Quo Vadis, Skeleton Action Recognition?

Pranay Gupta · Anirudh Thatipelli · Aditya Aggarwal · Shubh Maheshwari · Neel Trivedi · Sourav Das · Ravi Kiran Sarvadevabhatla

Received: 16 September 2020 / Accepted: 8 April 2021 / Published online: 5 May 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
In this paper, we study current and upcoming frontiers across the landscape of skeleton-based human action recognition. To study skeleton-action recognition in the wild, we introduce Skeletics-152, a curated and 3-D pose-annotated subset of RGB videos sourced from Kinetics-700, a large-scale action dataset. We extend our study to include out-of-context actions by introducing Skeleton-Mimetics, a dataset derived from the recently introduced Mimetics dataset. We also introduce Metaphorics, a dataset with caption-style annotated YouTube videos of the popular social game Dumb Charades and interpretative dance performances. We benchmark state-of-the-art models on the NTU-120 dataset and provide multi-layered assessment of the results. The results from benchmarking the top performers of NTU-120 on the newly introduced datasets reveal the challenges and domain gap induced by actions in the wild. Overall, our work characterizes the strengths and limitations of existing approaches and datasets. Via the introduced datasets, our work enables new frontiers for human action recognition.

Keywords Human action recognition · Human activity recognition · Skeleton · 3-D human pose · Deep learning

1 Introduction
Understanding human actions, especially from their 2-D and 3-D joint-based skeleton representations, has received a lot of focus recently. Joint-based representations have a small memory footprint which improves feasibility of on-board processing in compute-restricted environments (e.g. smartphones, cameras on IoT devices). The privacy-friendly nature of the skeleton representation is also an advantageous factor.

On the flip side, obtaining accurate 3-D skeleton data usually requires specialized capture mechanisms and constraints on the capture environment. Even after the capture hurdle is crossed, the sparsity of skeleton representation relative to denser counterparts (RGB, depth) induces ambiguity and imposes additional challenges. In addition, the lack of large-scale, diverse datasets remained a challenge until the advent of datasets such as NTU-60 (Shahroudy et al. 2016) and PKU-MMD (Chunhui et al. 2017). These datasets have prompted a number of diverse approaches for skeleton-based action recognition (Shi et al. 2019; Wu et al. 2019; Peng et al. 2020; Zhang et al. 2019; Shi et al. 2019; Song et al. 2019; Li et al. 2019). The introduction of the even larger NTU-120 dataset (Liu et al. 2019) is poised to continue this trend.

The datasets and capture methods for aforementioned works are confined to controlled, indoor settings. Naturally, this prompts the question regarding the ability to recognize human activities occurring outdoors, ‘in the wild’? Also, in recent times, a number of works on robust estimation of human 3-D pose from RGB data have emerged (Kocabas et al. 2020; Kolotouros et al. 2019; Rogez et al. 2019). These prompt yet another question: How well can human actions be recognized in terms of 3-D skeletal pose estimated from RGB videos? To answer these questions, we introduce Skeletics-152 (Sect. 3.1), a carefully curated and 3-D pose-annotated subset of videos sourced from Kinetics-700 (Carreira et al. 2019), a large-scale RGB action dataset.

Actions in NTU-120 and Kinetics datasets retain either full or partial context supplied by object interactions and background. In contrast, out-of-context actions represent an unconventional and challenging frontier for skeleton action recognition. To benchmark performance for such actions, we introduce the skeletal version of Mimetics (Weinzaepfel and Rogez 2019), a subset of Kinetics-400 containing exaggerated, out-of-context human actions (Sect. 3.2). Additionally,
we introduce Metaphorics, a new video dataset with detailed action phrase annotations for videos of the popular social game Dumb Charades and expert dance performances of popular songs (Sect. 3.3).

Typically, the introduction of a newer, larger dataset (NTU-120) is marked by a flurry of novel architectures which aim to solve challenging domain tasks. In this paper, we argue that this is also a good opportunity to evaluate approaches originally trained for earlier dataset versions and more generally, re-evaluate the status quo. This argument has already been made successfully for RGB action recognition (Sigurdsson et al. 2017). To this end, we benchmark state-of-the-art approaches on the NTU-120 dataset and analyze the results (Sect. 4). Subsequently, we evaluate the performance of top ranked models on our newly introduced datasets—Skeletics-152 (Sect. 5.1), Skeleton-Mimetics (Sect. 5.2), Metaphorics (Sect. 5.3).

Overall, our work characterizes the strengths and limitations of existing approaches and datasets. It also provides an assessment of top-performing approaches across a spectrum of activity settings and via the introduced datasets, proposes new frontiers for human action recognition.

Our primary contributions can be summarized as follows:

- We introduce Skeletics-152, a curated 3-D pose annotated subset of Kinetics-700 for benchmarking skeleton action recognition ‘in the wild’ (Sect. 3.1).
- We introduce Skeleton-Mimetics to recognize skeleton-based out-of-context and exaggerated actions (Sect. 3.2).
- We introduce Metaphorics, a new video dataset with phrase annotations for YouTube videos of Dumb Charades and interpretative dance to explore the new frontier of metaphor-style actions (Sect. 3.3).
- We benchmark current, past state-of-the-art skeleton action recognition approaches on in-lab datasets (Sect. 4) and also on our newly introduced datasets containing actions happening in-the-wild (Sect. 5). We also summarize trends within and across these various datasets (Sect. 6).

To enable a rich, interactive exploration of our contributions mentioned above, we have made them available at https://skeleton.iit.ac.in/. The website features an interactive data analytics dashboard, code and pre-trained models for top-performing skeleton action recognition models and new skeleton action datasets (Skeletics-152, Skeleton-Mimetics, Metaphorics) introduced by us for additional exploration and benefit of the community.

