https://babel.is.tue.mpg.de/

BABEL: Bodies, Action and Behavior with English Labels

Abhinanda R. Punnakkal*,[1], Arjun Chandrasekaran*,[1], Nikos Athanasiou[1], Alejandra Quirós-Ramírez [2] , Michael J. Black[1]
*equal contribution
[1]MPI Max Planck Institute for Intelligent Systems, [2]Universität Konstanz.

Human Movement & Semantics

Long term goal
- Understanding what actions are being performed, how, and why.
- Requires datasets of human actions with semantic labels.

Problem
- Existing mocap datasets only contain a few actions.
- 3D datasets only label 1 action in the entire sequence.

Idea
- People often perform multiple actions simultaneously, and sequentially, with transitions between them.
- Natural human movement => modeling relationship between:
  a) an action and its movement.
  b) different actions that occur simultaneously and sequentially.

Contribution
- Large dataset of diverse action labels for MoCap sequences in AMASS [1].
- Precise start and end of all actions in the sequence are labeled.

Data Collection

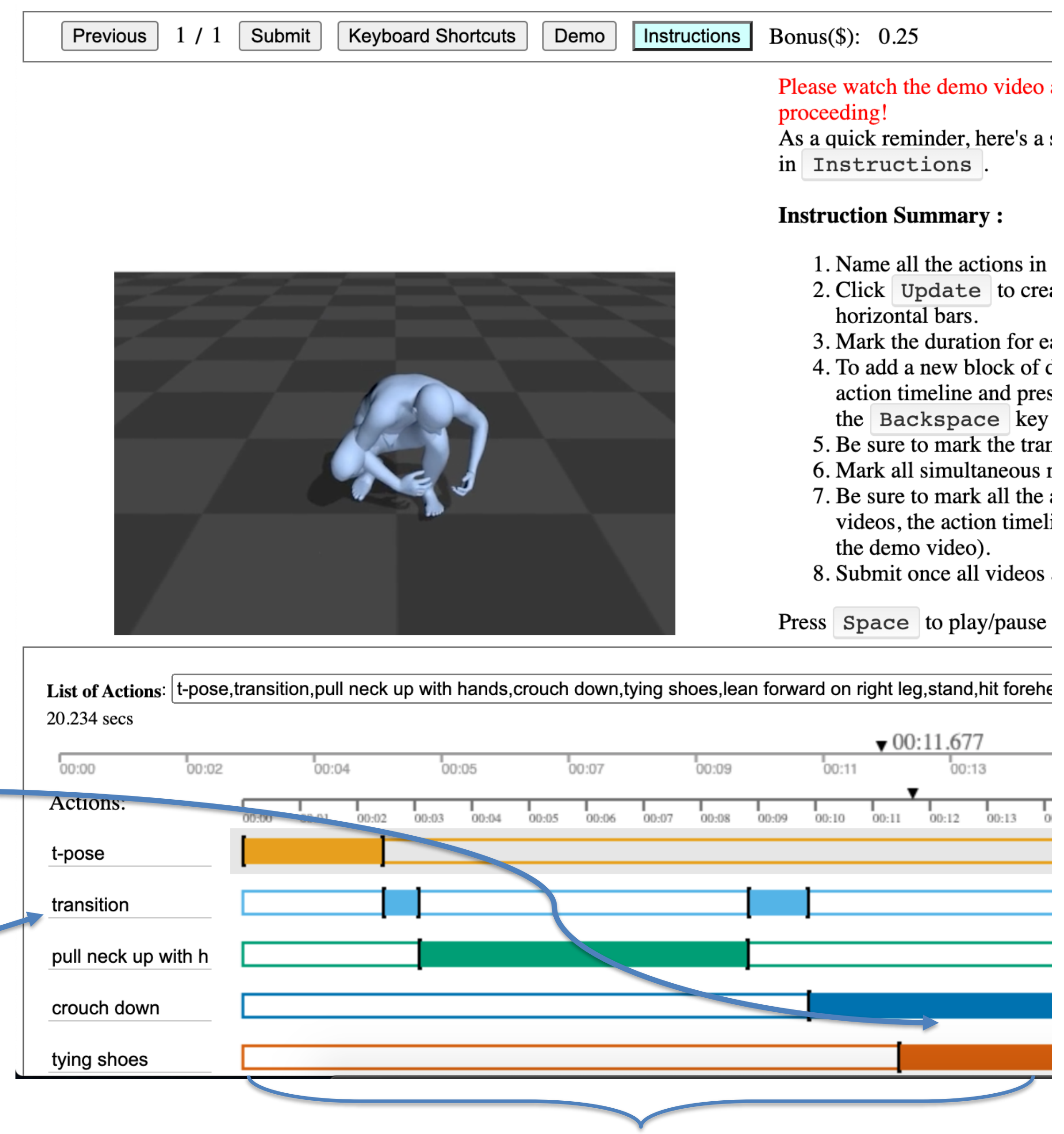
Annotators in Babel are from Amazon Mechanical Turk[1].

