[38]	91.50	**93.56**	86.60	**87.80**
4s-ShiftGCN[8]	90.70	**91.78**	85.90	**86.18**
MS-G3D[26]	91.50	**91.76**	86.9	**87.10**
PA-ResGCN[40]	90.9	**91.64**	**87.4**	86.42

Table 2.2 Results for top performing models of NTU60 and NTU120 dataset on NTU60-X dataset and NTU120-X (with finger joints) - see Section 2.4.1. The gray shaded columns show results on our newly introduced dataset. The blue highlighted cell corresponds to best overall performance for 60 and 120 class setups.

fied DSTA-Net, MS-G3D, 4s-Shift-GCN and PA-ResGCN outperform their counterparts' performance on the original NTU60 dataset by a significant margin. For NTU120-X, all three models except PA-ResGCN outperform their counterparts' performance on NTU120 dataset. PA-ResGCN fails to surpass the orginal accuracy for 120 class dataset by a small margin. We hypothesize that this could be due to PA-ResGCN's architecture being too specific for the original Kinect skeleton setup and unable to handle the addition of extra added finger joints in the large-category (120 class) setting. We can also see that DSTA-Net not only beats its numbers on the original dataset, but also achieves state of the art performance among all the models with a margin of more than 2% for NTU60 dataset.(See highlighted cells in Table 2.2).

These results also support the fact that existing approaches, if provided with better and richer joint data, have the capacity to perform better. A detailed analysis of each model's performance and category level improvements is discussed next.

2.4.2 Discussion

Table 2.3 and Table 2.4 list the five worst performing classes for all the four models on the original NTU60 and NTU120 datasets respectively along with their per class accuracy. The shaded columns in these tables provide the accuracy of these classes when the models are trained using the newly introduced NTU60-X and NTU120-X dataset. From these results, it is evident that most of the bottom performing classes for the original NTU datasets involve actions with fine finger movements (e.g. "writing", "type on keyboard", "eat meal", "make ok sign", "make victory sign"). When the models are provided with input data that includes finger joints, the per class accuracy for such categories is improved significantly. Figure 2.2 and Figure 2.4 also show that without inclusion of finger level joints, recognition of such action categories is ambiguous and difficult.

Model	Class name	NTU-60	NTU60-X
DSTA-Net [38]	Writing	67.04 %	79.41 %
	Reading	68.75 %	94.49 %
	Play with phone/tablet	71.06 %	86.91 %
	Type on a keyboard	71.64 %	93.45 %
	Sneeze or cough	73.55 %	80.07 %
4s-ShiftGCN [8]	Writing	65.19 %	76.23 %
	Reading	68.75 %	91.91 %
	Eat meal	72.89 %	80.22 %
	Type on keyboard	73.82 %	91.24 %
	Sneeze or cough	73.91 %	75.72 %
MS-G3D [26]	Writing	57.41 %	72.96 %
	Eat meal	71.43 %	79.85 %
	Reading	72.43 %	92.28 %
	Sneeze or cough	77.17 %	80.80 %
	Play with phone/tablet	78.75 %	79.41 %
PA-ResGCN [40]	Writing	63.97 %	78.89 %
	Reading	67.65 %	94.12 %
	Sneeze or cough	73.91 %	76.45 %
	Type on keyboard	74.91 %	91.61 %
	Eat meal	74.91 %	80.95%

Table 2.3 The NTU60 column shows accuracies of bottom 5 action classes for models trained on original NTU60 dataset. The NTU60-X column shows accuracies of the same classes but with models trained on our NTU60-X dataset (finger joints: Section 2.3). Thanks to availability of additional finger joint information in NTU-60X, we see visible performance improvement across all the models.

Model	Class name	NTU-120	NTU120-X
DSTA-Net [38]	Staple book	37.65 %	36.60 %
	Make ok sign	51.13 %	72.70 %
	Make victory sign	53.85 %	60.49 %
	Counting money	59.65 %	84.91 %
	Blow nose	64.94 %	70.73 %
4s-ShiftGCN [8]	Staple book	35.9 %	34.19 %
	Make victory sign	53.32 %	62.59 %
	Make ok sign	55.83 %	64.35 %
	Counting money	61.93 %	82.11 %
	Blow nose	63.07 %	67.60 %
MS-G3D [26]	Staple book	32.57 %	34.50 %
	Make victory sign	54.02 %	68.18 %
	Hit with object	60.03 %	69.98 %
	Blow nose	60.45 %	68.64 %
	Counting money	60.70 %	90.53 %
PA-ResGCN [40]	Staple book	40.63 %	36.08 %
	Make victory sign	59.79 %	60.49 %
	Hit with object	61.90 %	63.53 %
	Cutting paper	63.30 %	55.32 %
	Counting money	64.65 %	78.25 %

Table 2.4 The NTU120 column shows accuracies of bottom 5 action classes for models trained on original NTU120 dataset. The NTU120-X column shows accuracies of the same classes but with models trained on our NTU120-X dataset (finger joints: Section 2.3). Thanks to availability of additional finger joint information in NTU-120X, we see visible performance improvement across all the models.

	Bottom-5 NTU60-X	Bottom-5 NTU120-X
DSTA-Net [38]	Writing (79.41 %) Eat meal (80.0 %) Sneeze or cough (80.07 %) Touch head (headache) (80.73 %) Take off a shoe (84.67 %)	Staple book (36.6 %) Make victory sign (60.49 %) Cutting paper (61.26 %) Fold paper (68.17 %) Play magic cube (69.47 %)
4s-ShiftGCN [8]	Sneeze or cough (75.72 %) Writing (76.3 %) Eat meal (80.22 %) Touch head (headache) (80.36 %) Playing with phone/tablet (80.51 %)	Staple book (34.15 %) Cutting paper (52.01 %) Playing with phone/tablet (60.73 %) Fold paper (61.22 %) Make victory sign (62.59 %)
MS-G3D [26]	Writing (72.59 %) Eat meal (79.12 %) Wear a shoe (80.07 %) Punching/slapping other person (80.66 %) Sneeze or cough (81.52 %)	Staple book (34.5 %) Fold paper (61.91 %) Cutting paper (63.18 %) Playing with phone/tablet (64.73 %) Make ok sign (65.91 %)
PA-ResGCN [40]	Sneeze or cough (76.45 %) Writing (78.89 %) Touch head (headache) (78.91 %) Punching/slapping other person (80.66 %) Eat meal (80.95 %)	Staple book (38.53 %) Cutting paper (55.32 %) Make victory sign (61.19 %) Playing with phone/tablet (61.45 %) Fold paper (62.78 %)

Table 2.5 This table shows the bottom performing classes for all the models evaluated in this paper on the newly introduced NTU60-X and NTU120-X datasets. This table clearly indicates that the overall accuracy of bottom performing classes for the newly introduced NTU60-X and NTU120-X are higher than the overall accuracy of bottom performing classes for the original datasets, as shown in Table 2.3 and Table 2.4

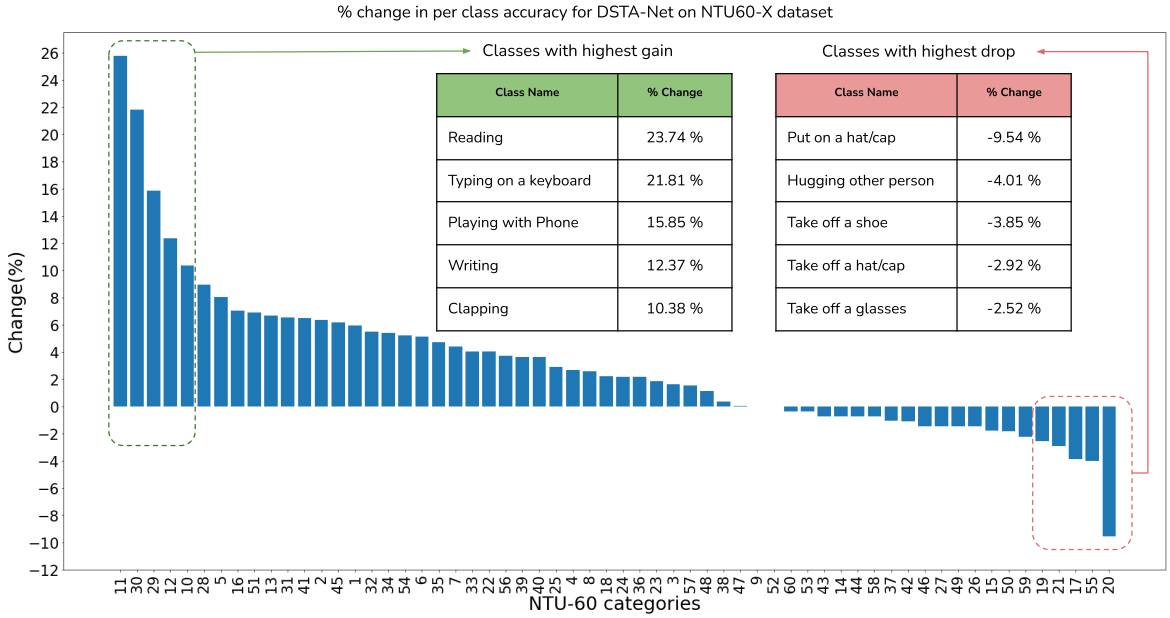


Figure 2.5 The % gain in per class accuracy for best performing model (DSTA-Net) after training on newly introduced NTU60-X dataset. The x-axis shows category id. The inset tables show actions with largest and least gain.

Table 2.5 shows the worst performing classes for the NTU-X dataset for all the models. Comparing with accuracy of worst performing classes of the original NTU dataset given in Table 2.3 and Table 2.4, we see that even the accuracy of worst performing classes of NTU-X dataset is, on average, higher compared to the original NTU dataset.

To further illustrate the performance boost we gain by including the finger level joints into the input skeleton, we show the change in per class accuracy when going from original NTU dataset to newly introduced NTU-X dataset in Figure 2.5 and Figure 2.6 for the top performing model DSTA-Net[38](based on Table 2.2).

The table provided in the inset of Figure 2.5 shows classes which are benefited the most when going from NTU60 to NTU60-X dataset for the state-of-the-art performer (DSTA-Net). It is easy to see that these classes predominantly involve finger level actions such as “reading”, “typing on keyboard”, “playing with phone” and “writing”. One can also observe that the gain for these classes is as high as **10-23 %**. A similar trend can be seen in Figure 2.6 which shows classes that benefit the most when going from NTU120 to NTU120-X dataset. Once again, the classes with highest gain involve finger level actions (e.g. “make of sign”,“hush”) and the gain for these classes is in the range **17-25%**. Both of these tables clearly indicate training recognition approaches on our proposed NTU-X dataset significantly boost the performance of classes involving finger movements.