## 2 Related Work

**Skeletal Datasets** Numerous 3-D skeletal datasets (Müller et al. 2007; Liu et al. 2019; Shahroudy et al. 2016; Chun-hui et al. 2017; Seidenari et al. 2013; Li et al. 2010) have been proposed over the last decade to further the advances in human action understanding. These datasets are focused on sequence based action detection and involve human subjects performing daily actions captured from multiple viewpoints. Sadeghipour and Morency (2011) introduce 3D Iconic gesture dataset where subjects outline the shape of virtual objects that is captured via Kinect v2 sensor. Recent work by Yan et al. (2018) introduces Skeleton-Kinetics using 2-D Open Pose on the large scale video dataset Kinetics-400. Weinzaepfel et al. take this idea further and introduce Mimetics (Weinzaepfel and Rogez 2019) containing a subset of Kinetics-400 with mimed actions.

**Skeleton Action Recognition** An earlier era of works serve to document handcrafted features for skeleton action recognition (Vemulapalli et al. 2014; Wang et al. 2012; Hussein et al. 2013; Halim et al. 2016). The recent class of approaches based on deep networks can be broadly categorized into three groups based on input skeleton data representation.

The first group explicitly consider the sequential nature of actions wherein the temporal dependencies are modelled using an RNN or an LSTM (zhang et al. 2017; Kim and Reiter 2017; Zhu et al. 2016). To further discriminate activities based on the joint dependencies, Song et al. (2016) introduce attention mechanisms at multiple levels in the network. Kundu et al. (2018) learn the action sequence as a trajectory in the pose manifold for the downstream activity classification task. Caetano et al. (2019) use CNN-based feature representation over a temporal window containing skeleton dynamics.

The second group of works model the input skeleton as a single spatio-temporal unit. In some instances, this unit is a tensor of the form frames × joints × coordinates which is subsequently processed by a CNN (Li et al. 2017; Du et al. 2015; Caetano et al. 2019; Li et al. 2018). More recently, a series of approaches use graph convolutions to model the (spatio-temporal) unit. Prominent examples include the STGCN framework introduced by Yan et al. (2018) and variants Wu et al. (2019); Song et al. (2019). In contrast to the fixed graph in ST-GCN, newer approaches involve adaptation to learn graph topology (Shi et al. 2019; Peng et al. 2020; Shi et al. 2019; Si et al. 2019; Tang et al. 2018).

In addition to the groups mentioned above, hybrid approaches also exist. Si et al. (2019) employ an attention-based graph convolutional LSTM to capture the spatio-temporal co-occurrence relationships. Zhang et al. (2019) propose a CNN-RNN late-fusion model with learnable view transformation. For a survey of 3-D skeleton action recognition, refer to Presti and La Cascia (2016) and Wang et al. (2018).

**Skeleton Action Recognition from RGB Video based Pose** In another class of approaches, human skeletal pose estimated from in-the-wild RGB video frames is used for action recognition. A number of approaches based on 2-D skeleton...
Fig. 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 (Sect. 3.1). To explore out-of-context action recognition in the wild, we introduce Skeleton-Mimetics (Sect. 3.2) and benchmark models trained on Skeletics-152. As a novel frontier for action recognition, we introduce Metaphorics (Sect. 3.2) which contains indirectly conveyed metaphor-style actions. Note that all datasets are skeleton-based – RGB background has been included to convey the original context.

3 Datasets

As mentioned in the Introduction (Sect. 1), large-scale datasets such as NTU-120 represent lab-style, controlled, indoor settings. In contrast, a much larger variety of human actions characterize in-the-wild RGB videos of human activities. To obtain 3-D skeleton representations from such videos, pose estimation techniques are applied on action sequences from large-scale activity datasets. In this section, we explore three diverse settings with progressively increasing level of complexity and abstractness in terms of human actions.

3.1 Skeletics-152

To source in the wild action videos, we use Kinetics-700 (Carreira et al. 2019) as the starting point. 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. However, unlike previously existing dataset with similar preparatory approach (Skeleton-Kinetics-400 Yan et al. 2018), we carefully omit categories from action settings which are incompatible for pose-based skeleton action recognition. Specifically,

- 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. ‘Standing’, ‘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 carefully removing such action categories, we use VIBE (Kocabas et al. 2020) on the remaining 274 categories to obtain the corresponding 3-D skeleton sequences. Within these sequences, we removed classes such as ‘Playing American football’, ‘Playing ice hockey’, ‘Doing aerobics’ containing large groups of people performing different activities within a single video. In addition, classes such as ‘Somersaulting’, ‘Springboard diving’ were removed since the VIBE model typically reported missing joints. Figure 2 shows some examples of omitted action classes.

For the case of multiple (> 2) skeleton detections in single video, we select the top two skeletons appearing in maximum number of frames. For intermediate frames with missing skeletons, we perform bounding box and joint interpolation. In the end, we obtain Skeletics-152, our curated 3-D skeleton dataset which contains 125,621 sequences spread over 152 classes. Refer to Fig. 3 (left) for some example sequences from Skeletics-152.
Fig. 2 Examples of classes from Kinetics-700 omitted for skeleton action recognition. In ‘Playing American football’, multiple people are detected. For ‘Playing ice hockey’ and ‘Somersaulting’, pose estimation is not accurate. In ‘Springboard diving’, the person performing the diving action is not tracked.

Fig. 3 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 (see Tables 5, 7). 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 Sect. 5 for details on the evaluation protocol and predictions.

It must be noted that Skeletics-152 is different from Skeleton-Kinetics-400 (Yan et al. 2018) which contains pseudo 3-D pose [2-D Pose + Joint-level confidence] obtained from the OpenPose (Cao et al. 2019) toolbox. Skeleton-Kinetics-400 indiscriminately includes all categories, without any curation. Unlike VIBE-based skeletons in our dataset, the pseudo 3D Skeleton-Kinetics-400 representations fail to capture the actual dynamics of 3D motion.

3.2 Skeleton-Mimetics

Human actions in the wild tend to be contextual. Context such as background, presence of certain objects can be very influential in certain traditional (RGB) action recognition approaches. This can lead to incorrect predictions when such actions are performed in out-of-context scenarios (Weinzaepfel and Rogez 2019). To this end, Mimetics dataset, consisting of 50 out-of-context action classes and derived from Kinetics-400 was introduced in Weinzaepfel and Rogez (2019). To explore skeleton action recognition for out-of-context, exaggerated action sequences, we introduce Skeleton-Mimetics. This dataset is derived from Skeletics-152, introduced previously. To create this new dataset, we shortlist classes with exaggerated movements and gestures. Instead of considering all the videos, we select specific action videos where action is performed by mimicry experts in out-of-context settings and without object interactions. The final dataset consists of 319 skeleton sequences across 23 classes.