Action labels at 2 levels of resolution:
1. **Sequence label** for overall action in entire sequence.
2. **Frame labels** describing precise span of all actions.

Simultaneous actions: All actions occurring in a frame are labeled.

Transitions between actions are explicitly labeled.

Web Interface(frame labels)



Please watch the demo video
proceeding!
As a quick reminder, here's a
in Instructions.

Instruction Summary :

1. Name all the actions in
2. Click Update to cre
horizontal bars.
3. Mark the duration for e
4. To add a new block of
action timeline and pres
the Backspace key
5. Be sure to mark the tra
6. Mark all simultaneous
7. Be sure to mark all the
videos, the action timel
the demo video).
8. Submit once all videos

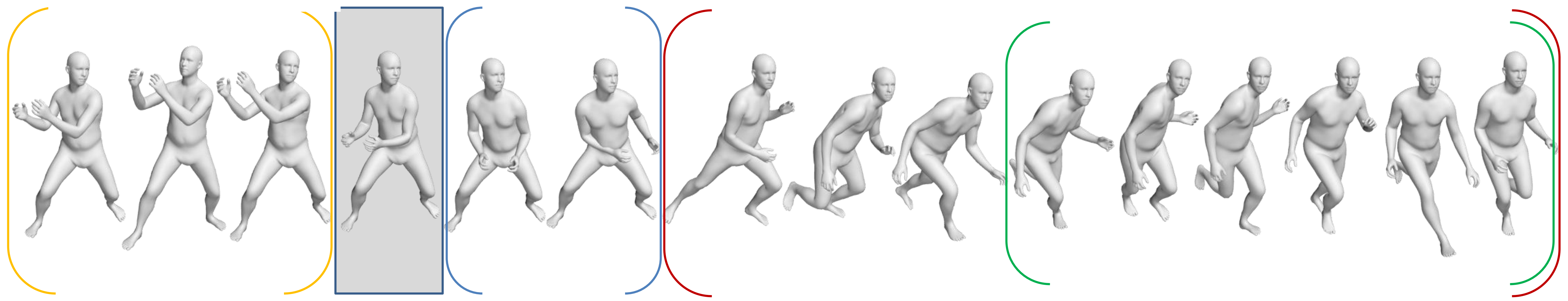
Press Space to play/pause

List of Actions: t-pose,transition,pull neck up with hands,crouch down,tying shoes,lean forward on right leg,stand,hit forehe
20.234 secs

Dense annotation: All frames are labeled with at least 1 action.

[1] https://www.mturk.com/

BABEL data



receive ball with both hands | transition | transfer ball to left hand | sprint | dribble ball with left hand

Dataset Comparison

Dataset	GT motion	# Actions	# Hours	Per Frame	Continuous
CMU Mocap [2]	✓	23	9	✗	✓
MoVi [3]	✓	20	9	✗	✓
Human3.6M [4]	✓	17	18	✗	✓
LaFan [5]	✓	12	4.6	✗	✓
HumanAct12 [6]	✗	12	6	✗	✗
NTU RGBD 60 [7]	✗	60	37	✗	✗
NTU RGBD 120 [8]	✗	120	74	✗	✗
BABEL (Ours)	✓	**> 250**	**43**	✗	✓
			37.5	✓	✓