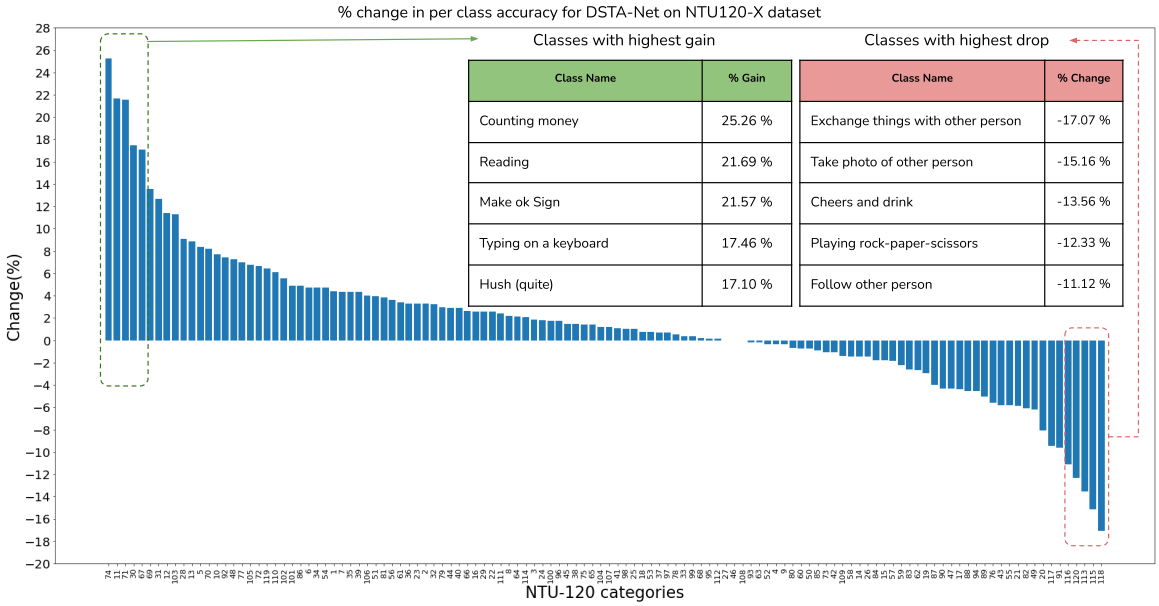


Figure 2.6 The % gain in per class accuracy for best performing model (DSTA-Net) after training on newly introduced NTU120-X dataset. The x-axis shows category id. The inset tables show actions with largest and least gain.

From Figure 2.5 and Figure 2.6, we also note that training on NTU-X dataset (finger joints) is not beneficial for all the classes, with a performance drop seen in some cases. Most of these classes involve another person (e.g. "hugging other person", "take photo of other person", "playing rock-paper scissors" etc.). As per our understanding, capturing accurate skeletons for multiple people in a RGB frame is difficult which leads to poor pose extraction and ambiguity in classifying the sequences involving multiple people.

However, it is clear that the overall magnitude of gain is higher than the magnitude of drop in per class accuracy. Hence, the average accuracy is higher for NTU-X dataset than the original NTU dataset.

2.4.3 Ablation Study

To examine the importance of body joints, finger joints and face joints individually, we also perform experiments with only body joints (25 joints), body + finger joints (67 joints) and body + fingers + face joints (118 joints) as well. The results of ablation study are shown in Table 2.6. The performance degrades when face joints are included with the body and finger joints. One reason for this could be that the actions in NTU dataset do not involve significant facial motion. Hence, the additional joints of the face make the skeleton graph larger than necessary and difficult for model optimization. Another possible reason could be that the existing models do not have a suitable architecture to handle the dense

	NTU60-X			NTU120-X	
model	body	body + fingers	body + fingers + face	body	body + fingers
DSTA-Net[38]*	89.69	**90.91**	88.97	84.82	**87.80**
4s-ShiftGCN[8]	89.56	**91.78**	89.64	84.78	**86.18**
MS-G3D[26]	91.26	**91.76**	91.12	84.50	**87.10**
PA-ResGCN[40]	89.98	**91.64**	89.79	82.85	**86.42**

Table 2.6 Results on different variants of NTU60-X and NTU120-X dataset to understand the contribution of the additional joints. (*: Ablations on DSTA-Net are done using only the Joint stream of the network which contributes most to its performance.)

subgraph arising from the presence of facial keypoints. The poor results of models trained with only body joints (25 joints) are also in line with our hypothesis that the inclusion of finger joints in the input skeleton is crucial for better performance. In other words, the performance gain is not merely due to the shift from Kinect-based to RGB-based skeleton generation process.

2.5 Limitations

As shown in Sec. 3.4.3, our introduced NTU60-X and NTU120-X provide gains in overall accuracy by introducing dense input skeleton to better capture subtle finger movements in actions. However, it can be argued that these gains are not significant given that the increase in input joints is very prominent. (25 joints in NTURGB+d against 67 or 118 joints in NTU-X datasets) One of the reasons could be the noise introduced by the pose estimators (i.e ExPose or SMPLx). Since these pose estimators are not always providing most accurate skeletons and are prone to noise when the input RGB frames have occlusion.

Further, almost all the existing architectures are essentially crafted to handle sparse input skeleton. These model when provided with a much denser input graph might not provide as significant jump as expected.

These limitations cause the gap in overall accuracy between our NTU-X and original NTURGB+d dataset to be not significant in some cases. However as shown in Tab. 2.3 and Tab. 2.4, NTU-X datasets show significant improvements for the actions which involve prominent finger level movements.

2.6 Conclusion

In this chapter, we have shown that the lack of hand-level joints is a fundamental performance bottleneck in the skeleton data of the largest action recognition dataset, NTU-RGBD. To address this bottleneck, we contribute a carefully curated skeleton dataset that provides finger-level hand joints and facial keypoint joints.

We appropriately modify the state of the art approaches to enable training using the introduced dataset. Our results demonstrate the effectiveness of the proposed dataset in enabling the modified approaches to overcome the aforementioned bottleneck and improve their performance, overall and on previously worst performing action categories. We also perform experiments to evaluate the relative importance of the introduced joints. We believe our contribution of a new, expanded joint dataset will meet the twin objectives of improving performance and encouraging novel approaches in future. Going forward, we expect the research community to devise novel and efficient approaches for tackling dense skeleton representations present in our dataset.

Our 118 joints dataset, consisting of full body, fingers and even face joints can improve the recognition of actions based on expressions. This can help in capturing subtle changes in expression that would help in recognizing fine-grained actions (e.g. 'moving head up' and 'shaking head'). The significance of our work also arises from the emerging trend of fusing skeletal representations with other modalities (depth, RGB) for better performance in out of context, in-the-wild action recognition scenarios [46, 14, 27]. The pre-trained deep networks we introduce serve as a good starting point for such fusion based approaches.

2.7 Acknowledgements

(Portions of) the research in this paper used the NTU RGB+D (or NTU RGB+D 120) Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

In the next chapter we propose a highly efficient and generalizable approach for sparse (NTURGB+d) and Dense (NTU-X) skeleton action recognition.

Chapter 3

Unified Modality Part Streams approach for Efficient Pose-based Action Recognition

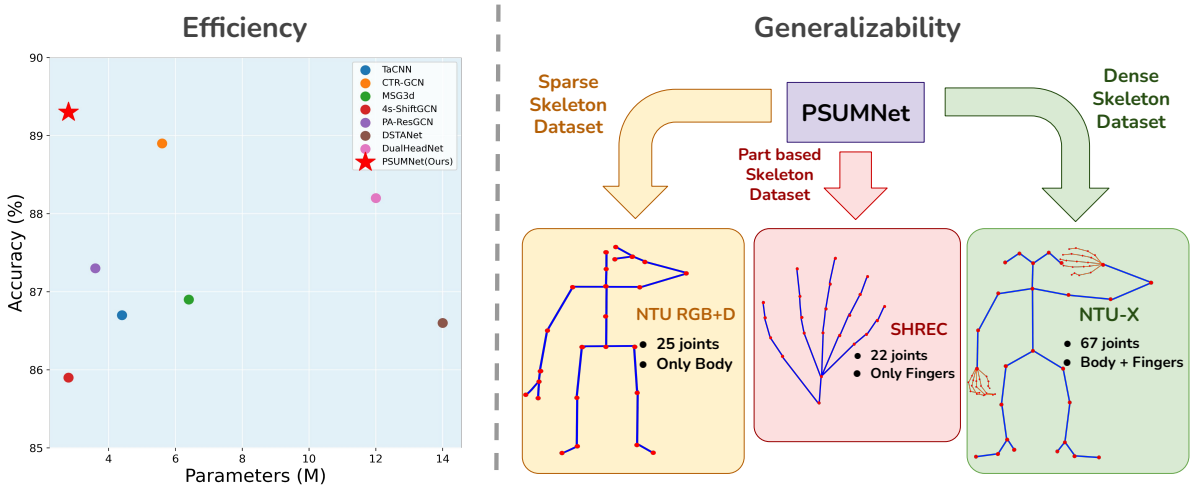


Figure 3.1 The plot on left shows accuracy against number of parameters for our proposed architecture PSUMNet ($red\star$) and existing approaches for the large-scale NTURGB+D 120 dataset (cross subject) of human actions. PSUMNet achieves state of the art performance while competing recent methods use 100%-400% more parameters. The diagram on right is to illustrate that PSUMNet scales to part based sparse pose (SHREC [10]) and dense pose (NTU-X [42]) configurations in addition to the popular NTURGB+D[35, 25] configuration.

3.1 Introduction

A plethora of RNN [12, 15], CNN [52, 16] and GCN [49, 26] based approaches have been proposed to tackle this important problem of skeleton based human action recognition at scale The success of

approaches such as ST-GCN [49] which modeled spatio-temporal joint dynamics using GCN has given much prominence to GCN-based approaches. Furthermore, approaches such as RA-GCN [39] and 2s-AGCN[37] built upon this success and demonstrated additional gains by introducing multi modal (bone and velocity) streams – see Fig. 3.2 (left). This multi stream approach has been adopted as convention by state of the art approaches.

However, the conventional setup has three major drawbacks. *First,* each modality stream is trained independently and the results are combined using late (decision) fusion. This deprives the processing pipeline of taking advantage of correlations across modalities. *Second,* with the addition of each new modality, the number of parameters increases by a significant margin since a separate network with the same model architecture is trained for each modality. *Third,* the skeleton is considered in a monolithic fashion. In other words, the entire input pose tree at each time step is treated as a whole and at once. This is counterintuitive to the fact that a lot of action categories often involve only a subset of the available joints. For example, action categories such as "Cutting paper" or "Writing" can be easily identified using only hand joints whereas action categories such as "Walking" or "Kicking" can be easily identified using only leg joints. Non-monolithic approaches which decompose the pose tree into disjoint part groups do exist [41, 40]. However, each part group is not considered within the global pose frame, causing such methods to fall short. Additionally, monolithic processing increases compute requirements when the number of joints in the pose representation increases [42].

Our proposed approach tackles all of the aforementioned drawbacks - see Fig. 3.2 (right). Overall, the high accuracy provided by PSUMNet, coupled with its efficiency in terms of compute (number of parameters and floating-point operations) makes our approach an attractive choice for real world deployment on compute restricted embedded and edge devices. Code and models will be released.

3.2 Related Work

Skeleton action recognition: Since the release of large scale skeleton based datasets [35, 25] various CNN [52, 16, 23], RNN [12, 15, 52, 53] and recently GCN based methods have been proposed for skeleton action recognition. ST-GCN [49] was the first successful approach to model the spatio-temporal relationships for skeleton actions at scale. Many state of the art approaches [8, 26, 38, 6] have adopted and modified this approach to achieve superior results. However, these approaches predominantly process the skeleton joints in a monolithic manner, i.e these approaches process the entire input skeleton at once which can create a bottleneck when the input skeleton becomes denser, e.g. NTU-X [42].