The proposed dataset differs from the existing Mimetics dataset in terms of accurate 3-D poses. Since Skeleton-
Fig. 4 An illustration of the annotation for a typical Charades episode using the Anvil interface. ‘Action’, ‘Ground Truth’, ‘Success/Fail’, ‘Special Actions’ are the annotation channels. In the ‘Action’ channel, ‘rabb...(rabbit)’ and ‘zorro’ are guesses that the guessing player makes for the first two actions performed by the actor upon being revealed the ground truth phrase ‘the vampire diaries’. The segment labelled ‘vampire’ in the ‘Ground Truth’ channel is the entire duration for which the actor tried to act out the word ‘vampire’. The ‘Success/Fail’ channel shows the success and failure for corresponding guesses present in the ‘Action’ channel. Here, ‘rabbit’ and ‘zorro’ are both incorrect and hence they are marked as ‘F’. The ‘Special Action’ channel has tabs containing ‘TV’ and ‘wo...(number of words)’. These are helping actions to indicate that the phrase is the name of a TV show and the number of words in the phrase respectively.

Mimetics is derived from a curated set of videos specially designed for skeleton action recognition, it eliminates the factor of unusable 3-D poses for action classification. Further, since we wanted to compare the efficacy of Skeletics-152 as a dataset for pre training skeleton action recognition models, we kept joint positions same for both the datasets. Refer to Fig. 3 (right) for some example sequences from Skeleton-Mimetics.

3.3 Metaphorics

The datasets encountered so far can be characterized as verb-based actions, since the action class is fundamentally incomplete without the verb or the activity being performed. However, humans also tend to associate actions to non-verb words or objects, Iconic Gestures dataset (Sadeghipour and Morency 2011) being a known example. In general, actions can be more abstract and used to convey metaphorical concepts. One such scenario is the popular social game of Dumb Charades. The game involves interactive and adaptive guessing of a target ‘phrase’ (usually a movie title) based on actions being performed by an ‘actor’. Unlike other datasets, nouns and adjectives can have action depictions. Moreover, the vocabulary is open-ended, further compounding the action understanding challenge. To study actions arising in this challenging scenario, we introduce Metaphorics, an even complex dataset. The dataset contains videos from two scenarios - dumb charades and interpretative dance.

Dumb Charades We first source Dumb Charade game episodes from YouTube. In the game episodes, one person (‘actor’) acts out a target phrase word by word while the other player tries to guess the target phrase solely from actions...
Fig. 5 Sample skeleton sequences from our Metaphorics dataset. The ground-truth phrase is color-coded green. The top-5 predictions by 4s-ShiftGCN are coded pink and those by MS-G3D are color-coded blue. Refer to Sect. 5.3 for details on the evaluation protocol and predictions performed by the ‘actor’. For annotation, we use the popular Anvil tool (Kipp 2001). We annotate (i) target phrase (ii) beginning and ending timestamps for each action segment (iii) guess phrase associated with a segments (iv) episode outcome (‘correctly guessed’, ‘incorrect’). We also annotate certain special actions such as number of words and the current word number for a multi-word target phrase. These special actions also include helping actions which the actor uses to convey some basic information to the guessers such as length of the word (‘long’, ‘short’). Figure 4 provides an illustration of a typical annotation for a Charades video episode.

To characterize performance of action recognition approaches, we associate each word with its corresponding temporal video segment. After removing instances where the actor is occluded, we obtain 716 segments across 28 game sessions.

Interpretative Dance We also source YouTube videos containing interpretative dances of popular songs. In these videos, the song is made audible only to the actor who then proceeds to enact real-time actions corresponding to song lyrics. In this case, the guesser needs to correctly guess the song title based on the performed lyric-based actions. Unlike Charades, the actor is required to act out the song lyrics in real time which increases the challenge since the actions are faster paced than Charades.
As part of the annotation process, we align the lyric subtitle file of the original song and the video based on the starting point of the song (in the video). Since the actions are performed in real time, we obtain the action level annotations by aligning the audio file of the video with the original audio file of the song. The temporal extents of the dance are thus annotated into word-level action video segments. We obtain a total of 129 video segments across two full-length music videos.

In total, our Metaphorics dataset contains 845 video clips. As with other RGB datasets, we obtain corresponding 3-D skeleton sequences using VIBE (Kocabas et al. 2020) — see Fig. 5 for examples. The proposed Metaphorics dataset is very diverse in terms of the action sequences and the labels due to its open-ended vocabulary (see Fig. 6). Compared to existing datasets, videos tend to be ‘bursty’ due to the extremely small temporal extents of the actions. The 3D-Iconic dataset (Sadeghipour and Morency 2011) is similar to Metaphorics in the sense that it contains object based gesture actions. However, subjects are prompted to explicitly outline the shape of the object categories unlike the unprompted actions seen in Metaphorics.

To gain an overall perspective about the datasets, both existing ones and those introduced in our work, we summarize their prominent attributes in Table 1.

### 4 Skeleton Action Recognition in the Lab

In this section, we look at the NTU-120 (Liu et al. 2019) dataset which is currently the largest lab styled 3-D skeleton action recognition dataset. It comprises 114,480 25-joint 3-D skeleton annotated videos of 120 human actions, performed by 106 subjects in a controlled indoor setting and captured from 32 different setups.
Table 2 Benchmarking comparison for NTU-120 test set (mean accuracy)

| Method                        | Cross setup | Cross subject |
|-------------------------------|-------------|---------------|
| DGNN Shi et al. (2019)        | 78.13       | 75.16         |
| GCN-NAS Peng et al. (2020)   | 85.29       | 81.99         |
| 2s-SDGCN Wu et al. (2019)    | 86.18       | 84.42         |
| VA-CNN (ResNeXt-101) Zhang et al. (2019)† | 86.90 | 84.88 |
| 4s-ShiftGCN Cheng et al. (2020) | **87.65** | 85.76 |
| MS-G3D Liu et al. (2020)     | 87.32       | **85.92**     |
| best-5 models average pooled | **88.80**   | **87.22**     |

Bold numbers represent the previous best and our best performance scores. Italic lines correspond to models which originally reported results on NTU-60 but retrained by us on NTU-120 for comparison. The last row of the table corresponds to the new state-of-the-art which is average pooled ensemble of best-5 models above.

