Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition

Mohammad A. Gowayyed[1], Marwan Torki[1], Mohamed E. Hussein[1], Motaz El-Sabban[2]

1 Alexandria University

2 Microsoft Research
Advanced Technology Labs Cairo

Agenda

- Introduction
- Related Work
- Approach
- Experiments
- Conclusion

Human Action Recognition

- **Given:** video of one or more humans performing an "action"
- **Output:** action label(what are they doing?)
- Examples of actions:
  - Walking
  - Running
  - Throwing a ball
  - Waving

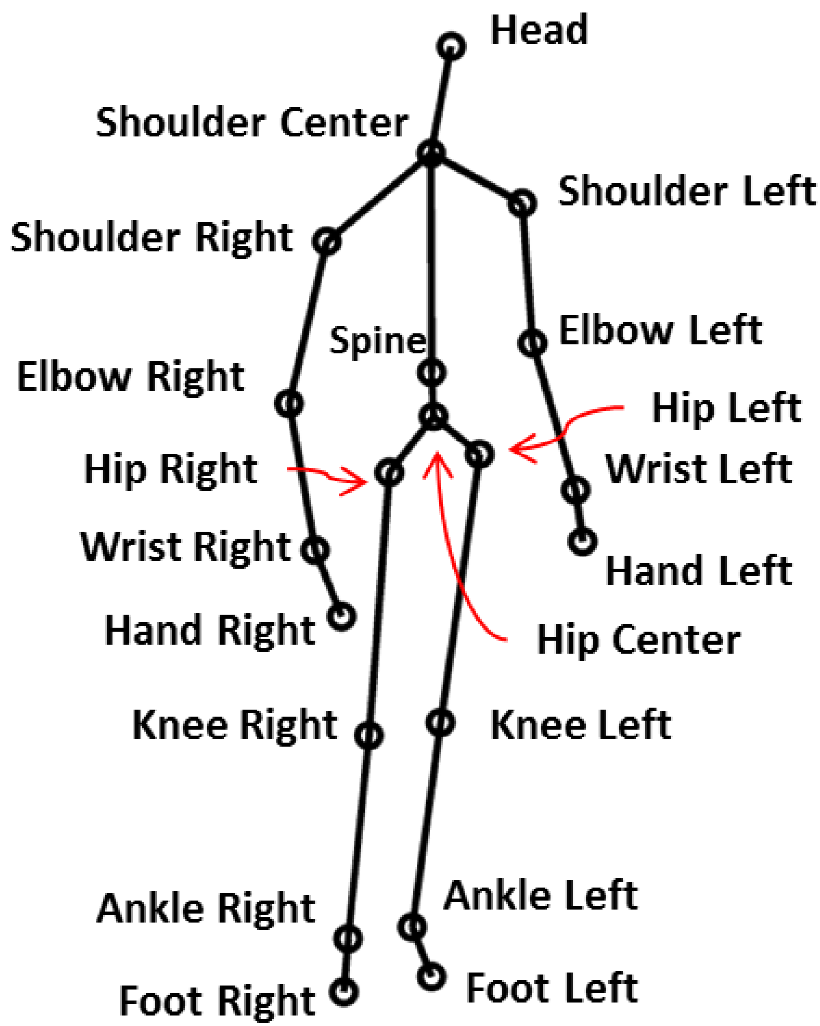
Human Action Recognition

Pose Estimation with Kinect

- [Shotton et al.]* introduced a real-time pose estimation framework using Kinect from a single depth image.
- Perform extensive training on synthetic data
- Provide joint positions at each frame
- We use these joints positions in our recognition approaches

*[Shotton et al.] Real-time human pose recognition in parts from single depth images. In CVPR, 2011.

Pose Estimation with Kinect



Problem Formulation

- Represent a sequence of skeletal joint motions over time using compact, efficient and discriminative descriptor.
- Input
  - Joints Positions
    - $X_{nJoints * nFrames}$
    - $Y_{nJoints * nFrames}$
    - $Z_{nJoints * nFrames}$
- Output
  - Descriptor to use as an input to a classifier

Agenda

- Introduction
- Related Work
- Approach
- Experiments
- Conclusion

Related Work

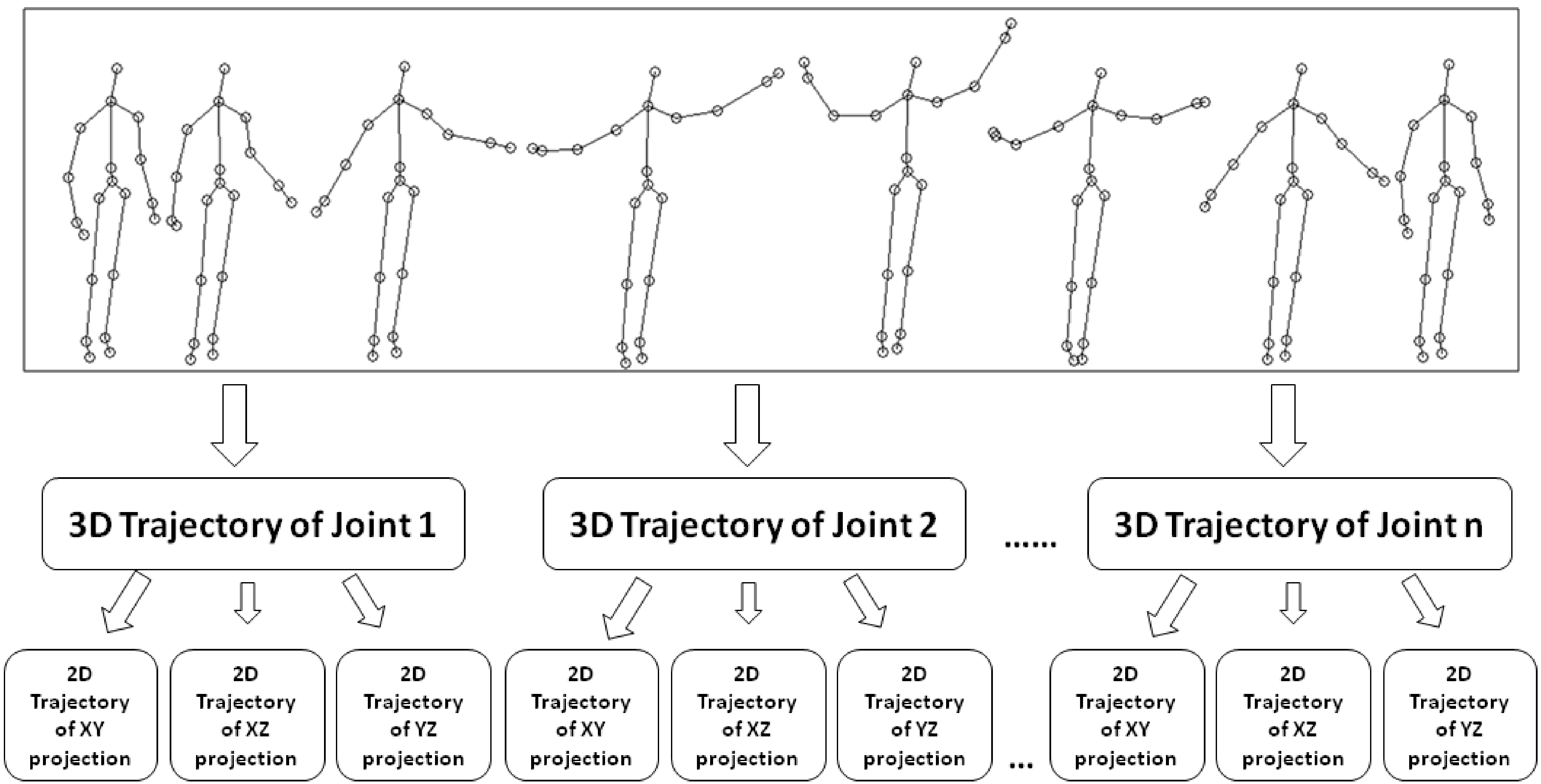
- Similarity measure
  - Dynamic Temporal Warping
- Deal with each frame as a state
  - Recurrent Neural Network
  - Hidden Markov Model
- State-of-the-art:-CVPR 2012
  - Actionlets Ensemble*