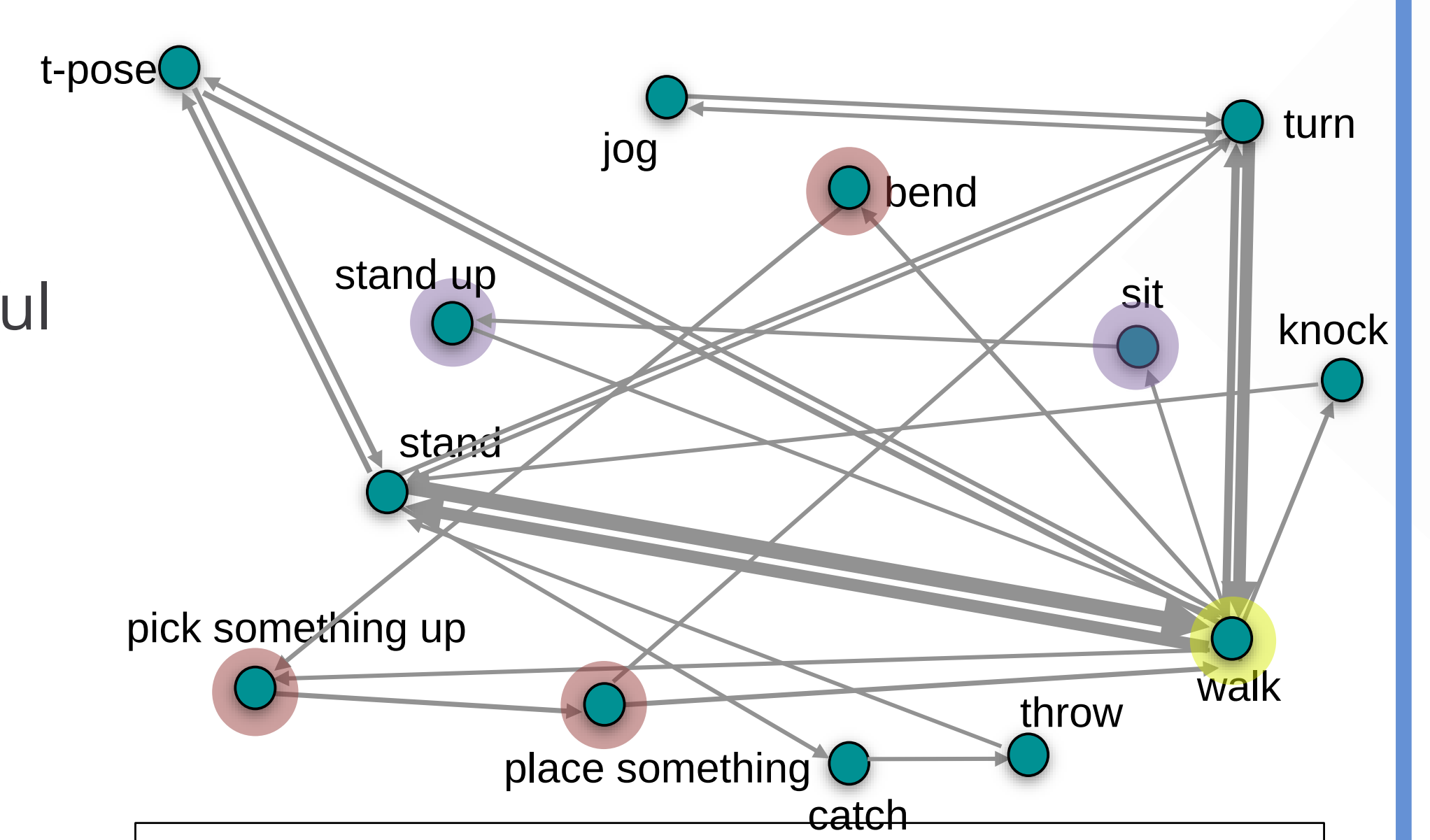
Temporally Adjacent Actions

We visualize the most frequent transitions between actions in BABEL.

"walk" has the most diverse set of adjacent actions.