†Our version of VA-CNN Zhang et al. (2017) with ResNeXt-101 backbone

Fig. 7 Class accuracy plots for NTU-120 with standard deviation

4.1 Evaluation Protocol

Two standard evaluation protocols are typically used for evaluation of multi-subject multi-viewpoint skeleton action recognition approaches. In the Cross Subject protocol, the train and test set are split based on performer id. Under the protocol proposed by Liu et al. (2019) for NTU-120, 53 subject ids out of 106 are allocated for training and the remaining for test. We use data from 11 (20%) randomly selected ids of original training set for validation.

The other protocol is Cross Setup. By default, action sequences from the 16 even-numbered camera setup ids are used for training and 16 odd setup ids are used for testing. As with cross subject protocol, we retain the original NTU120 test set and use 4 (25%) ids randomly chosen from even setup ids for validation.

4.2 Performance with Full Sequences

For benchmarking, we selected approaches which report performance on NTU-120 and top 5 approaches with the best performance on NTU-60 (Shahroudy et al. 2016), the precursor to NTU-120. The results on the test set of NTU-120 can be viewed in Table 2. The results show that 4s-Shift-GCN (Cheng et al. 2020) and MS-G3D (Liu et al. 2020) are the best performers for Cross Setup and Cross Subject respectively.

The italicized portion of Table 2 shows the performance of top performing NTU-60 models evaluated on the NTU-120 test set. Note that these models were not originally designed for NTU-120 and were retrained by us from scratch, for benchmarking purposes. From the results, we notice that our version of VA-NN (Zhang et al. 2019), retrained with a more powerful backbone (ResNeXt-101) performs competitively with state-of-the-art NTU-120 approaches (MS-G3D and 4s-Shift-GCN). VA-CNN has a relatively simpler architecture and is faster to train, further adding to its appeal. More significantly, the results underscore the importance of benchmarking existing approaches on newly introduced datasets while investing effort into creation of novel architectures.

Figure 7 shows the mean accuracy and associated standard deviation for the top-5 models. The significant magnitude of deviation indicates that additional progress is needed before mean accuracy can be considered a reliable measure of overall performance (Fig. 8).

To obtain a better understanding of performance, we list the 10 best recognized and worst recognized action classes in Tables 3 (cross subject) and 4 (cross setup). The results show that the best and worst performers largely stay same across all the models. The best performing classes (e.g. ‘Arm swings’, ‘Jump up’, ‘Walking towards’) have distinct actions...
The NTU-120 confusion matrix for MS-G3D model sorted by class-wise accuracy shows that the least accurately recognized classes are confused amongst each other (magnified inset) involving large joint-level movements. On the other hand, action classes containing subtle actions with fine-grained differences are the hardest to recognize and exhibit large intra-class confusion (Fig. 9). For instance, ‘Make ok sign’ and ‘Make victory sign’ get confused with each other significantly because the actions differ only in terms of hand joint movement which is not captured in adequate detail by the Kinect sensor. In addition, skeletons for classes such as ‘Reading’, ‘Writing’ have very low inter-class variability, resulting in poor performance.

The top 5 models exhibit similarities at the set level for best-10 and worst-10 classes. However, motivated by the variance in actual rank order (Tables 3, 4), we examine performance with an average pooled ensemble of top-5 models. The noticeably improved ensemble performance (bottom row of Table 2) suggests that the top-5 models span action classes in a complementary manner.

4.3 Performance on Partial Sequences

Action recognition from partially observed sequences has been an active area of research (Ryoo 2011; Cao et al. 2013; Ma et al. 2016; Cai et al. 2019) and has many practical applications in the field of video surveillance and human-computer interaction. The ambiguity induced by partial sequences naturally makes this a challenging problem. To study action recognition in this setting, we benchmark the top-3 models of Table 2 on partially observed skeleton sequences of NTU-120 using the Cross Subject protocol. The increase in accuracy is on expected lines, i.e. actions are generally better recognized when the extent to which they are accessible increases (see Fig. 8). Note that 4s-ShiftGCN and MS-G3D outperform the third best performer, VA-CNN, by a noticeable margin. This is likely due to the complementary features which are learnt via the multiple feature stream processing present in 4s-ShiftGCN and MS-G3D models.

To explore this at class level, we replicate the plot of Fig. 8, but now for the top-1 model (MS-G3D) and for individual action categories. The curves for best-5 and worst-5 classes sorted by overall accuracy can be viewed in Fig. 10. The closer an activity’s curve is to the top-left corner, the better is its ability to be recognized early. It is evident from the plot that most of the best-5 classes are not confidently recognizable until 50% of the action sequence is completed. From this viewpoint, it is also interesting to note that some activities (best-3, 4) ranked lower than the best-1 but have an earlier onset of recognition. The temporal effects of intra-class confusion on worst ranked classes can also be observed in this plot.