*[Wang et al.] Mining actionlet ensemble for action recognition with depth cameras, In CVPR, 2012.

Agenda

- Introduction
- Related Work
- Approach
- Experiments
- Conclusion

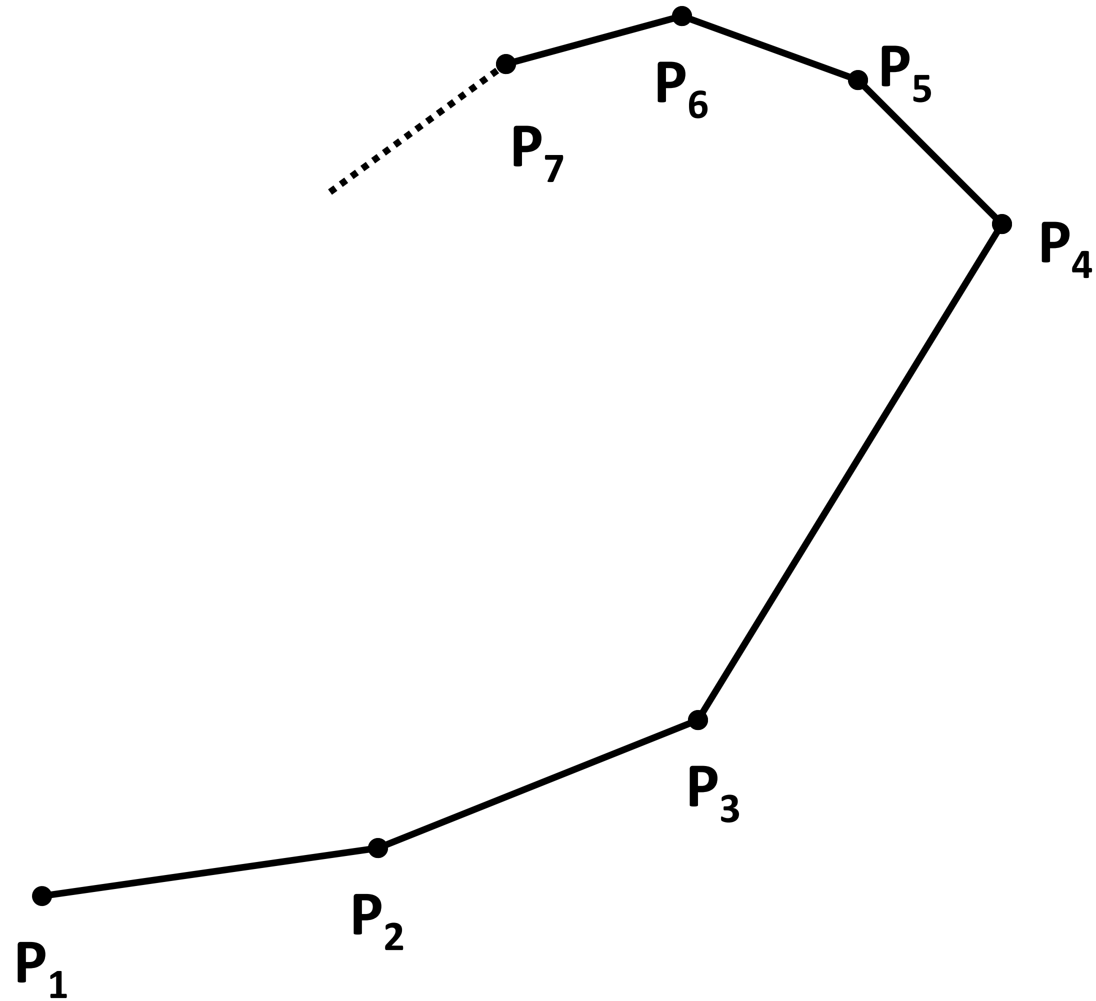
Approach


3D Trajectory of Joint 1
3D Trajectory of Joint 2
......
3D Trajectory of Joint n
2D Trajectory of XY projection
2D Trajectory of XZ projection
2D Trajectory of YZ projection
2D Trajectory of XY projection
2D Trajectory of XZ projection
2D Trajectory of YZ projection
...
2D Trajectory of XY projection
2D Trajectory of XZ projection
2D Trajectory of YZ projection