Part based approaches: The idea of grouping skeleton joints into different groups has few precedents. Du et al. [12] propose a RNN-based hierarchical grouping of part group representations. Thakkar et al. [41] propose a GCN based approach which applies modified graph convolutions to different part groups. Huang et al. [18] propose a GCN-based approach in which they utilize the higher order part level graph for better pooling and aligning of nodes in the main skeleton graph. More recently, Song et al. [40] propose a part-aware GCN method which utilizes part factorization to aid an attention mechanism to find

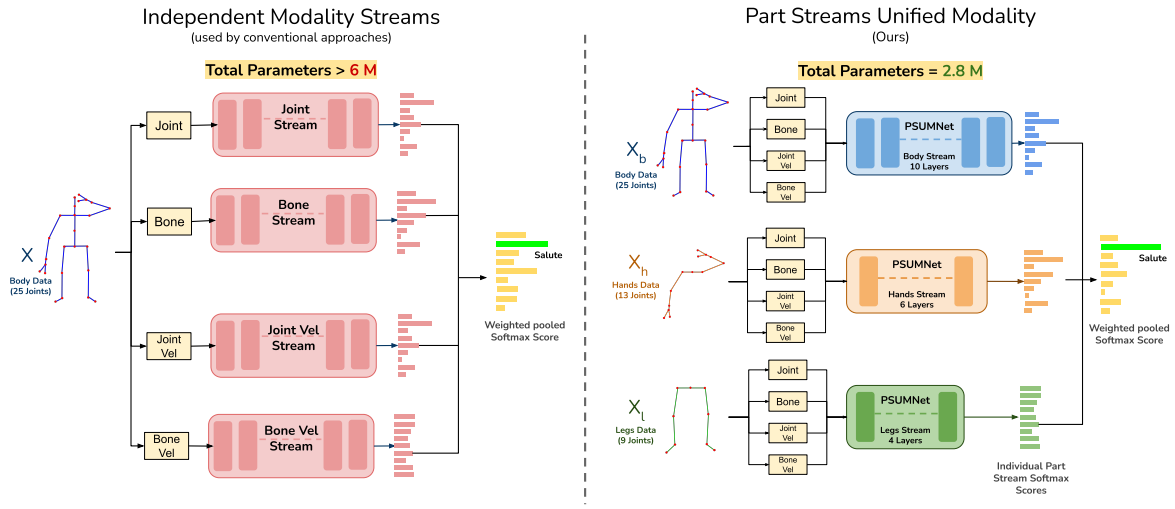


Figure 3.2 Comparison between conventional training procedure used in most of the previous approaches (left) and our approach (right). Conventional methods [6, 26] use dedicated independent streams and train separate instances of the same network for each of the four modalities, i.e joint, bone, joint velocity and bone velocity. This method increases the number of total parameters by a huge margin and involves a monolithic representation. Our method processes the modalities in a unified manner and creates part group based independent stream with a superior performance compared to existing methods which use 100%-400% more parameters - see Fig. 3.3 for architectural details of PSUMNet.

the most informative part. Some previous part based approaches segment the limbs based on left and right orientation as well (left/right arm, left/right leg etc.) [40, 41]. Such segmentation leads to disjoint part groups which contain very small number of joints and are unable to convey useful information. In contrast, our part stream approach creates overlapping part groups with sufficient number of joints to model useful relationships. Also, each individual part group in our setup is registered to the global frame unlike the per-group coordinate system setup in existing approaches. In addition, we employ a combination of part group and coarse version of the full skeleton instead of part-group only approach seen in existing approaches. Our part stream approach allows each part based sub-skeleton to contribute towards the final prediction via decision fusion. To the best of our knowledge, such globally registered independent part stream approach has never been used before.

Multi stream Training: Earlier approaches [37, 8, 26] and more recent approaches [6, 48] create multiple modalities termed joint, bone and velocity from the raw input skeleton data. The conventional method is to train the given architecture multiple times using different modality data followed by decision fusion. However, this conventional approach with multiple versions of the base architecture greatly increases the total number of parameters. Song et al. [40] attempt a unified modality pipeline wherein early fusion of different modality streams is used to achieve a unified modality representation. However,

before the fusion, each modality is processed via multiple independent networks which again increases the count of trainable parameters.

3.3 Methodology

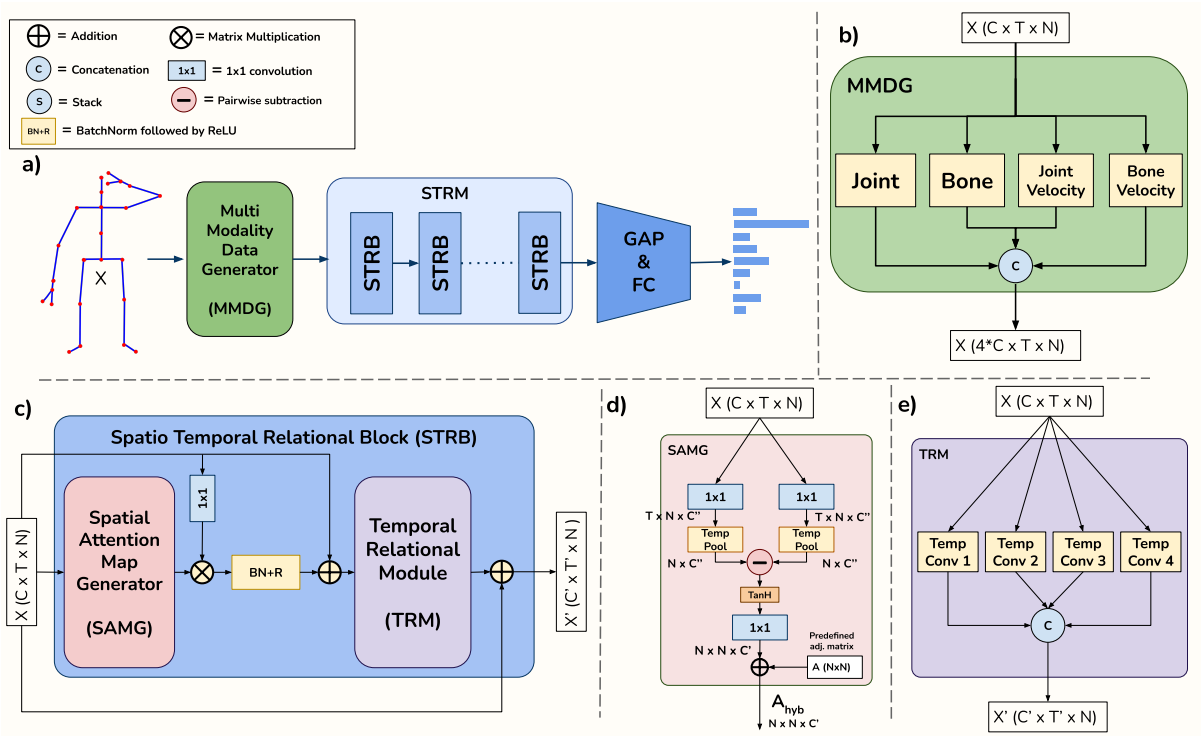


Figure 3.3 (a) Overall Architecture of one stream of the proposed architecture. The input skeleton is passed through Multi modality data generator (MMDG), which generates joint, bone, joint velocity and bone velocity data from input and concatenates each modality data into channel dimension as shown in (b). This multi-modal data is processed via Spatio Temporal Relational Module (STRM) followed by global average pooling and FC. (c) Spatio Temporal Relational Block (STRB), where input data is passed through Spatial Attention Map Generator (SAMG) for spatial relation modeling, followed by Temporal Relational Module. As shown in (a) multiple STRBs stacked together make the STRM. (d) Spatial Attention Map Generator (SAMG), dynamically models adjacency matrix (A_{hyb})to model spatial relations between joints. Predefined adjacency matrix (A) is used for regularization. (e) Temporal Relational Module (TRM) consists of multiple temporal convolution blocks in parallel. Output of each temporal convolution block is concatenated to generate final features.

We first describe our approach for factorizing the input skeleton into part groups and a coarser version of the skeleton (Sec. 3.3.1). Subsequently, we provide the architectural details of our deep network PSUMNet which processes these part streams (Sec. 3.3.2).

3.3.1 Part Stream Factorization

Let X ($\in \mathbb{R}^{3\times T\times N}$) represent the T-frames, N-joint skeleton configuration of a 3D skeleton pose sequence for an action. We factorize X into following three part groups – see Fig. 3.2 (right):

1. **Coarse body** (X_b): This is comprised of all joints in the original skeleton for NTURGB+D skeleton topology, 25 joints in total. For the 67-joint dense skeleton topology of NTU-X [42], this stream comprises of all the body joints but without the intermediate joints of each finger for each hand. Specifically, only 6 joints out of 21 finger joints are considered per hand resulting in total of 37 joints for NTU-X.

2. **Hands** (X_h): This contains all the finger joints in each hand and the arm joints. Note that the arms are rooted at the throat joint. For NTURGB+D dataset, the number of joints for this stream is 13 and for NTU-X, the total number of joints is 48.

3. **Legs** (X_l): This includes all the joints in each of the legs. Similar to X_h, the leg joints share the common hip joint. For NTURGB+D dataset the number of joints for this stream is 9 and for NTU-X, the total number of joints is 13.

We evaluate performance of PSUMNet on mainly two pose action datasets - NTURGB+D [35, 25] and NTU-X [42]. These datasets contain 25 joints and 67 joints skeleton representation respectively. Fig. 3.4, provides visual representation of how skeleton from both these datasets are divided into different part groups and used to train different part streams of PSUMNet.

As shown in Fig. 3.2 (right), the part group sub skeletons are used to train three independent streams of our proposed PSUMNet (Sec. 3.3.2). As explained previously, our hypothesis is that many of the action categories are dominated by certain part groups and hence can be classified using only a subset of the entire input skeleton. To leverage this, we perform late decision fusion by performing a weighted average of the prediction scores from each of the part streams to obtain the final classification. The weights used for each part stream scores are hyper parameters determined experimentally. Crucially, we change the number of layers in each of the streams in proportion to number of input joints - please see Tab. 3.1. This helps restrict the total number of parameters used for the entire protocol.

In contrast with other part based approaches, the part groups in our setting are not completely disjoint. More crucially, the part groups are defined with respect to a shared global coordinate space. Though seemingly redundant due to common joints across part groups, this design choice enables global motion information to propagate to the corresponding groups. Another significant advantage of such part stream approach is the better scalability to much denser skeleton datasets such as NTU-X [42] and to sparser datasets such as SHREC[10].