To understand recognition onset trends in a more fine-grained manner, we propose Area-under-curve (AUC) as a better alternative, i.e. the normalized area under the curve with % of elapsed sequence on x-axis and % of correctly recognized sequences on the y-axis. The closer a category’s
### Table 3 Best-10 and Worst-10 classes for models trained on NTU-120 (Cross Subject)

|     | MS-G3D       | 4s-ShiftGCN | VA-CNN       | 2s-SDGCN     | 2s-AGCN       |
|-----|--------------|-------------|--------------|--------------|--------------|
| **Best-10** |              |             |              |              |              |
|     | Hugging      | Jump up     | Falling down | Stand up     | Hugging      |
|     | Hopping      | Staggering  | Walking towards | Jump up     | Drink after cheers |
|     | Put on jacket | Take off jacket | Jump up     | Walking towards | Arm swings   |
|     | Walking Towards | Arm swings  | Arm swings   | Drink after cheers | Put on jacket |
|     | Drink after cheers | Put on jacket | Pushing     | Arm circles   | Run on the spot |
|     | Jump up      | Walking towards | Staggering  | Put on Jacket | Arm circles |
|     | Staggering   | Arm circles | Arm circles | Arm swings    | Jump up      |
|     | Arm circles  | Hugging     | Squat down  | Staggering   | Staggering  |
|     | Arm swings   | Run on the spot | Drink after cheers | Take off jacket | Walking towards |
|     | Capitulate   | Drink after cheers | Follow       | Hugging      | Follow       |

|     |              |             |              |              |              |
| **Worst-10** | Staple book  | Make ok sign | Staple book  | Staple book  | Staple book  |
|     | Make victory sign | Make victory sign | Make victory sign | Make ok sign | Make ok sign |
|     | Hit with object | Make ok sign | Staple book  | Writing      | Make victory sign |
|     | Blow nose    | Counting money | Counting money | Counting money | Counting money |
|     | Counting money | Blow nose    | Play with phone or tablet | Make ok sign | Writing      |
|     | Writing      | Reading      | Reading      | Cutting nails | Cutting paper |
|     | Reading      | Cutting paper | Writing      | Blow nose    | Hit with object |
|     | Make ok sign | Hit with object | Hit with object | Play with phone or tablet | Blow nose |
|     | Snap fingers | Cutting nails | Fold paper   | Wield Knife  | Cutting nails |
|     | Cutting nails | Play with phone or tablet | Play magic cube | Hit with object | Yawn |

### Table 4 Best-10 and Worst-10 classes for the models trained on NTU120 (Cross Setup)

|     | 4s-ShiftGCN | MS-G3D       | VA-CNN       | 2s-SDGCN     | 2s-AGCN       |
|-----|-------------|--------------|--------------|--------------|--------------|
| **Best-10** |              |             |              |              |              |
|     | Put on jacket | Stand up     | Walking towards | Walking towards | Walking towards |
|     | Walking towards | Nod head or bow | Falling down | Put on jacket | Put on jacket |
|     | Staggering   | Put on jacket | Jump up      | Stand up     | Stand up     |
|     | Jump up      | Walking towards | Staggering  | Nod head or bow | Hopping   |
|     | Stand up     | Arm circles  | Hopping      | Walking apart | Arm circles   |
|     | Walking apart | Hopping     | Stand up     | Hopping      | Staggering   |
|     | Take off jacket | Staggering  | Arm swings   | Arm circles   | Nod head or bow |
|     | Hopping      | Arm swings   | Walking apart | Jump up      | Cheer up      |
|     | Nod head or bow | High five   | Cross toe touch | Take off jacket | Drink after cheers |
|     | Cross toe touch | Walking apart | Nod head or bow | Sit down     | Walking apart |

|     |              |             |              |              |              |
| **Worst-10** | Staple book  | Writing     | Make ok sign | Staple book  | Cutting paper |
|     | Writing      | Staple book  | Staple book  | Writing      | Play magic cube |
|     | Make ok sign | Cutting paper | Writing      | Cutting paper | Make ok sign |
|     | Cutting paper | Make ok sign | Counting money | Yawn       | Staple book   |
|     | Yawn         | Counting money | Reading     | Counting money | Make victory sign |
|     | Make victory sign | Cutting nails | Yawn       | Cutting nails | Counting money |
|     | Wield knife  | Yawn         | Make victory sign | Make ok sign | Writing      |
|     | Counting money | Wield knife | Cutting paper | Make victory sign | Yawn       |
|     | Reading      | Reading      | Play with phone or tablet | Reading | Type on keyboard |
|     | Cutting nails | Make victory sign | Hit with object | Play magic cube | Cutting on nails |
AUC is to 1, the earlier it can be recognized. Refer to Fig. 11 where best-5 and worst-5 activities by AUC can be seen. Multiple interesting trends can be observed. Firstly, most of the best-5 and worst-5 classes are different from overall accuracy plot counterparts from Fig. 10. The AUC-wise top performing activities (e.g. ‘Shake head’, ‘Walking’, and ‘Take off jacket’) contain unique action sequences from the beginning and therefore, are more consistently identified with increasing number of frames. Also classes like ‘Walking Towards’ and ‘Walking Apart’ which show very minimal intra-class variation and differ only in terms of increasing distance between the subjects are easily differentiable with the increasing frames. The accuracy of ‘Point finger’ decreases in the first half due to barely discernible joint movement in the initial frames. Among the AUC-wise worst classes (‘Snap fingers’, ‘Make victory sign’, ‘Make OK sign’), the finger-joint motion is predominant. Since hand joints are not captured via Kinect v2 sensor in NTU-120 dataset, a fundamental bottleneck arises in recognizing these activities regardless of the elapsed time.

### 5.1 Skeletics-152

The train-test split in Skeletics-152 is created following the splits originally provided with Kinetics-700. We use the original validation set of Kinetics-700 as our test set. We randomly split the original Kinect-700 training set into training and validation sets in a 85:15 ratio. To address class imbalance, we employ class-frequency based mini-batch resampling and class-based loss weighting.

We evaluated the best two performers (MS-G3D and 4s-ShiftGCN) from NTU-120 in two training regimes. In one regime, we first extracted VIBE skeletons from RGB videos of NTU-120 and trained the models on these skeletons. This allowed us to eliminate the effect of different joint positions in NTU-120 dataset and VIBE pose estimation. The resulting models were ultimately fine-tuned on the Skeletics-152 data. In the second regime, we trained the models from scratch on Skeletics-152 data. We found that 4s-ShiftGCN provides the best performance (Table 5). Pre-training on NTU-120 provides a slight benefit compared to training from scratch. Comparing the performance in Tables 2 and 5, it is evident that skeleton-based action recognition in the wild is significantly more challenging given the inter/intra-category diversity and noise-inducing factors (e.g. occlusion, lighting, uncontrolled background context, interaction with unnecessary objects). In addition, we empirically observed that even the best 3-D pose estimators routinely generate poorly localized joint estimates, impacting performance.