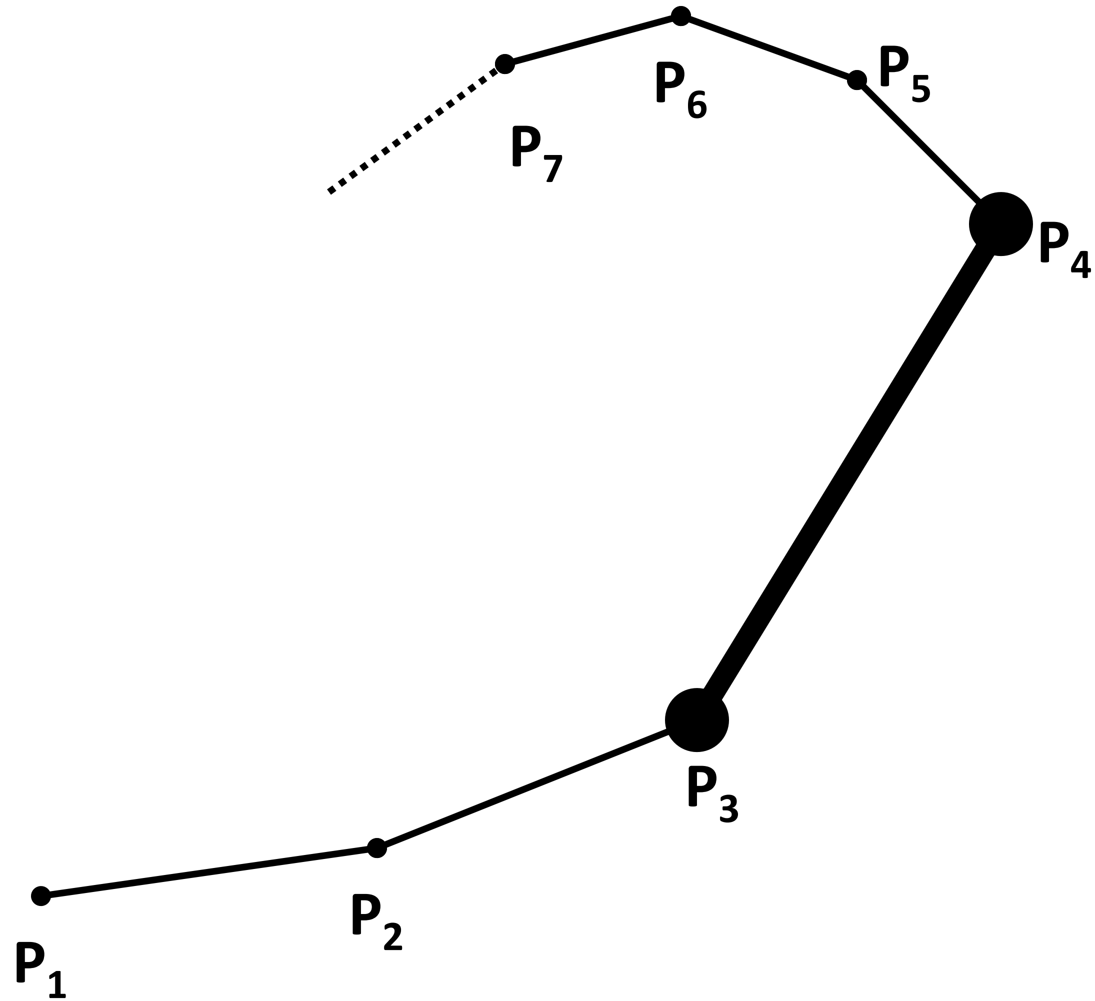
Histogram of Oriented Displacements (HOD)

- Describe a 2D trajectory using a histogram that records how long the object moved in which range of directions.
- This loses the temporal information.
- We use a temporal pyramid to capture the temporal evolution.
- What about 3D?
  – described using the HOD of their 3 2D projections: xy, xz, and yz.