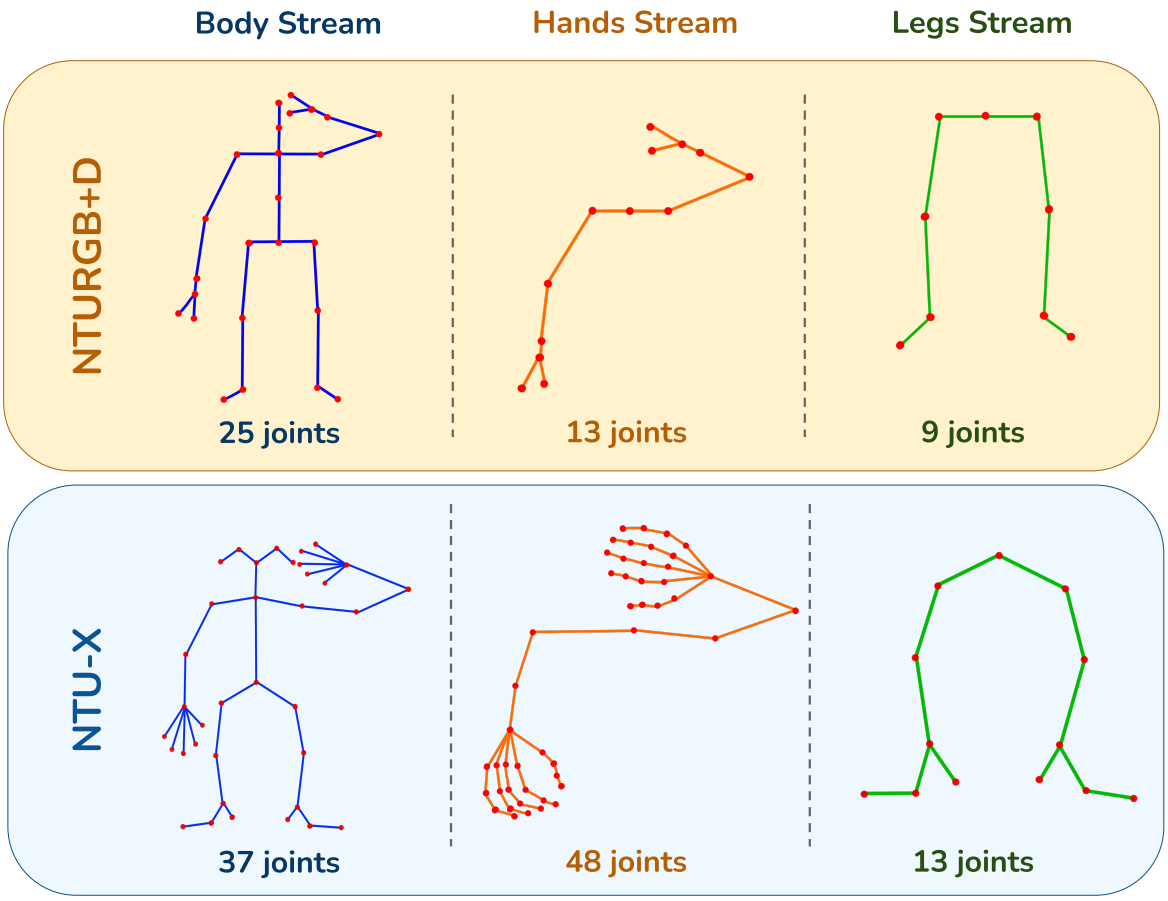


Figure 3.4 The part group factorization of 25 joints skeleton dataset NTURGB+D and 67 joints skeleton dataset NTU-X, used to train different part stream of PSUMNet.

3.3.2 PSUMNet

In what follows, we explain the architecture of a single part stream of PSUMNet (e.g. $X = X_b$) since all part streams essentially use the same architecture. An overview of PSUMNet's single stream architecture can be seen in Fig. 3.3 (a). First, the input skeleton X is passed through Multi Modality Data Generator (MMDG) to create a rich modality aware representation. This feature representation is processed by Spatio-Temporal Relation Module (STRM). Global average pooling (GAP) of the processed result is transformed via fully connected layers (FC) to obtain the per-layer prediction for the single part stream.

Next, we provide details for various modules included in our architecture.

3.3.3 Multi Modality Data Generator (MMDG)

As shown in Fig. 3.3 (b), this module processes the raw skeleton data and generates the corresponding multi modality data, i.e. joint, bone, joint-velocity and bone-velocity. The joint modality is the raw skeleton data represented by $X = \{x \in \mathbb{R}^{C \times T \times N}\}$, where C, T and N are channels, time steps and joints. The bone modality data is obtained using the following equation:

$$X_{bone} = \{x[:, :, i] - x[:, :, i_{nei}] \,|\, i = 1, 2, ..., N\} \tag{3.1}$$

where i_{nei} denotes neighboring joint of i based on predefined adjacency matrix. Next we create joint-velocity and bone-velocity modality data using following equations:

$$X_{joint-vel} = \{x[:, t+1, :] - x[:, t, :] \,|\, t = 1, 2, ..., T,\ x \in X_{joint}\} \tag{3.2}$$

$$X_{bone-vel} = \{x[:, t+1, :] - x[:, t, :] \,|\, t = 1, 2, ..., T,\ x \in X_{bone}\} \tag{3.3}$$

Finally, we concatenate all these four modality data into channel dimension to generate $X \in \mathbb{R}^{4C \times T \times N}$ which is fed as input to the network. Concatenating the modality data helps model the inter-modality relations in a more direct manner.

3.3.4 Spatio Temporal Relational Module (STRM)

This modality aware representation obtained from MMDG is processed by the Spatial Temporal Relational Module (STRM) as shown in Fig. 3.3 (a). STRM consists of multiple Spatio Temporal Relational Blocks (STRB) stacked one after another. The architecture of a single STRB is shown in Fig. 3.3 (c). Each STRB block contains a Spatial Attention Map Generator (SAMG) to dynamically model different spatial relations between joints followed by Temporal Relational Module (TRM) to model temporal relations between joints.

Spatial Attention Map Generator (SAMG): We dynamically model an Spatial Attention Map for the spatial graph convolutions [6, 36]. As shown in Fig. 3.3 (d), we pass the input skeleton through two parallel branches, each consisting a 1×1 convolution and a temporal pooling block. We pair-wise subtract outputs from the parallel branches to model the Attention Map. We add the predefined adjacency matrix A as a regularization to the Attention Map to generate the final hybrid adjacency matrix A_{hyb}, i.e.

$$A_{hyb} = \alpha M(X_{in}) + A \tag{3.4}$$

where α is a learnable parameter and A is the predefined adjacency matrix. M is defined as:

$$M(X_i) = \sigma(TP(\phi(X_{in})) - TP(\psi(X_{in}))) \tag{3.5}$$

where σ, ϕ and ψ are 1x1 convolutions, TP is temporal pooling.

Part Stream	# of STRBs	STRB output dimension
Body Stream	10	64, 64, 64, 64, 128, 128, 128, 512, 512, 512
Hands Stream	6	64, 128, 128, 128, 512, 512
Legs Stream	4	64, 128, 128, 512

Table 3.1 Architecture details for each part stream of PSUMNet. Last column denotes output dimensions of consecutive STRBs for each part stream.

Once we obtain this adjacency matrix A_{hyb}, we pass the original input through a 1×1 convolution and multiply the results with the dynamic adjacency matrix to characterize the spatial relations between the joints as follows:

$$X_{out} = A_{hyb} \bigotimes (\theta(X_{in})) \tag{3.6}$$

where θ is 1x1 convolution block. $\bigotimes$ is matrix multiplication operation.

Temporal Relation Module (TRM): We use multiple parallel convolution blocks to model the temporal relation between the joints of the input skeleton as shown in Fig. 3.3 (e). Each temporal convolution block is a standard 2D convolution with varying kernel sizes in temporal dimension and with dilation. This helps model temporal relations at multiple scales. For our experiments, TRM consist of four parallel temporal convolution block. TempConv1 and TempConv2 consist of a 1x1 convolution followed by batch normalization, ReLU and a 2d convolution with temporal kernel size of 5. The value of dilation for these blocks are 1 and 2 respectively. TempConv3 consist of a 1x1 convolution followed by batch normalization, ReLU and a 2d convolution with temporal kernel size 3 and a max pooling layer. TempConv4 consists of just 1x1 convolution followed by batch normalization and ReLU. The outputs from each of the temporal convolution blocks are concatenated. The result is processed by GAP and FC layers and mapped to a prediction (softmax) layer as mentioned previously.

Since each part group (body, hands, legs) contains significantly different number of joints, we adjust the number of STRBs and depth of the network for each stream accordingly as shown in Fig. 3.2 (Right). This design choice provides two advantages. *First,* it reduces the total number of parameters by a huge margin. *Second,* adjusting the depth of the network in proportion to the joint count enables richer dedicated representations for actions whose dynamics are confined to the corresponding part groups, resulting in better performance overall. Details about output dimensions of each STRB in each part stream can be found in Tab. 3.1.

3.4 Experiments

3.4.1 Datasets

NTURGB+D[35] is a large scale skeleton action recognition dataset with 60 different actions performed by 40 different subjects. The dataset contains 25 joints human skeleton captured using Microsoft Kinect V2 cameras. There are a total of 56,880 action sequences. There are two evaluation protocols for this dataset - First, Cross Subject (XSub) split where action performed by 20 subjects falls into training set and rest into the test set. Second, Cross View (XView) protocol where actions captured via camera ID 2 and 3 are used as training set and actions captured via camera ID 1 are used as test set.
NTURGB+D 120[25] is an extension of NTURGB+D dataset with additional 60 action categories and a total of 113,945 action samples. The actions are performed by a total of 106 subjects. There are two evaluation protocols for this dataset - First, Cross Subject (XSub) split where action performed by 53 subjects falls into training set and rest into the test set. Second, Cross Setup (XSet) protocol where actions even setup IDs are used as training set and rest as test set.
NTU60-X[42] is a RGB derived skeleton dataset for the sequences of the original NTURGB+D dataset. The skeleton provided in this dataset is much denser and contains 67 joints. There are total of 56,148 action samples and the evaluation protocols are same as the NTURGB+D dataset.
NTU120-X[42] is the extension of NTU60-x dataset and corresponds to the action sequences provided by NTURGB+D 120 dataset. There are total of 113,821 samples in this dataset and the evaluation protocols are same as the NTURGB+D 120 dataset. Following [42], we evaluate our model on only Cross Subject protocol of NTU60-X and NTU120-X datasets.
SHREC[10] is a 3d skeleton based hand gesture recognition dataset. There are a total of 2800 samples with 1960 samples in train set and 840 samples in test set. Each sample has 20-50 frames and gestures are performed by 28 participants ones using only one finger and ones using the whole hand. There are 14 gestures and 28 gestures splits provided by the creators and we report results on both of these splits.

3.4.2 Implementation and Optimization details

As shown in Fig. 3.2 (right), the input skeleton to each of the part streams contains different number of joints. For NTURGB+D dataset, the body stream has input skeleton with a total of 25 joints, hands stream has the input skeleton with a total of 13 joints and legs stream with a total of 9 joints. Within the PSUMNet architecture, we use 10 STRBs for the body stream, 6 STRBs for the hands stream and 4 STRBs to process the legs stream.

We implement PSUMNet using the Pytorch deep learning framework. We use SGD optimizer with 0.1 as the base learning rate and a weight decay of 0.0004. All the models are trained on 4 1080Ti 12 GB GPU systems. For training of 25 joints datasets-NTU60 and NTU120, we use a batch size of 200. For 67 joints datasets-NTU60-X and NTU120-X, due to much denser skeleton, smaller batch size of 65 is used.

