NTU60-X: TOWARDS SKELETON-BASED RECOGNITION OF SUBTLE HUMAN ACTIONS

Anirudh Thatipelli, Neel Trivedi, Ravi Kiran Sarvadevabhatla*

Center for Visual Information Technology
IIIT Hyderabad, INDIA 500032
ravi.kiran@iiit.ac.in
https://github.com/skelemoa/ntu-x

ABSTRACT

The lack of fine-grained joints such as hand fingers is a fundamental performance bottleneck for state of the art skeleton action recognition models trained on the largest action recognition dataset, NTU-RGBD. To address this bottleneck, we introduce a new skeleton based human action dataset - NTU60-X. In addition to the 25 body joints for each skeleton as in NTU-RGBD, NTU60-X dataset includes finger and facial joints, enabling a richer skeleton representation. We appropriately modify the state of the art approaches to enable training using the introduced dataset. Our results demonstrate the effectiveness of NTU60-X in overcoming the aforementioned bottleneck and improve state of the art performance, overall and on hitherto worst performing action categories.

Index Terms— human action recognition, skeleton, dataset, human activity recognition, pose estimation

1. INTRODUCTION

Understanding human actions from visual data is crucial for applications related to surveillance, interactive user interfaces and multimedia systems. This task is usually accomplished with RGB videos as input [1]. However, advances in technologies have enabled use of other modalities (e.g. depth [2]) and systems such as Microsoft Kinect which can provide skeleton-like human pose representations [3]. In particular, the introduction of large-scale skeleton-based human action datasets NTU RGB+D [4, 5] has shifted the focus towards skeleton-based human action recognition approaches. In contrast to full-frame RGB-based representations, 3D skeleton joints encode human body dynamics in a computationally efficient manner, preserve privacy and can offer greater robustness to view and illumination.

The recent adoption of graph neural networks which process the skeleton action sequence as a spatio-temporal graph has enabled a steady rise in performance scores for skeleton action recognition [6, 7, 8, 9]. However, a closer examination of the scores reveals a performance bottleneck across the approaches - see Table 2 (top). The worst performing classes

_____

* Equal contribution
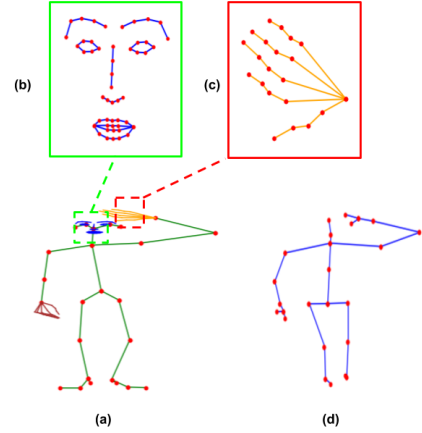
Fig. 1. (a) The 118 joint skeleton introduced in the new NTU60-X dataset. The 25 body joints are indicated by red dots.(b) 51 facial joints (c) 21 finger joints (d) 25 body joints present in NTU-60.

involve the usage of fingers. The reason is that hand joints in Kinect-based skeletons are represented by just two finger joints (Figure 1-d). As a result, actions involving subtle finger movements (e.g. 'eating', 'writing') often fail to be recognized correctly. Sometimes, even the non-hand, main body joints are often localized poorly by the Kinect-based capture system, as Figure 2 shows.

To address these issues, we introduce NTU60-X, a curated and extended version of the existing NTU dataset. Obtained from RGB videos, the skeletons 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, for a total of 118 joints per skeleton (see Figure 1). Additionally, we modify the state of the art approaches to enable experimental evaluation and benchmark the modified variants on NTU60-X. The results demonstrate the benefit of the newly introduced dataset for overcoming the performance bottleneck mentioned earlier and enabling recognition of subtle human actions involving hand-based joints.