As shown in the Table 6, the best-5 classes are all exercise based activities which tend to have very low intra class variability. On the other hand, sequences from the worst-5 classes exhibit a lot of diversity and intra-class variability (see Fig. 3).

### 5.2 Skeleton-Mimetics

Following the procedure of Weinaezpfel and Rogez (2019), we use our proposed Skeleton-Mimetics only for evaluation. Similar to Skeletics-152, we use 4s-ShiftGCN and MS-G3D as the base models. Due to the similarity of challenges faced in recording out-of-context skeleton actions and actions in the wild, we choose Skeletics-152 as the training dataset. The final prediction is obtained by considering the maximum softmax score among the 23 skeleton-mimetics classes out of the 152 in Skeletics. We perform another experiment where instead of training on samples from all the 152 classes of
Table 5 | Results on Skeletics-152 test set with mean accuracy as performance measure

| 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** |

Bold represents the best performance score

Table 6 | Best-5 and Worst-5 classes of all models trained on Skeletics-152 dataset

| Model | Best-5 | Worst-5 |
|-------|--------|--------|
| MS-G3D | Mountain climber (exercise) | High fiving |
| 4s-ShiftGCN | Clean and jerk | Falling off chair |
| | Front raises | Cumbia |
| | Jumping Jacks | Falling off chair |
| | Deadlifting | Hugging (not baby) |
| | Lunge | Combing hair |

| Model | Best-5 | Worst-5 |
|-------|--------|--------|
| MS-G3D | Mountain climber (exercise) | High fiving |
| 4s-ShiftGCN | Clean and jerk | Falling off chair |
| | Front raises | Cumbia |
| | Jumping Jacks | Falling off chair |
| | Deadlifting | Hugging (not baby) |
| | Lunge | Combing hair |

Table 7 | 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 extracted using VIBE

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

Bold value represents the best performance score
Refer to Sect. 5.2 for details

Skeletics-152, samples pertaining only to the 23 skeleton mimetics classes are used.

The results shown in Table 7 depict that for both the base models, training on the complete Skeletics-152 dataset, improves performance as compared to training on the smaller split containing the 23 skeleton-mimetics classes. The results suggest that training on a larger dataset (classes outside the test set) helps the models learn better generalisable features which boost the test performance, whereas a smaller dataset (classes specific to the test set) tends to learn features which offers limited generalisability (Table 8).

5.3 Metaphorics

Due to the small number of available action segments, we use the Metaphorics dataset only for evaluation. To begin with, we use the previously determined state-of-the-art skeleton action recognition models—MS-G3D and 4s-ShiftGCN to obtain the class predictions for the action segments. Since the label sets used to train the models and the ones from Metaphorics dataset are different, direct comparison is not possible. Therefore, to perform quantitative evaluation, we propose to compare lexical representations of the predicted label and the ground truth label from the Metaphorics dataset. To obtain the lexical representations, we average the word2vec embedding vectors of words that constitute the action description. Subsequently, we compute the cosine distance between these representations. These distances are averaged across the dataset. Table 9 reports the mean performance for two different sets of models—one set trained on NTU-120 and another set trained on Skeletics-152. The rather poor performance can be attributed to the fundamentally different nature of training and evaluation settings (see Table 1). In particular, the abstract, non-contextual, gesture-style actions in Metaphorics are qualitatively distinct from the explicit and contextual actions present in training datasets (NTU-120, Skeletics-152).

To obtain a qualitative perspective, we report top-5 predictions by MS-G3D, 4s-ShiftGCN pre-trained on Skeletics-152 for sample action segments in Fig. 5. As mentioned before,
Table 8  List of Best-5 and Worst-5 classes in terms of accuracy for NTU-60, NTU-120, Skeletics-152 and Skeleton-Mimetics datasets. The model associated with the best performance is in brackets alongside the dataset name

| Dataset          | Best-5 Classes                                      | Worst-5 Classes                                       |
|------------------|-----------------------------------------------------|------------------------------------------------------|
| NTU-60 [MS-G3D]  | Staggering (99.64%)                                 | Writing (57.41%)                                      |
|                  | Jump up (99.27%)                                     | Eat meal (71.43%)                                     |
|                  | Falling down (98.54%)                                | Reading (72.43%)                                      |
|                  | Put on jacket (98.53%)                               | Sneeze or cough (77.17%)                             |
|                  | Hopping (98.18%)                                    | Play with phone or tablet (78.75%)                    |
| NTU-120 [MS-G3D] | Hugging (99.64%)                                    | Staple book (32.57%)                                 |
|                  | Mountain climber (exercise) (92.59%)                | Falling off chair (11.5%)                            |
|                  | Clean and jerk (89.25%)                             | Cumbia (12.90%)                                      |
|                  | Front raises (87.37%)                               | Hit with object (60.03%)                             |
|                  | Lunge (87.14%)                                      | Passing American football (not in game) (16.28%)      |
|                  | Golf driving (93.33%)                               | Swinging baseball bat (14.29%)                        |
| Skeletics-152    |                                                      |                                                      |
|                  |                                                      |                                                      |
| Skeleton-Mimetics|                                                      |                                                      |
|                  |                                                      |                                                      |

Table 9  Benchmarking comparison for Metaphorics dataset (mean cosine similarity)

| Method                  | Training dataset | Mean cosine similarity |
|-------------------------|------------------|------------------------|
| 4s-ShiftGCN Cheng et al. (2020) | NTU-120          | 0.04                   |
| MS-G3D Liu et al. (2020)        | NTU-120          | 0.14                   |
| 4s-ShiftGCN Cheng et al. (2020) | Skeletics-152    | 0.08                   |
| MS-G3D Liu et al. (2020)        | Skeletics-152    | 0.10                   |

target phrases in Dumb Charades and interpretative dances are typically enacted indirectly using metaphors. Models trained on other datasets tend to map the skeleton sequence ‘literally’ to action labels. This explains some of the predictions seen in Fig. 5. For example, the model predictions for actions shown in the first row are related to the ground-truth tags (‘fight’, ‘swimming’). The predictions for the other example sequences highlight the shortcomings arising from the literal nature of actions in contextual action datasets as mentioned previously.