Type	Model	Params. (M) *	FLOPs (G) *	NTU60		NTU120	
				XSub	XView	Xsub	XSet
CNN	VA-Fusion [52]	24.6	-	89.4	95.0	-	-
based	TaCNN+ [48]	4.4	1.0	90.7	95.1	86.7	87.3
	ST-GCN [49]	3.1	16.3	81.5	88.3	70.7	73.2
	RA-GCN [39]	6.2	32.8	87.3	93.6	81.1	82.7
	2s-AGCN [37]	6.9	37.3	88.5	95.1	82.9	84.9
	PA-ResGCN[40]	3.6	18.5	90.9	96.0	87.3	88.3
GCN	DGNN [36]	26.2	-	89.9	95.0	-	-
based	MS-G3D[26]	6.4	48.8	91.5	96.2	86.9	88.4
	4s-ShiftGCN[8]	2.8	10.0	90.7	96.5	85.9	87.6
	DC-GCN+ADG [7]	4.9	25.7	90.8	96.6	86.5	88.1
	DualHead-Net [5]	12.0	-	92.0	96.6	88.2	89.3
	CTR-GCN [6]	5.6	7.6	92.4	**96.8**	88.9	90.6
Attention	DSTA-Net[38]	14.0	64.7	91.5	96.4	86.6	89.0
based	ST-TR [31]	12.1	259.4	89.9	96.1	82.7	84.7
	PSUMNet(Ours)	2.8	2.7	**92.9**	96.7	**89.4**	**90.6**

Table 3.2 Comparison with state of the art approaches for NTURGB+D and NTURGB+D 120 dataset. Model parameters are in millions ($\times 10^6$) and FLOPs are in billions ($\times 10^9$). *: These numbers are cumulative over all the streams used by respective models as per their training protocol.

3.4.3 Results

Tab. 3.2 compares the performance of proposed PSUMNet with other approaches on Cross Subject (XSub) and Cross View (XView) splits of NTURGB+D dataset [35] and Cross subject (XSub) and Cross Setup (Xset) splits of the NTURGB+D 120 dataset[25]. As can be seen from the Params. column in Tab. 3.2, PSUMNet uses the least number of parameters compared to other methods and achieves very comparable results across different splits of the datasets. For the hardest of the four splits - NTURGB+D 120 Cross Subject (XSub), PSUMNet achieves state of the art performance compared to other approaches which use 100%-400% more parameters. This shows the superiority of PSUMNet both in terms of performance and efficiency.

Model	Params. (M)	NTU60	NTU120
PA-ResGCN[40]	3.6	90.9	87.3
MS-G3D (Joint)[26]	3.2	89.4	84.5
1s-ShiftGCN (Joint)[8]	0.8	87.8	80.9
DSTA-Net (Bone)[38]	3.5	88.4	84.4
DualHead-Net (Bone)[5]	3.0	90.7	86.7
CTR-GCN (Bone)[6]	1.4	90.6	85.7
TaCNN+ (Joint)[48]	1.1	89.6	82.6
PSUMNet(Ours) (Body)	1.4	**91.9**	**88.1**

Table 3.3 Comparison of only body stream of PSUMNet with the best performing modality (i.e only joint, only bone) of state of the art approaches for NTURGB+D 60 and 120 dataset on Cross Subject protocol.

Model	Params. (M)	NTU60-X	NTU120-X
PA-ResGCN[40]	3.6	91.6	86.4
MS-G3D[26]	6.4	91.8	87.1
4s-ShiftGCN[8]	2.8	91.8	86.2
DSTA-Net[38]	14.0	93.5	87.8
CTR-GCN [6]	5.6	93.9	88.3
PSUMNet(Ours)	3.2	**94.7**	**89.1**

Table 3.4 Comparison with state of the art approaches for dense skeleton datasets NTU60-X and NTU120-X datasets.

We also compare the performance of only body stream of PSUMNet with single stream performance of other approaches in Tab. 3.3 for Xsub split of NTURGB+D and NTURGB+D 120 datasets. As can be seen, PSUMNet outperforms other approaches by a margin of 1-2% for NTURGB+D and by 2-3% for NTURGB+D 120 using almost the same or lesser number of parameters. This also supports our hypothesis that part stream based unified modality approach is much more efficient compared to conventional independent modality streams approach.

Trivedi et al.[42] introduced NTU60-X and NTU120-X, extensions of existing NTURGB+D and NTURGB+D 120 datasets with 67 joint dense skeletons containing fine-grained finger joints within the

Model	Params. (M)	14 gestures	28 gestures
Key-Frame CNN [10]	7.9	82.9	71.9
CNN+LSTM [29]	8.0	89.8	86.3
Parallel CNN[11]	13.8	91.3	84.4
STA-Res TCN [17]	6.0	93.6	90.7
DDNet [50]	1.8	94.6	91.9
DSTANet [38]	14.0	**97.0**	**93.9**
PSUMNet(Ours)	**0.9**	95.5	93.1

Table 3.5 Comparison with state of the art approaches for SHREC skeleton hand gesture recognition dataset.

full body pose tree. Handling such large number of joints while keeping the total parameters of the model in bounds is a difficult task. However, as shown in Tab. 3.4, PSUMNet achieves state of the art performance for both NTU60-X and NTU120-X datasets. Total parameters increase by a small margin for PSUMNet to handle the additional joints, yet it is worth noting that other competing approaches use 100%-400% more parameters as compared to PSUMNet. This shows the benefit of using part based streams approach for dense skeleton representation as well.

To further explore the generalization capability of our proposed method, we evaluate performance of PSUMNet for skeleton based hand gestures recognition dataset, SHREC[10]. Taking advantage of part based stream approach, we train only the hands stream of PSUMNet. As shown in Tab. 3.5, PSUMNet achieves comparable results to existing state of the art method (DSTANet[38]) which uses 1400% more parameters. PSUMNet outperforms the second highest approach (DDNet[50]) which uses 100% more parameters.

Overall, Tab. 3.2, 3.4 and 3.5 comprehensively show that proposed PSUMNet achieves state of the art performance, generalizes easily across a range of action datasets and uses a significantly smaller number of parameters compared to other methods.

3.4.4 Analysis

As explained in Sec. 3.3.1, we train PSUMNet using three part streams namely body, hands and legs streams and report the ensembled results from all the three streams. To understand the advantage of the proposed part stream approach, we compare stream wise per class accuracy for NTU120-X and NTU60-X of PSUMNet. Fig. 3.5 (left) depicts the per class comparison setting for per class accuracy comparison between the 'only hands stream' and 'only body stream' setting of PSUMNet for NTU120-X dataset. The classes shown correspond to those with largest (positive) gain in per class accuracy while

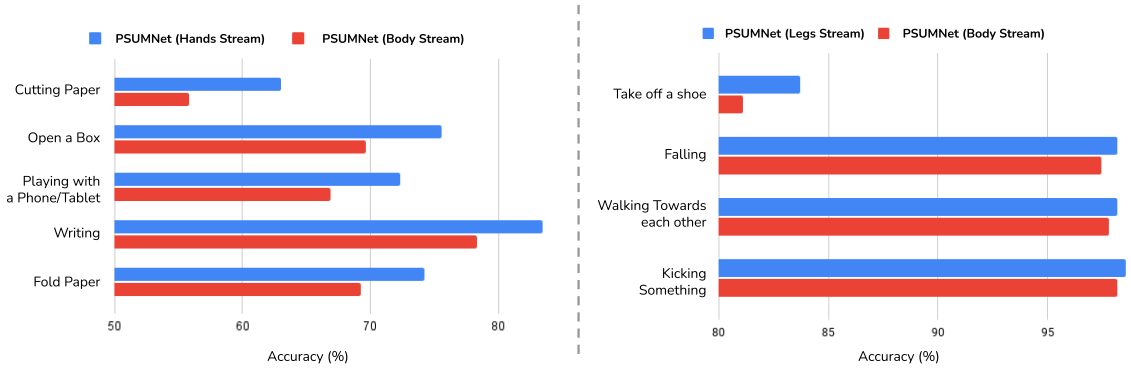


Figure 3.5 Comparing per class accuracy after training PSUMNet using only Hands stream and only body stream for NTU120-X dataset (Left) and only Legs stream with only body stream for NTU60-X datset (Right). On observing the class labels we can see that all the actions in the left plot are dominated by hand joints movements and all the actions in the right plot are dominated by leg joints movement and hence streams corresponding to these parts are able to classify these classes better which is in line with our hypothesis

using only hand stream. Upon observing the action labels of these classes, ("Cutting Paper", "Writing", "Folding Paper"), it is evident that these classes are dominated by hand joints movements and hence are better classified using only a subset of input skeleton which has dedicated representations for hand joints as opposed to using entire skeleton in a monolithic fashion.

Similarly, we also compare the per class accuracy while using only legs stream against only body stream of PSUMNet for NTU60-X dataset as shown in Fig. 3.5 (right). In this case too, the class labels with highest positive gain while using only legs stream are dominated correspond to expected classes such as "Walking", "Kicking".

The above results can also be appreciated better by studying the number of parameters in each of the part based stream. The body stream in PSUMNet has $1.4M$ parameters, Hands stream has $0.9M$ and legs stream has $0.5M$ parameters. Hence, hands stream while using only 65% of the total parameters of the body stream and legs stream while using only 35% of the body stream parameters can identify those classes better which are dominated by joints corresponding to each part stream.

Early action recognition: In the experiments so far, evaluation was conducted on the full action sequence. In other words, the predicted label is known only after all the frames are provided to the network. However, there can be scenarios we would like to know the predicted action label without waiting for the entire action to finish. To examine the performance in such scenarios, we create test sequences whose length is determined in terms of a fraction of the total sequence length. We study the trends in accuracy as the % of total sequence length is steadily increased. For comparison, we study