For resources (pre-trained models, analysis and videos)
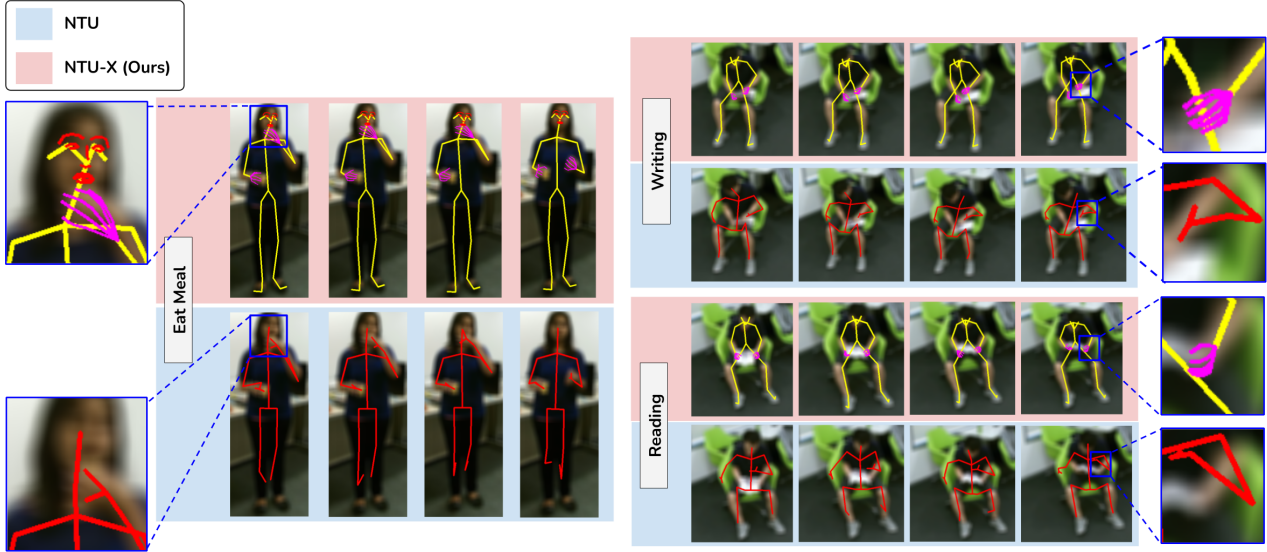
Fig. 2. Sample skeletons from original NTU60 Kinect dataset (blue background) and proposed NTU60-X dataset (pink background). Note that RGB frame is included only for reference and is not part of skeleton data. The three classes mentioned - 'eat meal', 'writing' and 'reading' are the most confused classes for NTU60 dataset (see Table 2). As the zoomed insets illustrate, the quality of joints captured by NTU60-X dataset is better compared to NTU60.

related to NTU60-X, please visit https://github.com/skelemoa/ntu-x.

2. RELATED WORK

Prior to creation of NTU60 dataset, a number of datasets enabled progress for skeleton-based human action recognition. MSR-Action3d [10] 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 dataset [11] scaled up the diversity to include videos from multiple views and with actions performed by 10 different actors. The NTU RGB-D dataset [4, 5] is the largest and widely used skeleton action dataset with a large diversity in terms of action classes, performers and capture viewpoints. Varying-view RGB-D Action Dataset (VAD) [12] 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 [6], the 2D skeleton joint coordinates predicted from RGB frames are combined with the joint estimation confidence to obtain a

Dataset	Body	Face	Fingers	Sequences	Classes	Joints
MSR-Action3D[10]	✓			567	20	20
Northwestern-UCLA[11]	✓			1,475	10	21
VAD[12]	✓			25,600	40	25
NTU RGB+D[4]	✓			56,880	60	25
NTU60-X (Ours)	✓	✓	✓	56,148	60	118

Table 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.

pseudo 3D skeleton representation on videos from Kinetics-400 action dataset [13]. However, the resulting skeleton dataset contains many invalid sequences [14].

We summarize the salient aspects of these datasets and our proposed NTU60-X in Table 1.

3. NTU60-X

The NTU dataset [4] 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 [15]. SMPL-X uses strong 2D pose priors estimated using Openpose [16] 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 [17]. ExPose uses a part-wise attention-based model that feeds high reso-

	PA-ResGCN	4s-ShiftGCN	MS-G3D
NTU60	Writing (63.97 %) Reading(67.65 %) Sneeze or cough (73.91 %) Type on keyboard (74.91 %) Eat meal (74.91 %)	Writing (65.19%) Reading (68.75%) Eat Meal (72.89%) Type on keyboard (73.82%) Sneeze or cough (73.91%)	Writing (57.41%) Eat meal (71.43%) Reading (72.43%) Sneeze or cough (77.17%) Play with phone or tablet (78.75%)
NTU60-X	Writing (78.89 %) Reading (94.12 %) Sneeze or cough (76.45 %) Type on keyboard (91.61%) Eat meal (80.95 %)	Writing (76.23%) Reading (91.91%) Eat Meal (80.22%) Type on Keyboard (91.24%) Sneeze or cough (75.72%)	Writing (72.96%) Eat meal (79.85 %) Reading (92.28 %) Sneeze or cough (80.80%) Play with phone or tablet (79.41 %)

Table 2. The columns correspond to the state of the art models on the NTU60 dataset. In the first row, the action classes with the lowest accuracy are shown. The second row is similar, except that the results correspond to our newly introduced dataset NTU60-X (Section 3), with models being modified appropriately (Section 4). Thanks to the availability of additional finger joints on hands in NTU60-X, performances for lowest accuracy classes improve by 10-20% on average.

lution 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.