6 Discussion

In this section, we analyze the salient trends for skeleton action recognition approaches across different datasets. The list of best-5 and worst-5 classes for various scenarios (datasets) can be viewed in Table 8.

The first two columns correspond to the lab-based indoor datasets—NTU-60 and NTU-120. It is interesting to note that even the worst performing classes of NTU-60 have accuracy in the range 50–70% while the counterparts in NTU-120 exist in a much lower range (32–60%). One reason is that the introduction of NTU-120 resulted in an increase of action classes with subtle, finger-level movements which impacts performance as mentioned previously (Sect. 4.3). Overall, our analysis motivates the need for approaches which can explicitly focus on boosting the performance for classes ranked lowest. Another concurrent requirement arising from our analysis is for skeleton representations which provide finger-level joint information.

We have already seen that average performance in the wild is relatively lower compared to lab-based settings (Tables 2, 5). The results from Table 8 for Skeletics-152 reflect this trend as well. Actions in the wild exhibit large intra-class variability which affects even the best-5 classes (cf. best-5 of NTU-120). Actions belonging to the worst-5 classes in Skeletics-152 and Skeleton-Mimetics are characterized either by high intra-class variability or by containing subtle, finger-dominant motions which cannot be captured...
by existing skeleton representations. Additionally, action sequences in NTU-120 are somewhat choreographed, having a defined starting pose and ending pose, but this is absent in Skeletics-152, where 4s-ShiftGCN is the best performer. A pictorial illustration of performance trends in the top-2 models for selected action classes from Skeletics-152 and Skeleton-Mimetics can be viewed in Fig. 3. Interestingly, even for the classes with lowest performance (worst-3), the correct prediction for Skeletons is often in the list of top-5 model predictions. This is similar to the trend already observed for NTU-120 (Sect. 4).

Such class-level insights cannot be deduced for the proposed Metaphorics dataset since its label set is open-ended, i.e. does not contain a fixed set of action categories. We believe this setting represents an open frontier for skeleton action recognition. The performance on the skeleton version of Metaphorics provides an opportunity to study the generalization capabilities and limitations of existing approaches which are typically optimized for non-interactive, category-based, closed-world recognition paradigms.

7 Conclusion

In this paper, we have examined multiple existing and upcoming frontiers in the landscape of skeleton-based human action recognition. As an important facet of establishing new frontiers for skeleton action understanding in the wild, we curate and introduce three new datasets—Skeletics-152, Skeleton-Mimetics and Metaphorics. Our experiments and benchmarking reveal the capabilities and shortcomings of state-of-the-art recognition models. In addition, the results also highlight the bias induced by processing components (e.g. RGB 3-D pose estimation) and the task paradigm (classification). We hope these findings and the newly introduced datasets will spur the design of better models for ‘in the wild’ actions, both contextual and non-contextual – see the work of Moon et al. (2020) as a preliminary representative example.

As mentioned earlier, our findings can be interactively explored at https://skeleton.iiit.ac.in/. The website features an interactive data analytics dashboard, code and pre-trained models for top-performing skeleton action recognition models and new skeleton action datasets (Skeletics-152, Skeleton-Mimetics, Metaphorics) introduced by us for additional exploration and benefit of the community.

In our current work, we have not examined approaches which map skeleton actions to lexical phrase representations (cf. class labels) Hahn et al. (2019) and Jasani and Mazagonwalla (2019) in detail. We intend to study this promising frontier in the future.

Acknowledgements We wish to thank the anonymous reviewers for their detailed and constructive feedback. We also wish to thank Kalyan Aditya and Sai Shashank Kalakonda for their efforts in creating the project page. This work is partly supported by MeitY, Government of India.

References

Angelini, F., Fu, Z., Long, Y., Shao, L., & Naqvi, S. M. (2018). Actionxpose: A novel 2d multi-view pose-based algorithm for real-time human action recognition.

Ayumi, V. (2016). Pose-based human action recognition with extreme gradient boosting. In 2016 IEEE student conference on research and development (SCOREd), pp. 1–5. IEEE.

Caetano, C., Brémont, F., Schwartz, W. R. (2019). Skeleton image representation for 3d action recognition based on tree structure and reference joints. In Conference on graphics, patterns and images (SIGGRAPI).

Caetano, C., Sena, J., Brémont, F., dos Santos, J. A., & Schwartz, W. R. (2019). Skelenotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In IEEE international conference on advanced video and signal-based surveillance (AVSS).

Cai, Y., Li, H., Hu, J. F., & Zheng, W. S. (2019). Action knowledge transfer for action prediction with partial videos. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 8118–8125. https://doi.org/10.1609/aaai.v33i01.33018118.

Cao, Y., Barrett, D., Barbu, A., Narayananwar, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J., & Wang, S. (2013). Recognize human activities from partially observed videos. In: Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society conference on computer vision and pattern recognition, pp. 2658–2665. https://doi.org/10.1109/CVPR.2013.343.

Cao, Z., Hidalgo Martínez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Carreira, J., Noland, E., Hillier, C., Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. CoRR arXiv:1907.00987.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 4724–4733.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Chéron, G., Laptev, I., Schmid, C. (2015). P-cnn: Pose-based cnn features for action recognition. In Proceedings of the IEEE international conference on computer vision, pp. 3218–3226.

Chunhui, L., Yueyu, H., Yanghao, L., Sijie, S., & Jiaying, L. (2017). Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. ACM multimedia workshop.

Du, Y., Fu, Y., & Wang, L. (2015). Skeleton based action recognition with convolutional neural network. In 2015 3rd IAPR Asian conference on pattern recognition (ACPR), pp. 579–583. https://doi.org/10.1109/ACPR.2015.7486569

Eweisi, A., Cheema, M. S., Bauckhage, C., & Gall, J. (2014). Efficient pose-based action recognition. In Asian conference on computer vision (pp. 428–443). Springer.