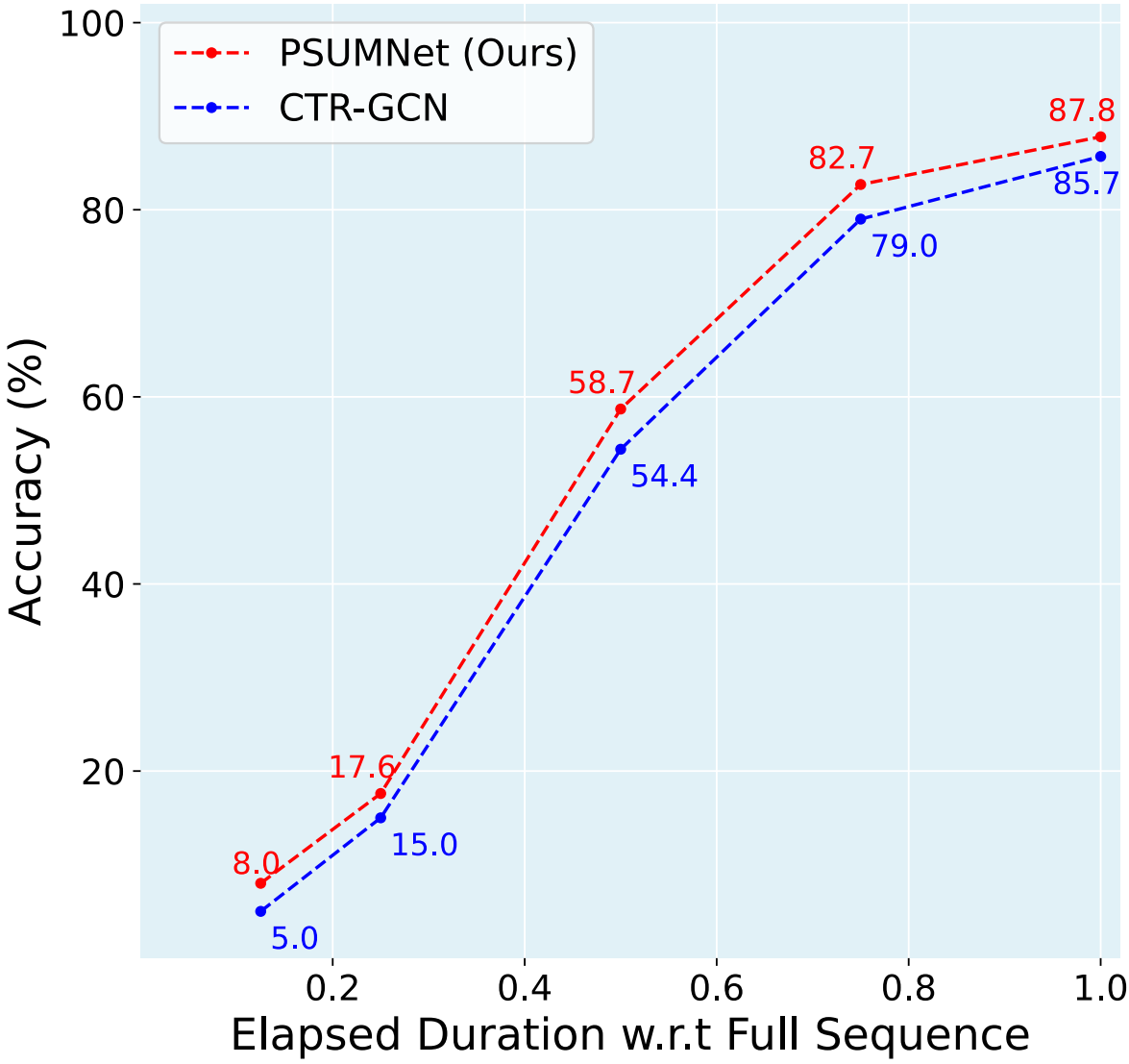
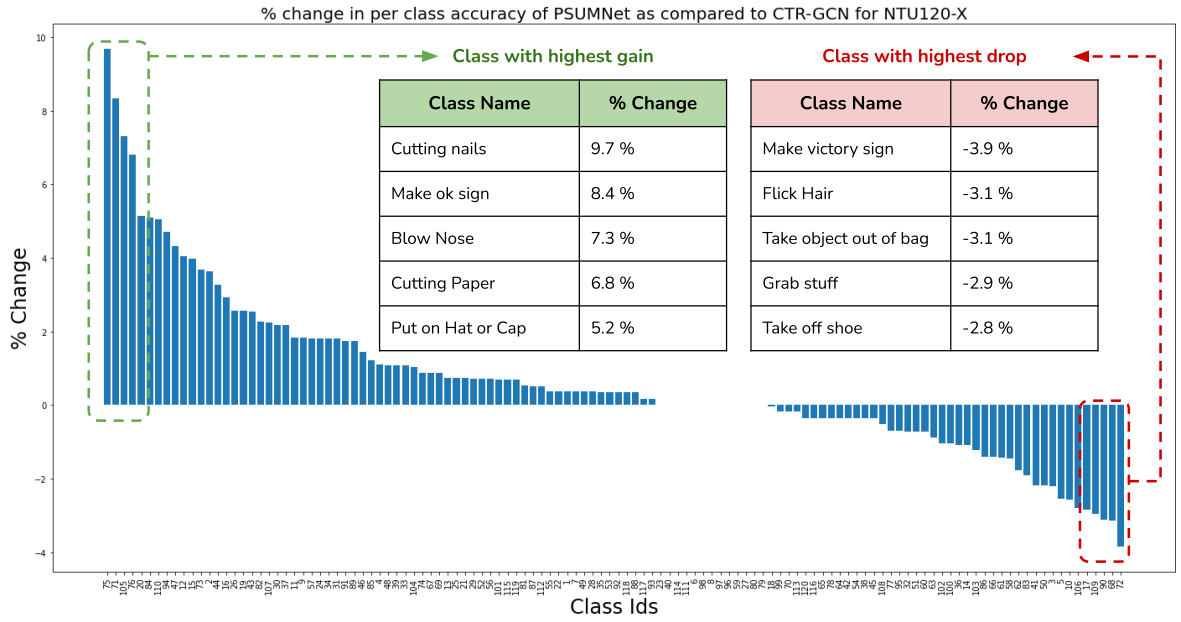


Figure 3.6 Comparing PSUMNet with current state of the art method, CTR-GCN on partially observed sequences for NTURGB+D 120 (XSub) dataset. Annotated numbers for each line plot denote accuracy of both models on partial sequences.

PSUMNet with the state of the art network, CTR-GCN [6]. As can be seen in Fig. 3.6, PSUMNet consistently outperforms CTR-GCN for partially observed sequences, indicating its suitability for early action recognition.

Per Class accuracy analysis: We analyze the per class accuracy change between our PSUMNet and existing state of the art CTR-GCN[6] for NTU120-X dataset in Fig. 3.7.



Class Name	% Change
Cutting nails	9.7 %
Make ok sign	8.4 %
Blow Nose	7.3 %
Cutting Paper	6.8 %
Put on Hat or Cap	5.2 %

Class Name	% Change
Make victory sign	-3.9 %
Flick Hair	-3.1 %
Take object out of bag	-3.1 %
Grab stuff	-2.9 %
Take off shoe	-2.8 %

Figure 3.7 Change in per class accuracy for PSUMNet against CTR-GCN [6] for NTU120-X dataset.

To further understand per class performance of PSUMNet, we provide 5 best performing classes and 5 worst performing classes of PSUMNet for NTU60, NTU120, NTU60-X and NTU120-X datasets in Tab. 3.6.

	NTU60	NTU120	NTU60-X	NTU120-X
Top 5	Take of Jacket(100%)	Jump up(100%)	Pick up(99.6%)	Arm circles(100%)
	Jump Up(100%)	Take off Jacket(99.6%)	Staggering(98.9%)	Arm swings(100%)
	Staggering(100%)	Walking(99.6%)	Jump Up(98.9%)	High Five(99.7%)
	Falling Down(99.6%)	Arm Swing(99.5%)	Kicking(98.9%)	Stand up(99.6%)
	Hugging(99.6%)	Cheers(99.3%)	Stand Up(98.9%)	Move Objects(99.5%)
Bottom 5	Writing(56.9%)	Staple book(48.7%)	Take of shoe(80.7%)	Staple Book(45.0%)
	Reading(73.5%)	Writing(62.9%)	Sneeze(83.3%)	Vicotry Sign(63.5%)
	Take off shoe(77.0%)	Victory Sign(63.5%)	Eat meal(84.0%)	Cutting Paper(63.9%)
	Eat meal(77.5%)	Counting Money(64.9%)	Put on shoe(84.9%)	Hit with Object(67.0%)
	Play with phone(77.5%)	Reading(66.5%)	Headache(86.2%)	Fold Paper(73.4%)

Table 3.6 Best and worst performing classes of PSUMNet for NTU60, NTU120, NTU60-X and NTU120-X dataset. The numbers in parenthesis indicate per class accuracy for that action label.

3.4.5 Ablations

Type	Streams	Params. (M)	NTU60	NTU120
Part Streams	Body	1.4	91.0	87.8
	Hands	0.9	89.6	85.8
	Legs	0.5	59.8	50.6
	Body + Hands	2.3	91.8	89.0
	Body + Legs	1.9	91.2	87.9
	Hands + Legs	1.4	90.4	86.5
Disjoint Parts	Body + Hands + Legs	2.8	89.6	86.1
Modalities in PSUMNet	Joint	2.8	90.3	86.1
	Bone	2.8	90.1	87.6
	Joint-Vel	2.8	88.5	82.7
	Bone-Vel	2.8	87.6	83.2
	Joint + Bone	2.8	91.4	88.6
FC Fusion	Body + Hands + Legs	2.8	90.9	86.3
PSUMNet		2.8	92.2	89.3

Table 3.7 Ablation experiments on NTURGB+D and NTURGB+D 120 Cross Subject dataset.

To understand the contribution of each part stream in PSUMNet, we provide individual stream wise performance of PSUMNet on NTU60 and NTU120 datasets Cross Subject splits as ablations in Tab. 3.7.

At a single stream level, the body stream achieves higher accuracy compared to hands and legs stream. This is expected since the body stream includes a coarse version of all the joints. However, as mentioned previously (Sec. 3.4.4), hands and legs streams classify actions dominated by respective joints better. Therefore, accuracies of Body+Hands (row 4 in Tab. 3.7) and Body+Legs (row 5) variants are higher than only the body stream. Legs stream achieves lower accuracy as compared to body and hands stream because there are only a small subset of action categories which are dominated by leg joints movements. However, as with hands stream, legs stream benefits classes which involve significant leg joints movements.

Our proposed part groups factorization registers each group's sub-skeleton in a global frame of reference (see Fig. 3.2). Further, all the part groups are not disjoint and have overlapping joints to better propagate global motion information through the network (Sec. 3.3.1). To justify our choice of globally registered part groups, we perform an ablation with a different part grouping strategy, with each part

group being disjoint and in a local frame of reference. Specifically, the ablation setup for body stream includes on 9 torso joints (including shoulders and hips joints), hands stream includes only 12 joints and legs stream includes only 8 joints. It is important to notice here that unlike our original strategy, both legs and hands in corresponding part stream are not connected. As expected, such strategy fails to capture global motion information unlike our proposed method (c.f. 'Disjoint parts' row and last row in Tab. 3.7).

To further investigate contribution of each data modality in our proposed unified modality method, we provide ablation studies with PSUMNet trained on single and two modalities instead of four (c.f. 'Modalities in PSUMNet' rows and last row in Tab. 3.7). We notice that PSUMNet benefits most from joint and bone modalities compared to velocity modalities. However, the best performance is obtained by utilizing all the modalities.

To justify the weighted late fusion of class scores predicted by each individual part streams (See Sec. 3.3.1), we also tried using a FC layer to combine final layer representation of each individual stream before computing the final accuracy. For this experiment we concatenated final layer features from each part stream and passed this fused representation through a FC layer followed by softmax to get the final class scores. However, as can be seen from penultimate row of Tab. 3.7, such fusion under performs the weighted late fusion of class scores as used in PSUMNet. One of the reason for lower performance for this ablation could be the weighing assigned by FC layers to the features of each part streams.

3.5 Conclusion

In this chapter, we present Part Streams Unified Modality Network PSUMNet to efficiently tackle the challenging task of scalable pose-based action recognition. PSUMNet uses part based streams and avoids treating the input skeleton in monolithic fashion as done by contemporary approaches. This choice enables richer and dedicated representations especially for actions dominated by a small subset of localized joints (hands, legs). The unified modality approach introduced in this work allows PSUMNet enables efficient utilization of the inter-modality correlations. Overall, the design choices provide two key benefits – (1) they help attain state of the art performance using significantly smaller number of parameters compared to existing methods (2) they allow PSUMNet to easily scale to both sparse and dense skeleton action datasets in distinct domains (full body, hands) while maintaining high performance. We expect PSUMNet to be an attractive choice for pose-based action recognition especially in real world deployment scenarios involving compute restricted embedded and edge devices.

In the next chapter we explore different challenges and shortcomings of existing skeleton action recognition paradigm specifically for in the wild and non-contextual skeleton action recognition.