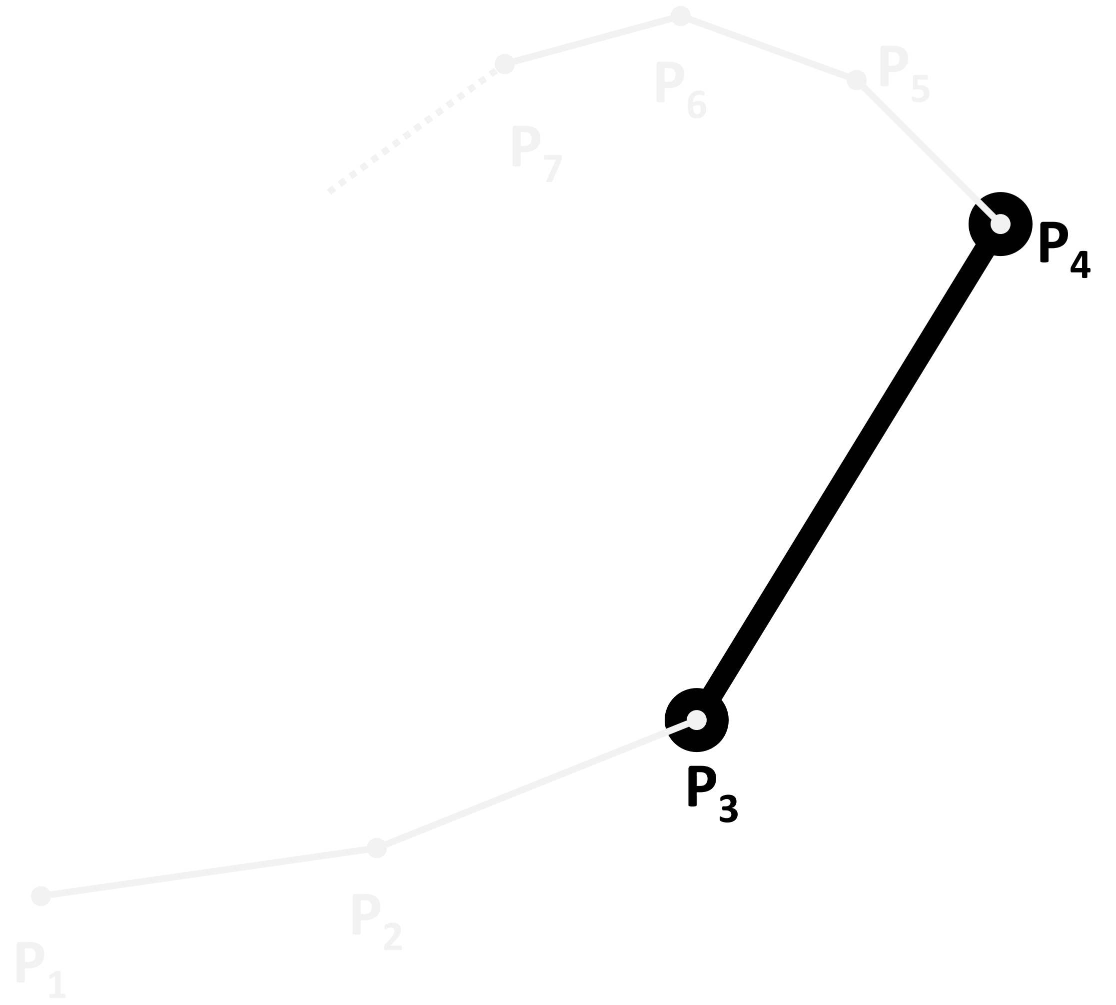
Approach



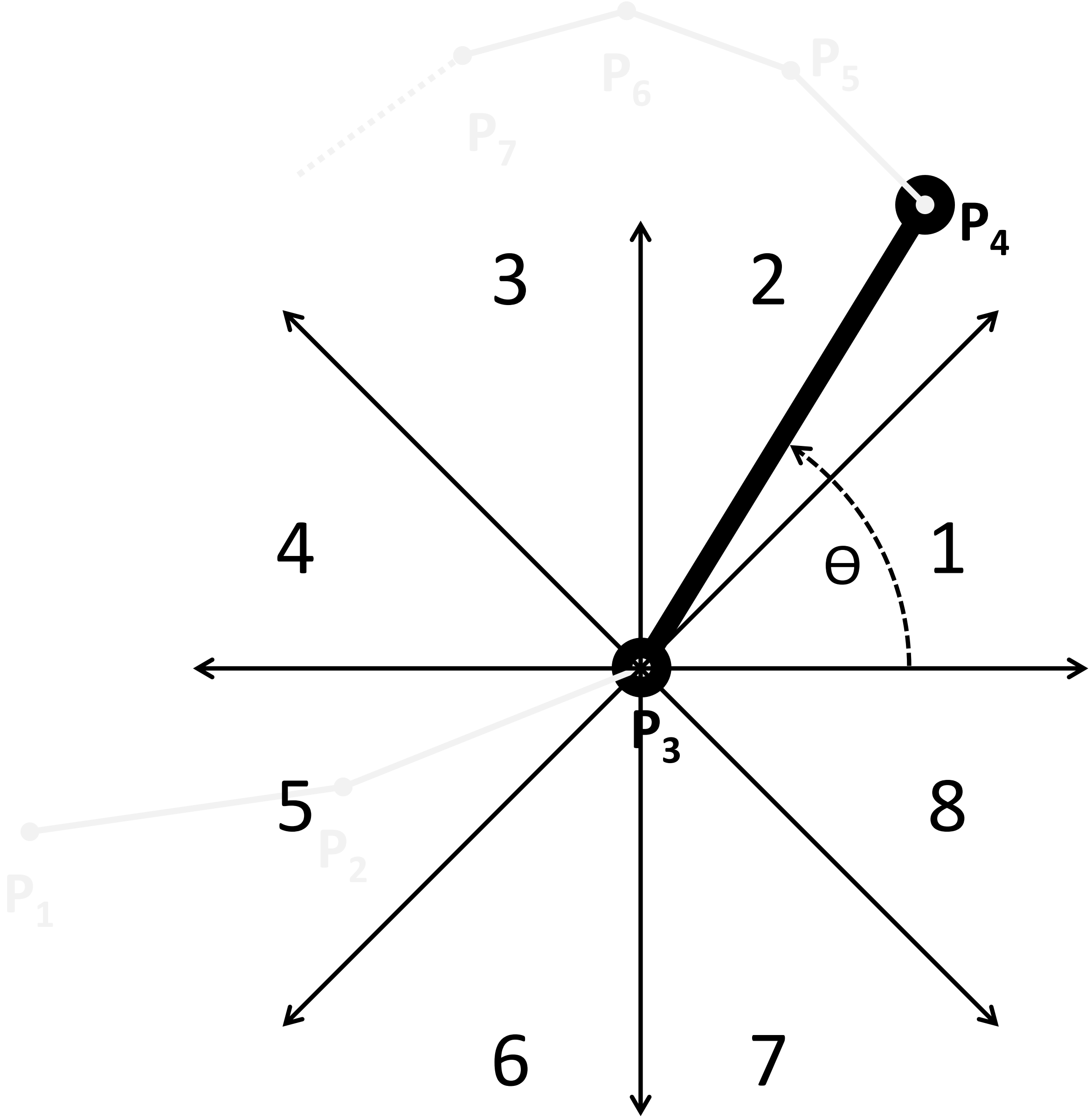
Approach



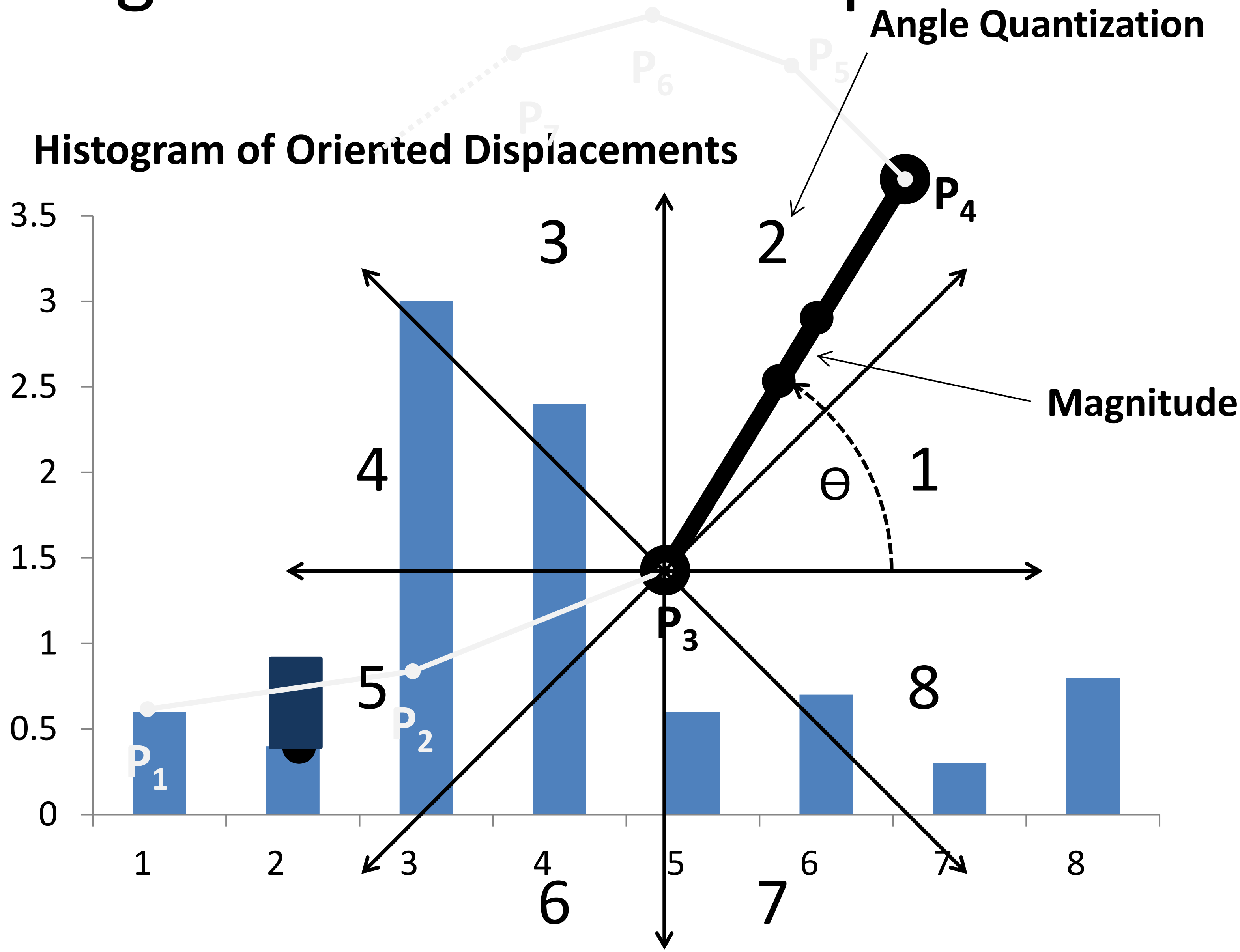
Approach



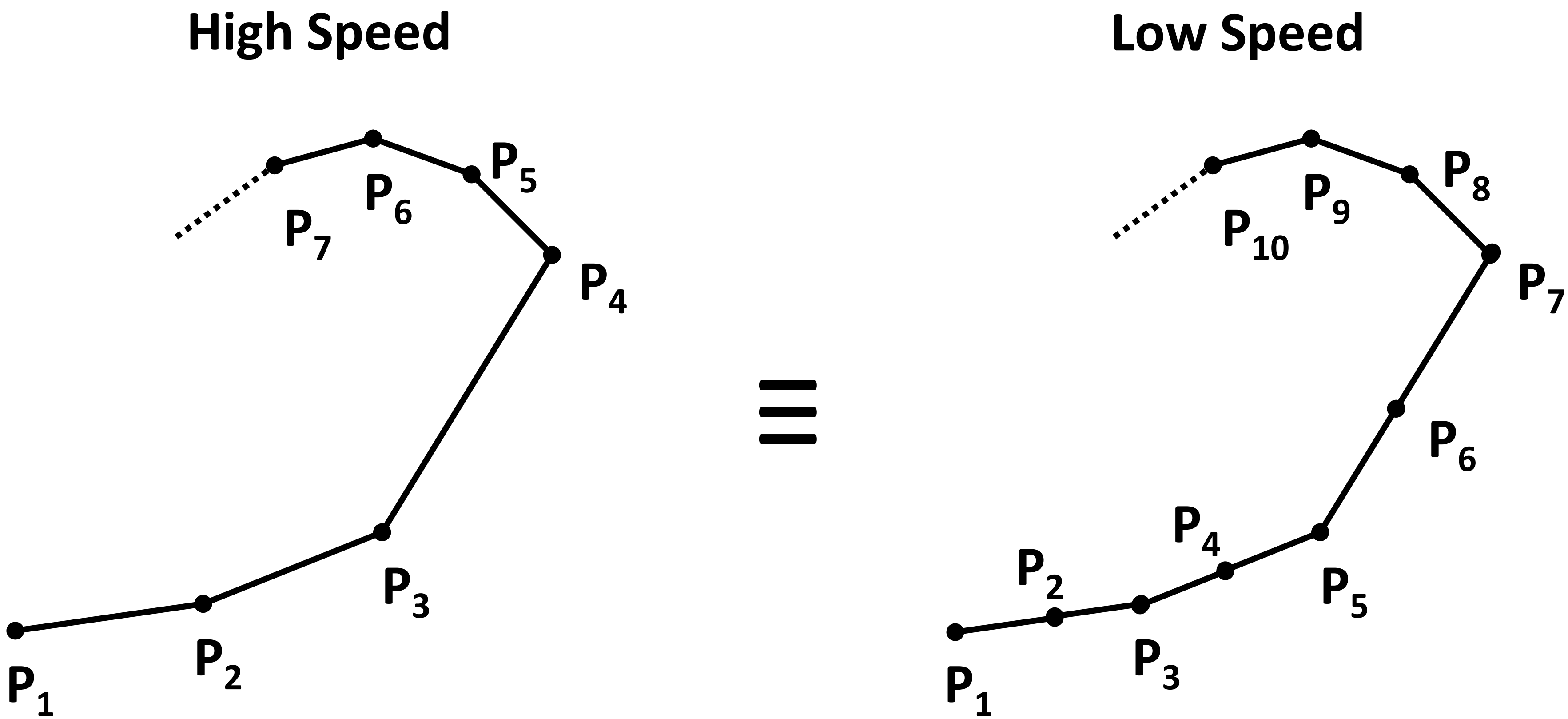
Approach



Histogram of Oriented Displacements

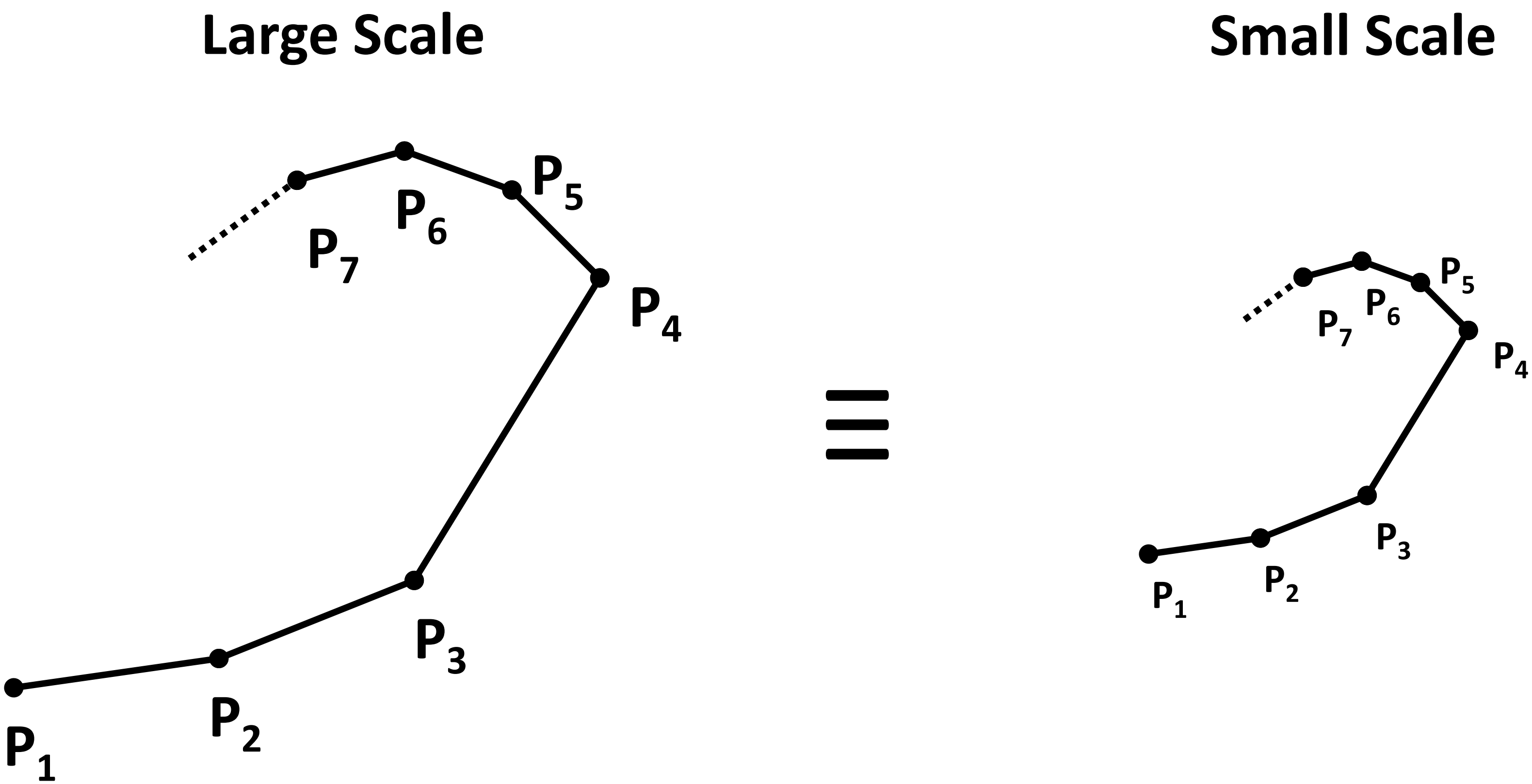


