Temporal Segmentation of Egocentric Videos

Yair Poleg Chetan Arora Shmuel Peleg

CVPR 2014

Presenter: Hsin-Ping Huang

Egocentric Video

Policeman

UN Inspectors in Syria

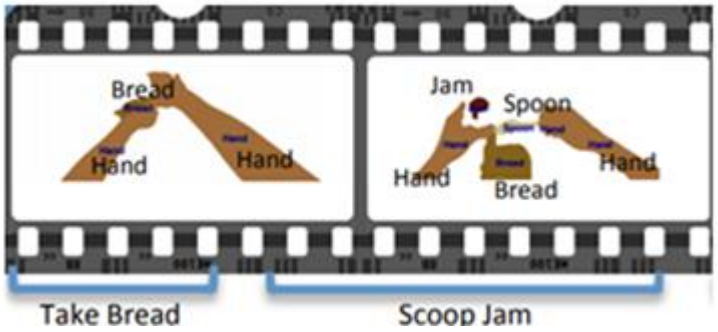
Google Glass

- Browsing **long unstructured** videos is time consuming!
- Video

Video credit: HUJI EgoSeg Dataset

Related Work

Understanding Objects and Activities



[Fathi et al., ICCV 2011]

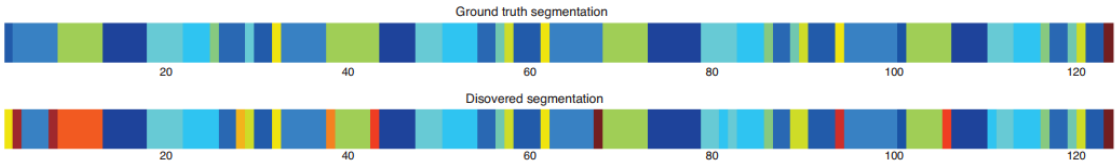
Hard to generalize

[Ryoo et al., CVPR 2013]