Hahn, M., Silva, A., & Rehg, J. M. (2019). Action2vec: A cross-modal embedding approach to action learning. arXiv preprint arXiv:1901.00484.
Halim, A. A., Dartigues-Pallez, C., Precioso, F., Riveill, M., Benslimane, A., & Ghoneim, S. (2016). Human action recognition based on 3d skeleton part-based pose estimation and temporal multi-resolution analysis. In 2016 IEEE international conference on image processing (ICIP) (pp. 3041–3045). IEEE.

Hussein, M., Torki, M., Gowayyed, M., & El Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In IJCAI international joint conference on artificial intelligence.

Jasani, B., & Mazonwalla, A. (2019). Skeleton based zero shot action recognition in joint pose-language semantic space. arXiv preprint arXiv:1911.11344.

Jhuang, H., Gall, J., Zufii, S., Schmid, C., & Black, M. J. (2013). Towards understanding action recognition. In The IEEE International Conference on Computer Vision (ICCV).

Kim, T. S., & Reiter, A. (2017). Interpretable 3d human action analysis with temporal convolutional networks. CoRR arXiv:1704.04516.

Kipp, M. (2001). Anvil—a generic annotation tool for multimodal dialogue. In Seventh European conference on speech communication and technology.

Kocabas, M., Athanasiou, N., & Black, M. J. (2020). Vibe: Video inference for human body pose and shape estimation. In The IEEE conference on computer vision and pattern recognition (CVPR).

Kolotouros, N., Pavlakos, G., Black, M. J., & Daniilidis, K. (2019). Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In ICCV.

Kundu, J. N., Gor, M., Uppala, P. K., & Babu, R. V. (2018). Unsupervised feature learning of human actions as trajectories in pose embedding manifold. CoRR arXiv:1812.02592.

Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., & He, M. (2017). Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 601–604. IEEE.

Li, C., Zhong, Q., Xie, D., & Pu, S. (2017). Skeleton-based action recognition with convolutional neural networks. CoRR arXiv:1704.07595.

Li, C., Zhong, Q., Xie, D., & Pu, S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18, pp. 786–792. International joint conferences on artificial intelligence organization. https://doi.org/10.24963/ijcai.2018/109. https://doi.org/10.24963/ijcai.2018/109.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2019). Actional-structural graph convolutional networks for skeleton-based action recognition. In The IEEE conference on computer vision and pattern recognition (CVPR).

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3d points. In 2010 IEEE Computer Society conference on computer vision and pattern recognition-workshops (pp. 9–14). IEEE.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence., https://doi.org/10.1109/TPAMI.2019.2916873.

Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In The IEEE conference on computer vision and pattern recognition (CVPR).

Ma, S., Sigal, L., & Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1942–1950.

Moon, G., Kwon, H., Lee, K. M., & Cho, M. (2020). Integral action: Pose-driven feature integration for robust human action recognition in videos.

Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., & Weber, A. (2007). Documentation mocap database hdm05. Tech. Rep. CG-2007-2, Universität Bonn.

Peng, W., Hong, X., Chen, H., Zhao, G. (2020). Learning graph convolutional network for skeleton-based human action recognition by neural searching. The thirty-fourth AAAI conference on artificial intelligence, AAAI.

Presti, L. L., & La Cascia, M. (2016). 3d skeleton-based human action classification: A survey. Pattern Recognition, 53, 130–147.

Rogez, G., Weinaefp, P., & Schmid, C. (2019). LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ryoo, M. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. Proceedings of the IEEE international conference on computer vision, 1036–1043. https://doi.org/10.1109/ICCV.2011.6126349.

Sadeghipour, A., & Morency, L. P. (2011). 3d iconic gesture dataset. https://doi.org/10.4119/UNIBI/2683224, https://pub.uni-bielefeld.de/record/2683224.

Seidenari, L., Varano, V., Berretti, S., Bimbo, A., & Pala, P. (2013). 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, pp. 479–485.

Shahroudy, A., Liu, J., Ng, T. T., Wang, G. (2016). Nut rgb+d: A large scale dataset for 3d human activity analysis. In The IEEE conference on computer vision and pattern recognition (CVPR).

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Skeleton-based action recognition with directed graph neural networks. In The IEEE conference on computer vision and pattern recognition (CVPR).

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In CVPR.

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In The IEEE conference on computer vision and pattern recognition (CVPR).

Sigurdsson, G. A., Russakovskiy, O., & Gupta, A. (2017). What actions are needed for understanding human actions in videos? In ICCV.

Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2016). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. CoRR arXiv:1611.00607.

Song, Y. F., Zhang, Z., & Wang, L. (2019). Richly activated graph convolutional network for action recognition with incomplete skeletons. In International conference on image processing (ICIP). IEEE.

Tang, Y., Tian, Y., Lu, J., Li, P., & Zhou, J. (2018). Deep progressive reinforcement learning for skeleton-based action recognition. In The IEEE conference on computer vision and pattern recognition (CVPR).

Vemulpalari, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. 2014 IEEE conference on computer vision and pattern recognition, pp. 588–595.

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In IEEE International conference on computer vision and pattern recognition (CVPR).

Wang, P., Li, W., Ogumbona, P., Wan, J., & Escalera, S. (2018). Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 171, 118–139.

Weinaefp, P., & Rogez, G. (2019). Mimetics: Towards understanding human actions out of context. arXiv preprint arXiv:1912.07249.

Weinaefp, P., & Rogez, G. (2021). Mimetics: Towards understanding human actions out of context. International Journal of Computer Vision., https://doi.org/10.1007/s11263-021-01446-y.
Wu, C., Wu, X.J., & Kittler, J. (2019). Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *The IEEE international conference on computer vision (ICCV) workshops*.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Zhang, J., Li, W., Wang, P., Ogunbona, P., Liu, S., & Tang, C. (2016). A large scale rgb-d dataset for action recognition. In *International workshop on understanding human activities through 3D sensors* (pp. 101–114). Springer.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*, pp. 2117–2126.

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

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. arXiv:1603.07772.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.