HOD is speed-invariant*



*Given that movement is not far from linearity between positions in the lower resolution.

HOD is scale-invariant*



*Given that the histogram is L2 normalized at the end.

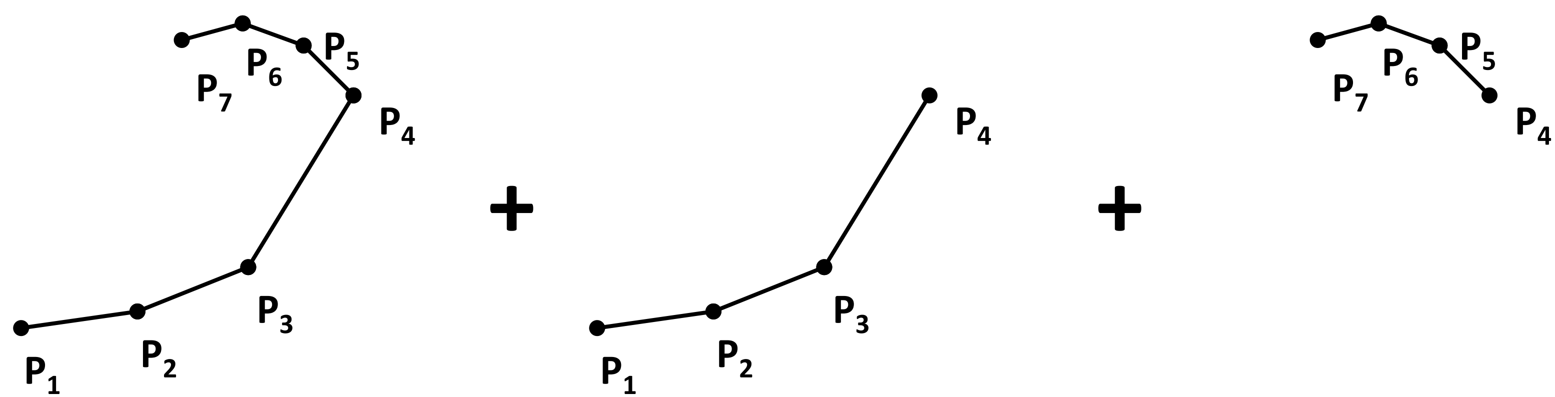
Temporal Information

- If we used HOD to just describe the entire trajectory we will lose the temporal information.
- We solve this by applying a temporal pyramid:
  - describing it all, halves, and quarters (for 3-level pyramid).
- The final HOD is the concatenation of the all descriptors (7 in case of a 3-level HOD).