Chapter 4

New Frontiers: In the wild and non-contextual pose based action recognition

4.1 Introduction

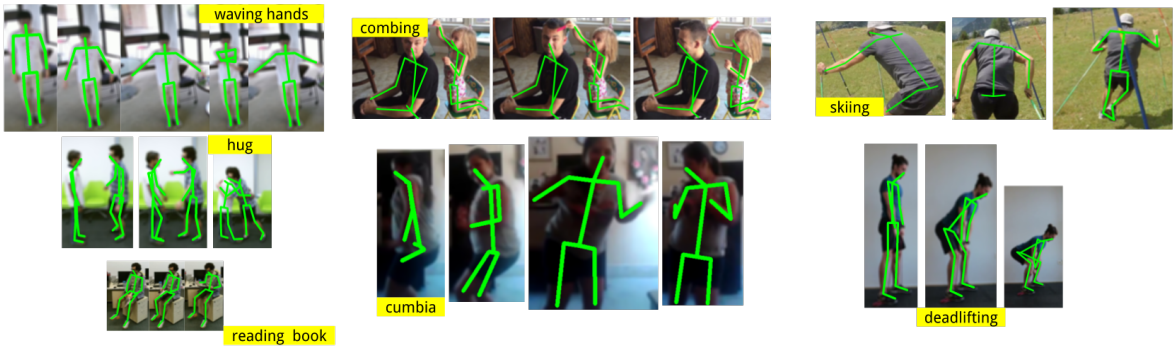


Figure 4.1 A pictorial illustration of the landscape for skeleton-based action recognition. Datasets such as NTU-120 characterize actions in controlled lab-like settings. We use state-of-the-art RGB 3-D pose estimation to obtain skeletons and benchmark recognition models 'in the wild' by introducing SKELETICS-152 dataset (Sec. 4.2). To explore out-of-context action recognition in the wild, we introduce SKELETON-MIMETICS (Sec. 4.3) and benchmark models trained on SKELETICS-152. Note that all datasets are skeleton-based – RGB background has been included to convey the original context.

The pose based skeleton datasets discussed thus far in our discussion, mainly NTURGB+d [35, 25] and NTU-X [42] involve the actions which are captured in indoor and controlled settings. Hence, all the samples have the similar environment leaving very little variations in terms of visibility of actors, number of actors in the frame and complex interactions between actors. Consequently it is worth exploring how well do the methods designed for such controlled setup perform when exposed to "in the

wild" datasets. To understand and analyze the performance and challenges that come along with in the wild pose based action recognition, we propose a new curated dataset dubbed Skeletics-152. (See Sec. 4.2). Skeletics-152 is a carefully curated and 3-D pose-annotated subset of videos sourced from Kinetics-700 [3], a large-scale RGB action dataset.

It is worth noting that for indoor datasets (NTURGB+d, NTU-X) and in the wild datasets (Kinetics), all the actions retain partial or full context provided by the objects and environment in the RGB sequences. We try to understand the performance of action recognition methods for non-contextual action sequences by introducing a non-contextual skeleton action dataset dubbed Skeleton-Mimetics (Sec. 4.3), the skeletal version of Mimetics [47], a subset of Kinetics-400 containing exaggerated, out-of-context human actions.

A brief overview of the action sequences with overlayed skeleton poses for samples from indoor dataset (NTU-120), in the wild dataset (Skeletics-152) and non-contextual action dataset (Skeleton-Mimetics) is shown in Fig. 4.1.

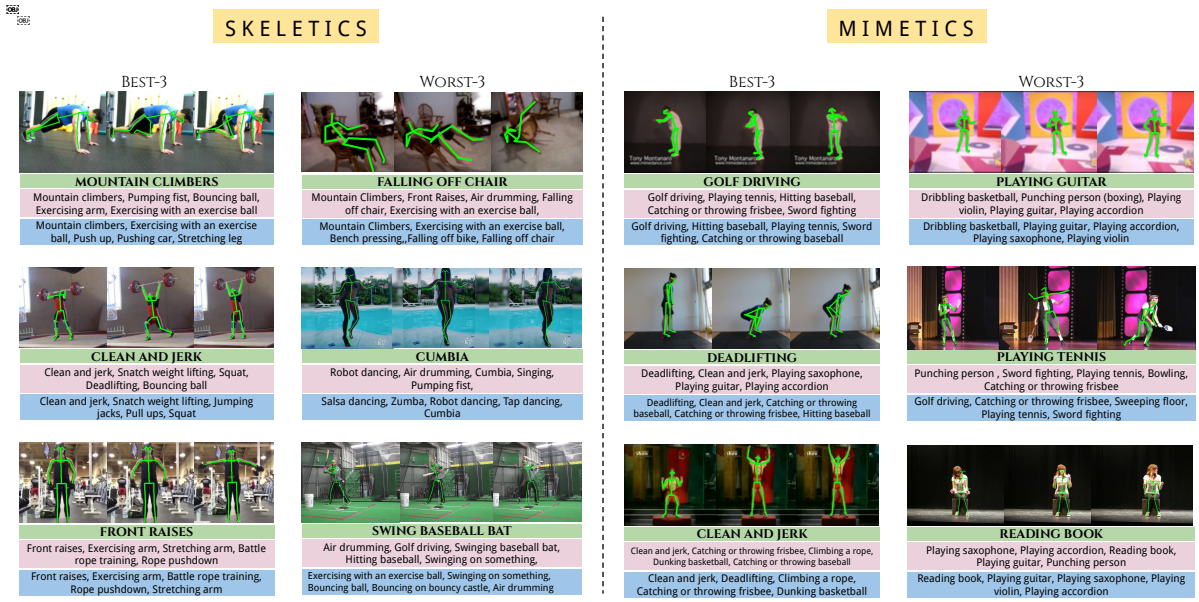


Figure 4.2 Sample skeleton sequences from Skeletics-152 and Mimetics-Skeleton. The sequences are chosen from best-3 and worst-3 classes in terms of performance achieved by best models on these datasets . The ground-truth phrase is color-coded green. The top-5 predictions by 4s-ShiftGCN are coded pink and those by MS-G3D are coded blue. Refer to Section 4.2 for details on the evaluation protocol and predictions.

4.2 Skeletics-152

Skeletics-152 is a "in the wild" skeleton action dataset, containing 125,657 skeleton sequences spanning over 152 action categories. Skeletics-152 dataset is created from a large in the wild video dataset Kinetics-700 [3]. Kinetics-700 is a large-scale video dataset consisting of over 650,000 YouTube video clips spanning over 700 action categories ranging from daily routine activities, sports and other fine-grained actions. Unlike the creators of a similar action dataset Skeleton-Kinetics 400 [49], we carefully curate and omit certain categories from Kinetics-700 dataset which are not relevant to pose based recognition. Following is the list of criteria we established for this curation process,

- A number of classes (e.g. 'Petting cat', 'Scrubbing face') were removed because most of the videos contain occluded poses which make the 3D pose estimation unviable.
- Some classes (e.g. 'Cooking eggs', 'Wrapping presents', 'Clay pottery making') were removed as they were captured from egocentric views.
- Some classes (e.g. 'Peeling apples', 'Peeling potatoes', 'Baking cookies') are highly object-centric and hence, irrelevant for skeleton based action recognition.
- Classes involving no substantial movement (e.g. 'Staring', 'Attending a conference') cannot be recognised solely based on human pose.
- Classes where the labels differ solely due to scene background were removed (e.g. 'Walking through snow' is same as 'Walking').

After removing aforementioned categories we extracted 3D pose sequences for the RGB sequences of the remaining 274 classes using VIBE [21] pose estimator. It is worth noting that VIBE provides real 3D skeleton joint estimation as opposed to pseudo 3D joint representation (2D + joint-level confidence score) provided in Skeleton-Kinetics-400 [49], which uses OpenPose [2] toolbox for pose estimation. The curated list of classes was further refined by removing classes involving a large group of people ("Playing ice hockey", "Playing american football",... etc) and classes for which VIBE reported significant missing joints. Ultimately, we present a highly reliable and focused in the wild skeleton action dataset containing 152 action categories. Few samples from this dataset along with their corresponding skeleton overlays are provided in Fig. 4.2 (Left).

4.2.1 Results

To understand the challenges involved with handling in the wild dataset, we benchmark existing action recognition approaches on Skeletics-152 dataset. We use the same train/test split as provided in Kinetics-700 dataset. We train two models, MS-G3D[26] and 4s-ShiftGCN[8], in two different experiment setups as under,

1. We first train the models on NTU-120 dataset and then fintune with Skeletics-152 before reporting the final accuracy. To maintain the skeleton topology between NTU-120 and Skeletics-152, we extract VIBE skeleton sequences for NTU-120 samples for the pretraining.

2. We train the models directly from scratch on Skeletics-152 and report the test set accuracy.

The results for these experiments are provided in Tab. 4.1.

Model	Accuracy	F-1 score
MS-G3D (trained from scratch)	56.39	50.80
4s-ShiftGCN (trained from scratch)	56.15	50.41
MS-G3D (pretrained on NTU-120 + finetuned)	55.75	49.57
4s-ShiftGCN (pretrained on NTU-120 + finetuned)	**57.01**	**51.13**

Table 4.1 Results on Skeletics-152 test set with mean accuracy as performance measure.

It is evident from Tab. 4.1 that skeleton action recognition for in the wild dataset is essentially much more challenging as compared to conventional indoor datasets. Inter/Intra class imbalance, poor joint estimation even by the best pose estimation method and a variety of noise inducing factors such as occlusion, uncontrolled lighting and environment setups, unnecessary interaction with objects, and many more make this problem much harder and open doors for thorough investigation and research in this domain of "in the wild" skeleton action recognition.

4.3 Skeleton-Mimetics

The action sequences contained in the datasets like NTURGB+d, NTU-X or Skeletics-152 are contextual where the environment and objects included in the RGB frames provide the necessary and relevant context for the action being performed. Mimetics dataset, consisting of 50 out-of-context action classes and derived from Kinetics-400 was introduced in [47] to investigate the out of context action recognition scenario. To mimic the similar scenario of out of context action recognition for skeleton sequences we introduce Skeleton-Mimetics.

Skeleton-Mimetics is derived from Skeletics-152 by selecting the action sequences which are out of context and performed by mimicry experts and mime artists without object interactions. We again use VIBE [21] to obtain the 3D pose estimation for these selected samples resulting in a dataset containing 319 out of context skeleton action sequences spanning over 23 classes. Few samples from this dataset along with their corresponding skeleton overlays are provided in Fig. 4.2 (Right).

4.3.1 Results

Following [47], we use Skeleton-mimetics only for the evaluation. Since Skeleton-mimetics is derived from Skeletics-152, we use the Skeletics-152 as our training set. We evaluate the performance of MS-G3D[26] and 4s-ShiftGCN[8], on Skeleton-mimetics for the following two experiments setup,

1. We train the models on entire Skeletics-152 dataset and evaluate on the Skeleton-Mimetics dataset. To commute the final accuracy we consider softmax score of the 23 classes corresponding to Skeleton-mimetics dataset.

2. We train the models using the samples of Skeletics-152 for only the 23 classes included in Skeleton-mimetics and then evaluate on the samples from Skeleton-mimetics dataset.

The results for these experiments are provided in Tab. 4.2.

Training set / Base model	Skeletics-152 (Complete)	Skeletics-152 (Skeleton-Mimetics classes)
MS-G3D	**57.37**	49.22
4s-ShiftGCN	56.11	51.10

Table 4.2 Performance summary in terms of mean accuracy for Skeleton Mimetics dataset as the test set. The 25 joints skeletons for both skeletics-152 (train set) and Skeleton-Mimetics (test set) are exctracted using VIBE.

It is evident from the results provided in Tab. 4.2 that training on entire Skeletics-152 dataset provided overall better evaluation performance for Skeleton-mimetics. This suggests that training on a larger dataset also helps models learn features which are much more generalizable to smaller evaluation datasets.