Semantically meaningful action chains:
- ❑ sit → stand up → walk
- ❑ walk → bend → pick → something up → place something

Frequent action transitions in BABEL

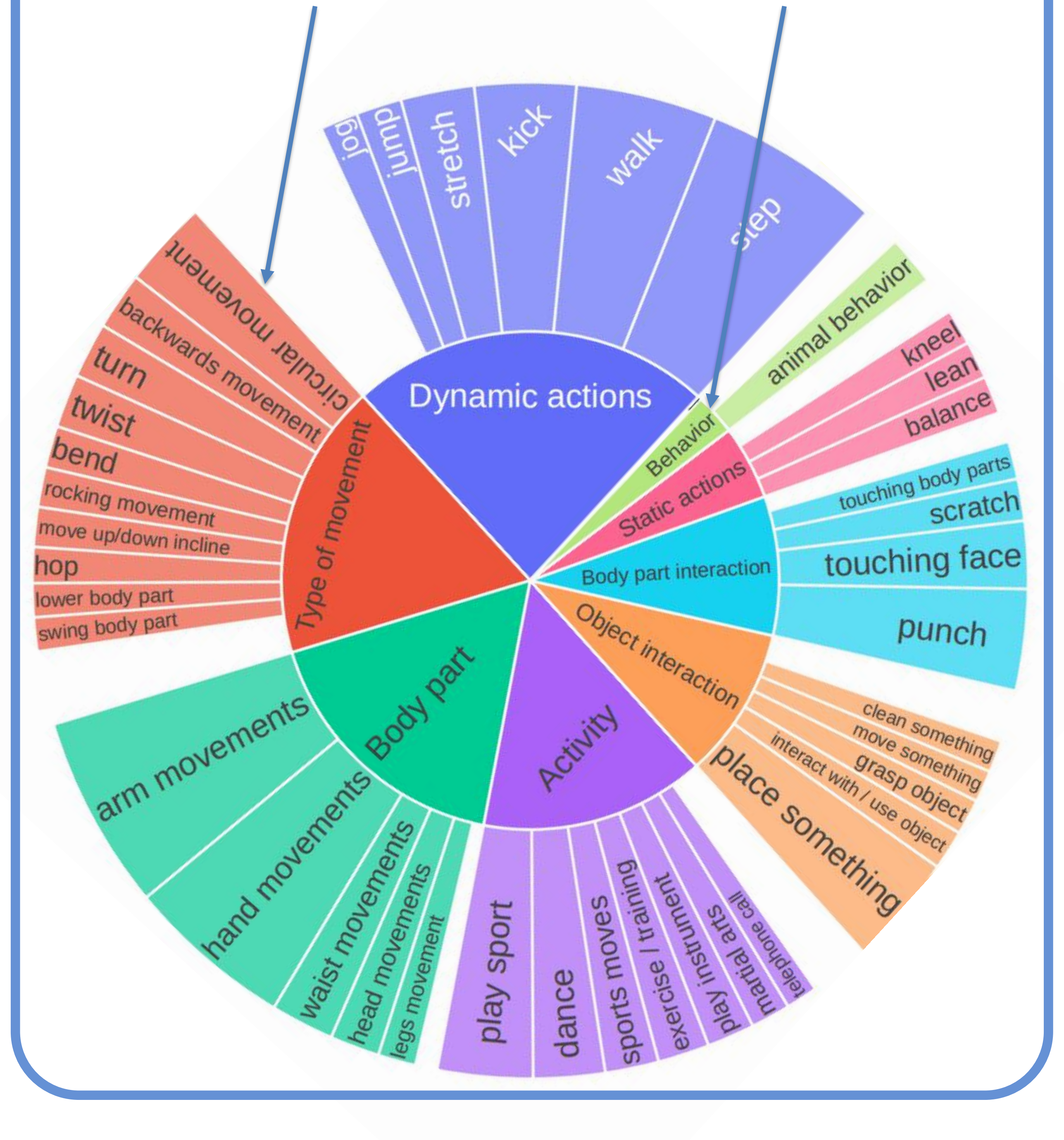


Node: Action
Edge (a → b): action b follows action a

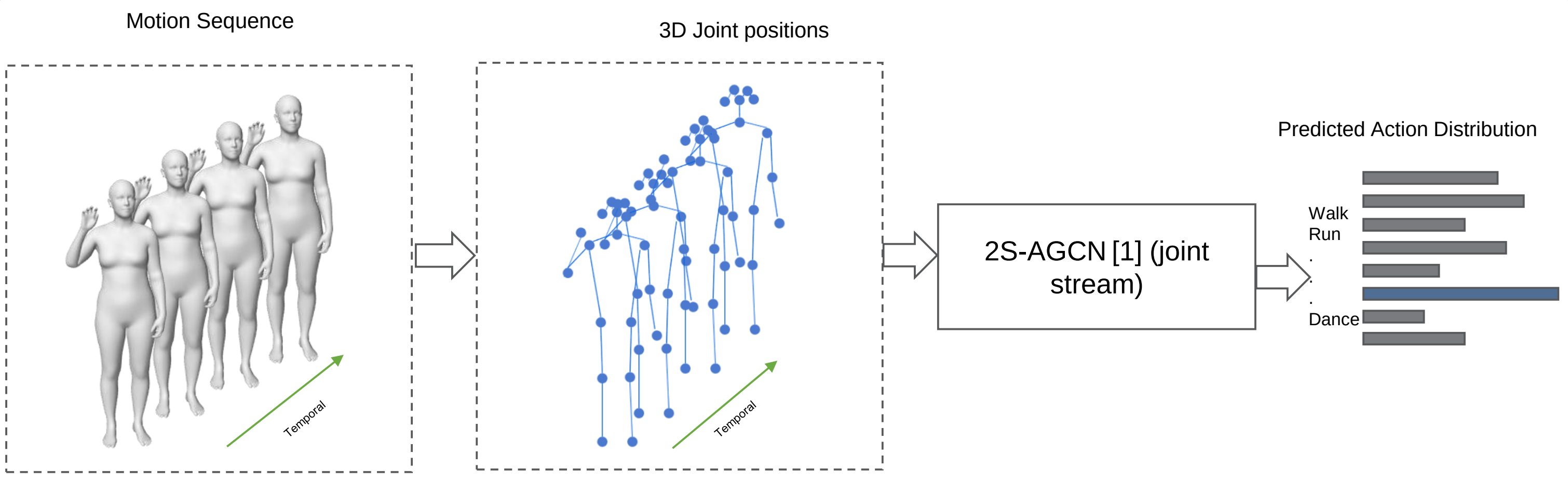
Action Categories

- Action labels for 13220 mocap sequences.
- More than 100k action labels from over 250 unique action categories.
- Long tailed distribution of actions.
- English labels from annotators are organized into:

1. Action Categories   **2. Semantic Categories**



3D Action Recognition Benchmark



- **Task**: 3D Action Recognition
- **Model**: 2s-AGCN [10] (joint stream)
- **Dataset**: BABEL-60 (subset of BABEL)

CE = Cross Entropy
Top-1 norm = Avg. Top-1% across categories.

Dataset	Loss	Top-5%	Top -1%	Top-1 norm%
NTU-RGBD 60 [7]	CE	97.00	85.72	85.79
BABEL-60	CE	73.18	41.14	24.0
	Focal	67.83	33.41	28.10

- **Long tail in BABEL =>** Top-1 norm% << Top-1%.
- **Focal loss =>** Reduction in class bias.
- **Saturating performance on NTU-RGBD 60 =>** BABEL is a more challenging benchmark.

Applications of BABEL: action recognition, motion synthesis, temporal action localization, etc.

References

1. Mahmood N. et al, AMASS: Archive of motion capture as surface shapes, ICCV 2019.
2. CMU Graphics Lab. (Date last accessed 13-November-2020).
3. Ghorbani S. et al, MoVi: A Large Multipurpose Motion and Video Dataset, arXiv 2020.
4. Ionescu C. et al, Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE PAMI 2014.
5. Harvey F. G. et al, Robust Motion In-betweening, SIGGRAPH-TOG 2020.
6. Guo C. et. al, Action2Motion: Conditioned Generation of 3D Human Motions, MM-ACM 2020.
7. Shahroudy A. et al, NTU RGB+D: A large scale dataset for 3d human activity Analysis, CVPR 2016.
8. Liu J. et al, NTU RGB+D 120 : A large scale benchmark for 3d human activity understanding., IEEE PAMI 2016.
9. Lin T. et al, Focal loss for dense object detection. IEEE PAMI 2020.
10. Shi L. et al, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, CVPR 2019