Temporal Information

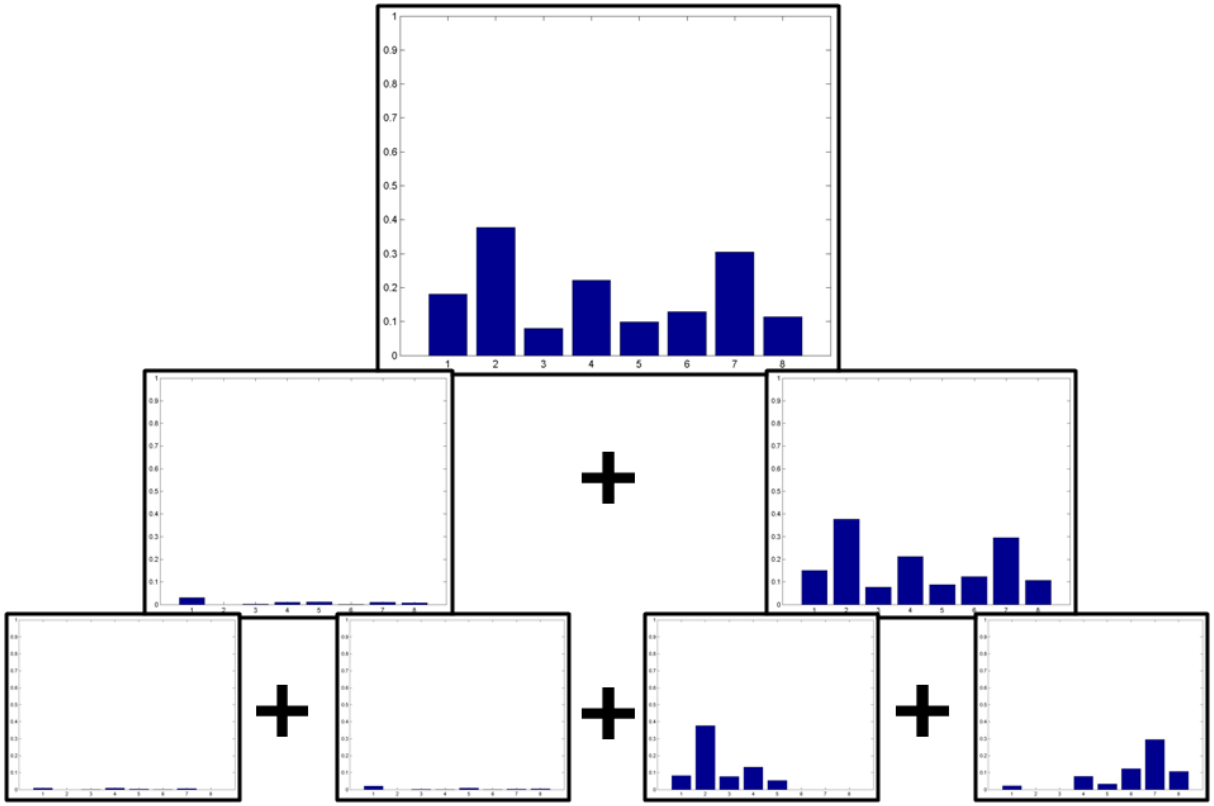
- For a 2-level HOD, the final descriptor is the concatenation of the next three trajectories:

The entire trajectory **First half** **Second half**



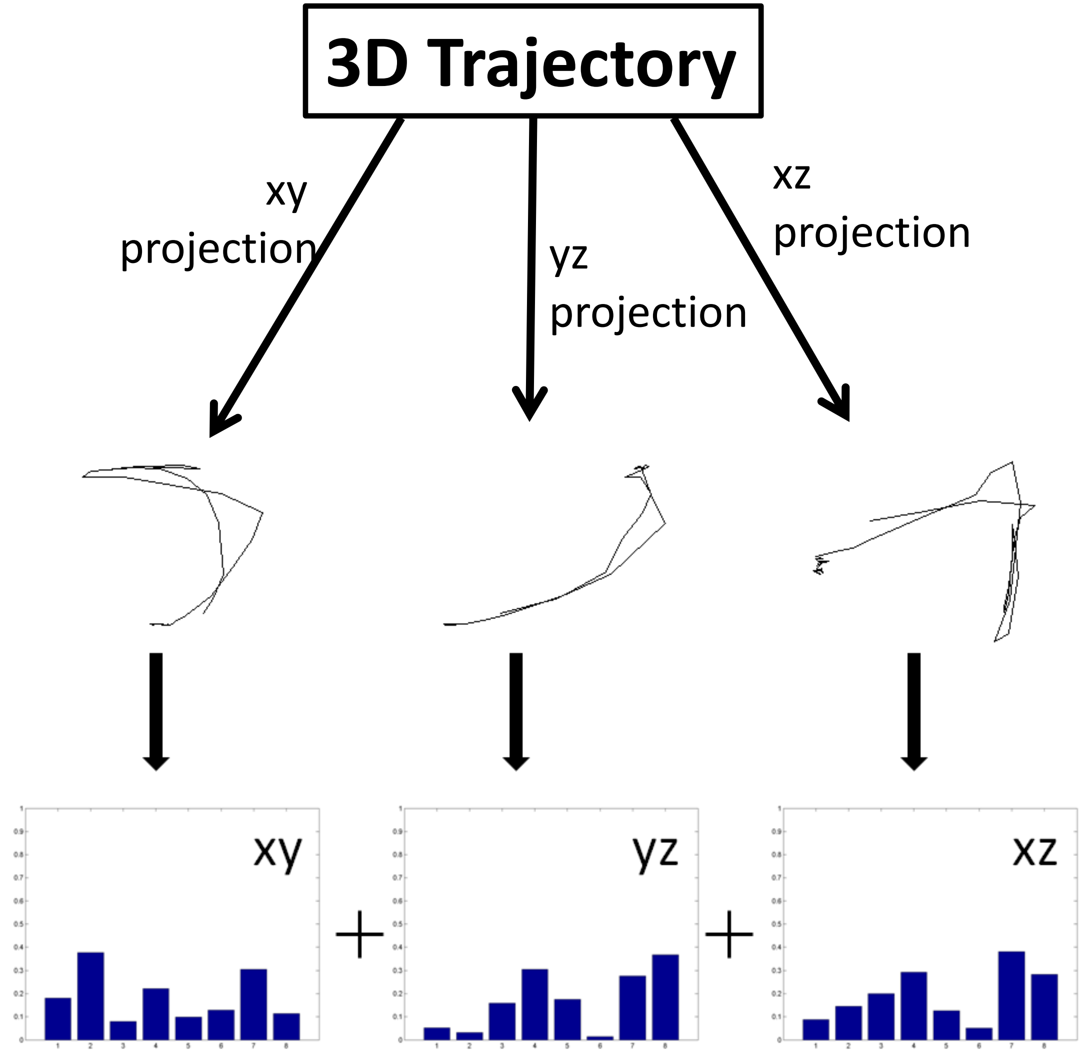
Temporal Pyramid

- 3-level HOD



Using HOD for 3D Trajectories

- Our approach is to describe the 3D trajectories by the HOD of their 3 2D projections (xy, yz, and xz).



Agenda

- Introduction
- Related Work
- Approach
- Experiments
- Conclusion

Datasets

- MSR-Action3D
  - 20 Joints locations are available using a **kinect** sensor.
  - 567 videos.
  - Same setup as in *

Action Set 1	Action Set 2	Action Set 3
Horizontal Wave	High Wave	High Throw
Hammer	Hand Catch	Forward Kick
Forward Punch	Draw X	Side Kick
High Throw	Draw Tick	Jogging
Hand Clap	Draw Circle	Tennis Swing
Bend	Hand Wave	Tennis Serve
Tennis Serve	Forward Kick	Golf Swing
Pickup and Throw	Side Boxing	Pickup and Throw

*[Wang et al.] Mining actionlet ensemble for action recognition with depth cameras, In CVPR, 2012.

Datasets

- HDM05
  - 30 Joints locations are available using a **Motion Capture** system.
  - Actions:
    - deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat, and throw basketball
  - Same setup as in *

*[Ofli et al.] Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In CVPRW, 2012

Results

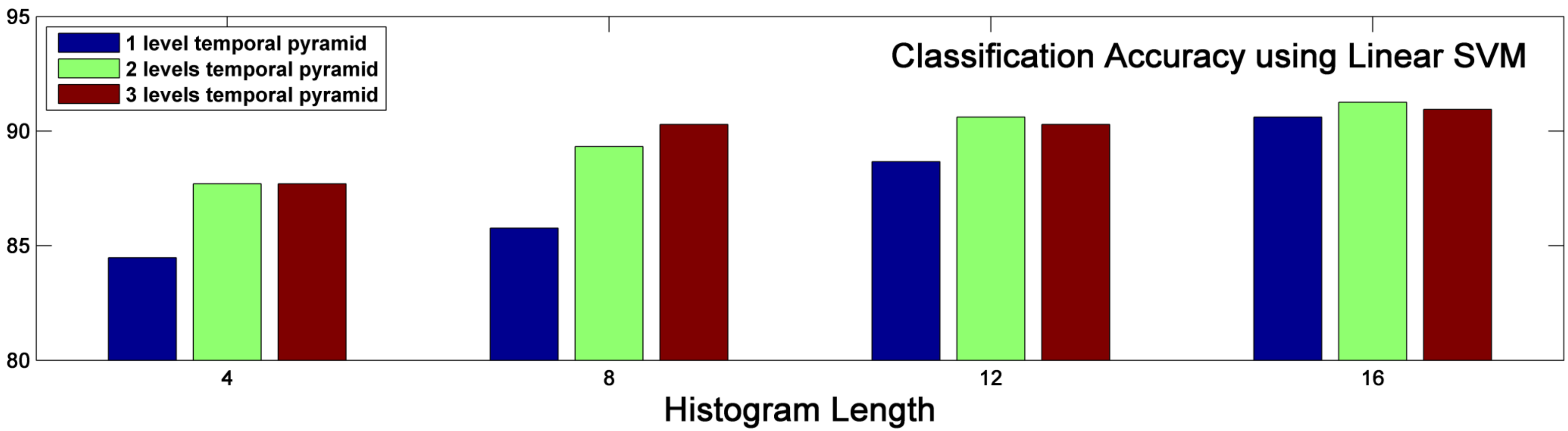
- MSR-Action3D

Method	Accuracy (%)
Actionlets Ensemble*	88.2
2-level 16-bin HOD (20 joints)	**91.26**
2-level 16-bin HOD (right hand joint only)	**74.07**
1-level 4-bin HOD (weakest configuration)	**84.47**

*[Wang et al.] Mining actionlet ensemble for action recognition with depth cameras, In CVPR, 2012.

Results

- MSR-Action3D



Results

- HDM05 – clean data

Method	Accuracy (%)
Sequence of Most Informative Joints*	84.4
3-level 4-bin HOD (20 joints)	**97.27**
3-level 8-bin HOD (right elbow joint only)	**82.72**
1-level 4-bin HOD (weakest configuration)	**80.0**

*[Ofli et al.] Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In CVPRW, 2012

Comparison with the Actionlets Ensemble*

- Their approach:
  - Use Fourier coefficients of relative positions of the whole set of joints as their main descriptor.
  - Introduced a mining algorithm to extract a set of actionlets for each action (each actionlet is a set of joints).
  - Multiple Kernel Learning to combine the actionlets.
  - Has a lot of parameters that are not easy to tune: ambiguity and confidence.

*[Wang et al.] Mining actionlet ensemble for action recognition with depth cameras, In CVPR, 2012.

Comparison with the Actionlets Ensemble*

- Ours:
  - Simpler framework!
  - No ensemble, the descriptor is used directly.
  - We have only two parameters (number of pyramid levels and number of histogram bins), easier to tune.
  - Our weakest configuration still performs very well.

*[Wang et al.] Mining actionlet ensemble for action recognition with depth cameras, In CVPR, 2012.

Agenda

- Introduction
- Related Work
- Approach
- Experiments
- Conclusion

Conclusion

- Introduced HOD: a novel 2D trajectory descriptor.
- Used it to efficiently describe the 3D trajectories of human body joints for action recognition.
- HOD is scale-invariant and speed-invariant.
- Outperformed the state-of-the-art on two popular datasets: MSR-Action3D and HDM05 using Linear SVM.

Thanks, Questions?