Since it is difficult to automatically select between SMPL-X and ExPose pose representations, we employ a semi-automatic approach to curate the final dataset. We use Openpose [16] toolbox to estimate the 2D pose and associated confidence for the full-body joints. Openpose provides total 70 joints for face out of which we use 51 major joints as shown in Figure 1(b) to make the final skeleton of 118 joints (25 body + 21 $\times$ 2 fingers + 51 face) for each frame of the clips. Keeping the intra-view and intra-subject variance of the NTU dataset in mind, we sample random videos covering each view per class of NTU and estimate the SMPL-X and ExPose outputs. We examine the quality of the skeleton backprojected to RGB frame and use the accuracy of alignment to select between ExPose and SMPL-X. Empirically, we observe that ExPose and SMPL-X perform equally well for single-person actions but SMPL-X, though slow, provides better pose estimates for multi-person action class sequences.

To ensure good dataset quality, we remove corrupted videos from the original dataset, using a procedure similar to one adopted for the original dataset [4]. We also omit videos in which people are completely absent. Additionally, for some classes such as 'hugging', OpenPose provides poor estimates when one of the two people involved are in close proximity. We discard instances of such videos as well.

4. EXPERIMENTS

To evaluate the impact of NTU60-X on overall performance, we benchmarked models with state-of-the-art performance on

model	NTU60	NTU60-X
MS-G3D[7]	91.5	**91.76**
PA-ResGCN[9]	90.9	**91.64**
4s-ShiftGCN[8]	90.7	**91.78**

Table 3. Results of top performing models of NTU60 dataset on NTU60-X dataset (with finger joints) - see Section 4.1.

NTU60. For the models MS-G3D [7] and 4s-ShiftGCN [8], we updated the graph structure of the skeletons to incorporate the newly introduced joints.

PA-ResGCN [9], 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 were also changed based on the NTU60-X skeleton. In case of NTU60 skeleton, PA-ResGCN defines total 5 parts: torso, left arm, right arm, left leg and right leg. In the new NTU60-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. And for 118 joints (body + fingers + face) skeleton along with these 3 additional parts, one more part of face was added resulting in a total of 9 parts. The source code and pre-trained models are available in our github repository.

NTU60-X				
model	body	body + fingers	body + fingers + face
PA-ResGCN[9]	89.98	**91.64**	89.79
4s-ShiftGCN[8]	89.56	**91.78**	89.64
MS-G3D[7]	91.26	**91.76**	91.12

Table 4. Results on different variants of NTU60-X dataset to understand the contribution of the additional joints.

4.1. Results

The results of training the three models on the new NTU60-X dataset with finger joints included for Cross Subject protocol are shown in Table 3. From the results, it is evident that our modified MS-G3D, 4s-Shift-GCN and PA-ResGCN outperform their counterparts' performance on the original NTU dataset.

To further demonstrate the importance of including finger joints in the input skeleton, we analyze the worst performing classes for MS-G3D, PA-ResGCN and 4s-ShiftGCN on both original NTU60 and new NTU60-X dataset. As Table 2 shows, the performance for the worst classes (e.g. 'writing', 'reading', 'eat meal') , involving subtle finger-based movements, improves by 10-20% on average when trained on NTU60-X dataset.

4.2. 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 4. 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 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.

5. CONCLUSION

In this paper, 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 which 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 hitherto worst performing action categories. We also perform experiments to evaluate the relative importance of the introduced joints.

Going forward, we plan to expand our dataset to include the full complement of 120 action classes present in NTU-RGBD collection [5]. We also expect the research community to devise novel and efficient approaches for tackling dense skeleton representations present in our dataset.

For resources (pre-trained models, analysis and videos) related to NTU60-X, please visit https://github.com/skelemoa/ntu-x.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] Ronald Poppe, "Poppe, r.: A survey on vision-based human action recognition. image and vision computing 28(6), 976-990," *Image Vision Comput.*, vol. 28, pp. 976–990, 06 2010. 1

[2] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," June 2012, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). 1

[3] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu, "A survey on 3d skeleton-based action recognition using learning method," 2020. 1

[4] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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. 1, 2, 3

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 4

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018. 1, 2

[7] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 4

[8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 4

[9] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang 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)*, New York, NY, USA, 2020, pp. 1625–1633, Association for Computing Machinery. 1, 3, 4

[10] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14. 2

[11] Jiang wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning and recognition," 2014. 2

[12] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng, "A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition," *arXiv preprint arXiv:1904.10681*, 2019. 2

[13] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. 2

[14] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla, "Quo vadis, skeleton action recognition?," *arXiv preprint arXiv:2007.02072*, 2020. 2

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael 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. 2

[16] 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. 2, 3

[17] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black, "Monocular expressive body regression through body-driven attention," in *European Conference on Computer Vision (ECCV)*, 2020. 2