4.4 Conclusion

In this chapter, we describe different challenges that the skeleton action recognition domain faces and need further deeper understanding and exploration. To that extent, we examine two major challenging sub problems within the umbrella of skeleton action recognition, *First,* in the wild skeleton action recognition and *Second,* non-contextual skeleton action recognition.

We introduce, Skeletics-152 dataset, a curated and focused in the wild skeleton action recognition dataset and Skeleton-mimetics dataset, a non-contextual skeleton action recognition dataset. We provide salient features of these newly proposed datasets along with existing popular action recognition datasets in Tab. 4.3 for comparison. As can be seen from the highlighted rows of Tab. 4.3, Skeletics-152 and

Skeleton-Mimetics are one of the kind datasets in the domain of in the wild skeleton recognition and non-contextual skeleton action recognition respectively.

	No of. Classes	No. of. sequences	Action Vocabulary	Action setting	Action environment	Typical action duration	Level of action explicitness	Camera
HDM05 [28]	100	1500	Fixed	Lab/Prompted	Non-contextual	Not specified	High	Fixed
3D-Iconic dataset [33]	20	1739	Fixed	Lab/Prompted	Non-contextual	Not specified	High	Fixed
Florence-3D [34]	9	215	Fixed	Lab/Prompted	Non-contextual	Not specified	High	Fixed
NTU-60 [35]	60	56880	Fixed	Lab/Prompted	Non-contextual	1-10 seconds	High	Fixed
Large-RGB+D [51]	94	4953	Fixed	Lab/Prompted	Non-contextual	Not specified	High	Fixed/Moving
Kinetics-skeleton [49]	400	300,000	Fixed	Wild	Contextual	10 seconds	High	Fixed/Moving
NTU-120 [25]	120	114,480	Fixed	Lab/Prompted	Non-contextual	1-10 seconds	High	Fixed
Mimetics [46]	50	713	Fixed	Wild	Non-contextual	1-10 seconds	Moderate	Fixed
Skeletics-152	152	125,657	Fixed	Wild	Contextual	10 seconds	High	Fixed/Moving
Skeleton-Mimetics	23	319	Fixed	Wild	Non-contextual	1-10 seconds	Moderate	Fixed

Table 4.3 Attributes of different datasets in the skeleton based action recognition domain. Gray-shaded rows correspond to the new datasets introduced in this paper. Prompted means that subjects were instructed what action to perform. N.A means that there is no notion of classes. Not Specified means that the duration of an action is not specified in the respective paper.

We benchmark and analyze the performance of different skeleton action recognition approaches for these two proposed datasets. Our results and analysis provide insights into the shortcomings and challenges of the existing paradigm and approaches and encourage further investigation in these new frontiers.

Chapter 5

Conclusions

In this thesis, first we elaborate the significant shortcomings of conventional skeleton action datasets, NTURGB+D [35] and NTURGB+D 120 [25] in terms of noisy joints estimations and more importantly lack of fine finger and facial joints in the pose tree. We experimentally show that without fine finger level joints available in the input data, action categories like "Thumbs up", "Thumbs down", "Make ok sign", "Make victory sign" and many more can not be distinguished from one another. To overcome this major drawback, we propose extended versions of NTURGB+D and NTURGB+D 120, namely NTU60-X and NTU120-X, which includes a dense 118 joints pose tree for the action sequences as opposed to 25 joints pose tree provided in original Kinect based datasets. We use advanced RGB based 3d skeleton extraction methods like SMPL-X [30] and ExPose [9] to obtain fine finger joints (21 joints per hand) and facial joints (51 joints) along with standard 25 body joints. Qualitatively, the skeleton sequences provided in NTU-X are much more stable, noise free and comprehensive as compared to original NTURGB+D datasets. We also demonstrate quantitative superiority of proposed NTU-X dataset by benchmarking existing skeleton action recognition approaches on the proposed dataset. We observe that all the methods outperform their NTURGB+D counterparts when trained using proposed NTU-X datasets. Significant improvements in per class accuracy of classes involving subtle finger movements also justify the superiority of datasets like NTU-X and the shortcomings of conventional datasets like NTURGB+D. NTU-X opens new doors to explorations of RGB derived skeleton action datasets and need of finer and richer pose tree representation in these datasets. Even though NTU-X provides much better skeleton representation compared to its contemporary - kinect based NTU datasets, many skeletons included in NTU-X datasets still suffer from noisy joints estimates, especially at the fine finger level. Due to occlusion of certain limbs in the RGB frames and due to noise introduced by pose estimation architectures, many sequences in the NTU-X dataset are prone to noise and unreliable joint estimations. However, since the sequences in NTU-X dataset are captured from multiple camera views points, we lay future directions for works which could incorporate such multi-view data to better estimate the skeleton pose mainly for the limbs which are occluded in one of the view points.

Further, we propose an efficient and highly generalizable method, dubbed PSUMNet, to tackle the problem of both sparse and dense pose based action recognition. Instead of using conventional modality

based streams approach, PSUMNet combines all the modalities in a single pipeline and proposes novel part based streaming approach leading to extremely efficient model which achieves state of the art performance across various sparse, dense action recognition and also gestures datasets. We expect PSUMNet to be an attractive choice for pose-based action recognition especially in real world deployment scenarios involving compute restricted embedded and edge devices. Even though PSUMNet promises significant reduction in total trainable parameters while attaining very high accuracy, PSUMNet still requires three different part streams to be trained independently. As a future direction, if the part streams of PSUMNet could be combined into a single pipeline, allowing a one-go training of the entire architecture while maintaining the same performance - could be considered a significant contribution.

Lastly, we explore new frontiers within the paradigm of skeleton action recognition, specifically "in the wild" and "non-contextual" skeleton action recognition. We propose to new curated datasets, namely Skeletics-152 and Skeleton-Mimetics to explore the two mentioned frontiers respectively. We benchmark different approaches on these two datasets and provide insights of their performances. We discuss the difficulties raised by these new frontiers and lay the groundwork for the future works in these new and more challenging directions.

We provide an interactive dashboard to further investigate our findings and download the proposed datasets at https://skeleton.iiit.ac.in/. Along with access to newly proposed datasets (Skeletics-152 and Skeleton-Mimetics), we provide the code and pre-trained weights for the models used throughout the experiments mentioned in this thesis for the benefit of the community.

Related Publications

- **Neel Trivedi**, Anirudh Thatipelli, Ravi Kiran Sarvadevabhatla. *NTU-X: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions*. Twelfth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '21). Article 13, 1–9. https://doi.org/10.1145/3490035.3490270

- **Neel Trivedi**, Ravi Kiran Sarvadevabhatla. *PSUMNet: Unified Modality Part Streams are All You Need for Efficient Pose-based Action Recognition*. 1st International Workshop and Challenge on People Analysis: From Face, Body and Fashion to 3D Virtual Avatars (WCPA) at European conference on Computer Vision 2022.

- Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheswari, **Neel Trivedi**, Sourav Das, Ravi Kiran Sarvadevabhatla. *Quo Vadis, Skeleton Action Recognition?*. International Journal of Computer Vision 129, 2097–2112 (2021). https://doi.org/10.1007/s11263-021-01470-y

Bibliography

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 4

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. x, 4, 9, 10, 39

[3] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987, 2019. 38, 39

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 8

[5] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, and E. Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4334–4342, 2021. 29, 30

[6] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. xi, 21, 22, 26, 29, 30, 33, 34

[7] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, pages 536–553. Springer, 2020. 29

[8] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 10, 12, 13, 14, 15, 18, 21, 22, 29, 30, 39, 41

[9] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. x, 4, 9, 10, 43

[10] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat. 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, pages 33–38, 2017. xi, 5, 20, 24, 28, 31

[11] G. Devineau, W. Xi, F. Moutarde, and J. Yang. Convolutional neural networks for multivariate time series classification using both inter-and intra-channel parallel convolutions. In *Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP'2018)*, 2018. 31

[12] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 20, 21

[13] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 4

[14] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, May 2021. 8, 19

[15] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, 158:85–105, 2017. 20, 21

[16] A. Hernandez Ruiz, L. Porzi, S. Rota Bulò, and F. Moreno-Noguer. 3d cnns on distance matrices for human action recognition. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 1087–1095, New York, NY, USA, 2017. ACM. 20, 21

[17] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang. Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 31

[18] L. Huang, Y. Huang, W. Ouyang, and L. Wang. Part-level graph convolutional network for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11045–11052, Apr. 2020. 21

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 4

[20] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv preprint arXiv:1904.10681*, 2019. 8, 9

[21] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 39, 40

[22] C. G. Lab. Cmu graphics lab motion capture database converted to fbx. 1

[23] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 786–792. AAAI Press, 2018. 21

[24] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010. 2, 6, 9

[25] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. xi, 2, 5, 8, 9, 20, 21, 24, 28, 29, 37, 42, 43

[26] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. xi, 6, 10, 12, 13, 14, 15, 18, 20, 21, 22, 29, 30, 39, 41

[27] G. Moon, H. Kwon, K. M. Lee, and M. Cho. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2021. 19

[28] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007. 1, 42

[29] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018. 31

[30] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. x, 4, 9, 10, 43

[31] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 29

[32] H. Rahmani, A. Mahmood, Q. Du Huynh, and A. Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014. 2

[33] A. Sadeghipour and L.-P. Morency. 3d iconic gesture dataset, 2011. 42

[34] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–485, 2013. 42

[35] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. xi, 2, 5, 8, 9, 10, 20, 21, 24, 28, 29, 37, 42, 43

[36] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 26, 29

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 21, 22, 29

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *ACCV*, 2020. 6, 10, 12, 13, 14, 15, 16, 18, 21, 29, 30, 31

[39] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1915–1925, 2020. 21, 29

[40] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, pages 1625–1633, New York, NY, USA, 2020. Association for Computing Machinery. 6, 10, 12, 13, 14, 15, 18, 21, 22, 29, 30

[41] K. Thakkar and P. Narayanan. Part-based graph convolutional network for action recognition. *arXiv preprint arXiv:1809.04983*, 2018. 21, 22

[42] N. Trivedi, A. Thatipelli, and R. K. Sarvadevabhatla. NTU-X: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions. *arXiv preprint arXiv:2101.11529*, 2021. xi, 5, 20, 21, 24, 28, 30, 37

[43] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2012. 2

[44] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014. 2

[45] J. wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition, 2014. 8, 9

[46] P. Weinzaepfel and G. Rogez. Mimetics: Towards understanding human actions out of context. *arXiv*. 19, 42

[47] P. Weinzaepfel and G. Rogez. Mimetics: Towards understanding human actions out of context. *arXiv preprint arXiv:1912.07249*, 2019. 38, 40, 41

[48] K. Xu, F. Ye, Q. Zhong, and D. Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 2022. 22, 29, 30

[49] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6, 8, 20, 21, 29, 39, 42

[50] F. Yang, Y. Wu, S. Sakti, and S. Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM Multimedia Asia on ZZZ*, pages 1–6. 2019. 31

[51] J. Zhang, W. Li, P. Wang, P. Ogunbona, S. Liu, and C. Tang. A large scale rgb-d dataset for action recognition. In *International Workshop on Understanding Human Activities through 3D Sensors*, pages 101–114. Springer, 2016. 42

[52] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 20, 21, 29

[53] R. Zhao, K. Wang, H. Su, and Q. Ji. Bayesian graph convolution lstm for skeleton based action recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 21