Short-term: seconds

Long-term: minutes/hours

Unsupervised Segmentation

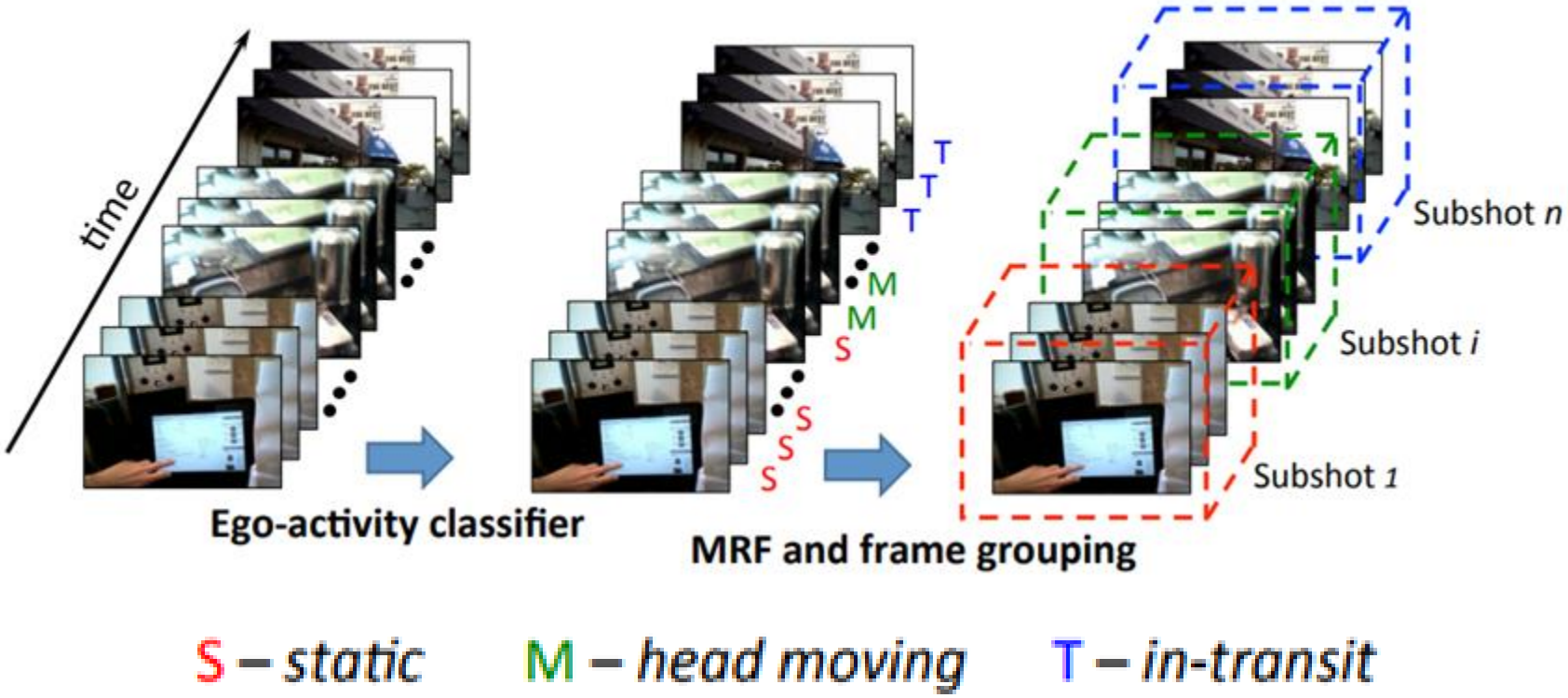


[Kitani et al., CVPR 2011]

Clustering:
no semantic meanings

Related Work

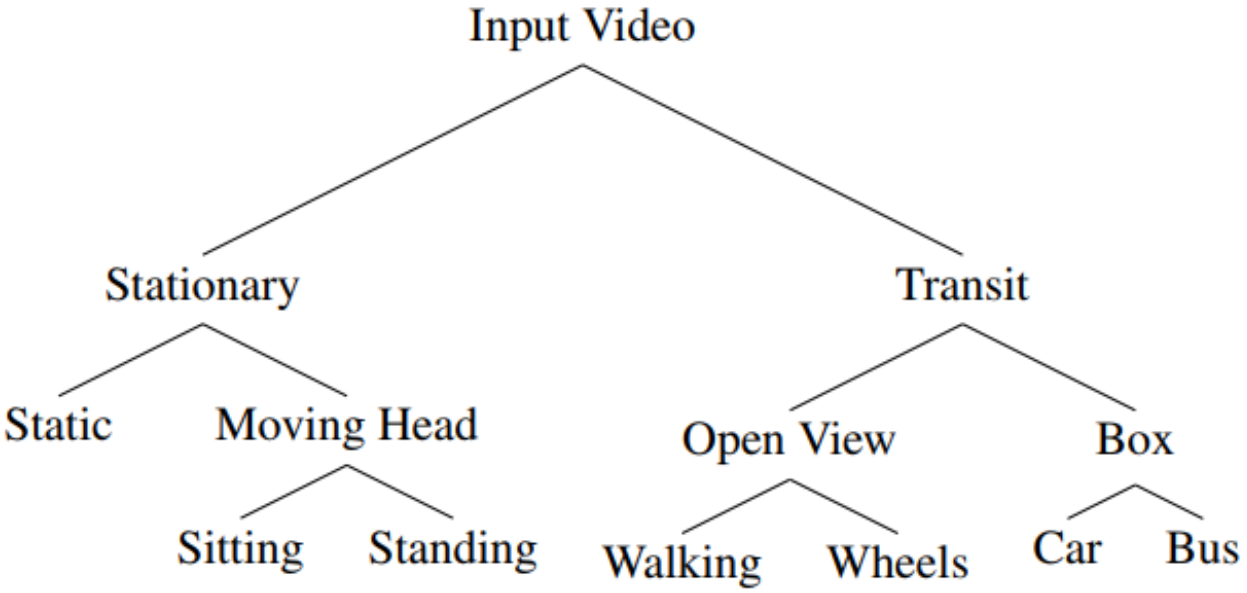
Story-Driven Summarization



[Lu et al., CVPR 2013]

Contribution

- Do **temporal segmentation** into hierarchy of motion classes
- Detect **fixation of wearer's gaze**

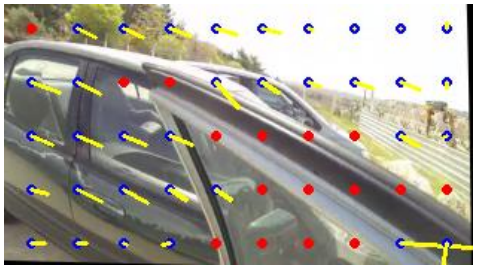


Difficulty

- Two sources of information
  - ***Motion of the wearer***
  - objects and activities

Feature Tracking

- Hard to find ego-motion
  - Head rotation
  - Depth variations
  - Dynamic objects



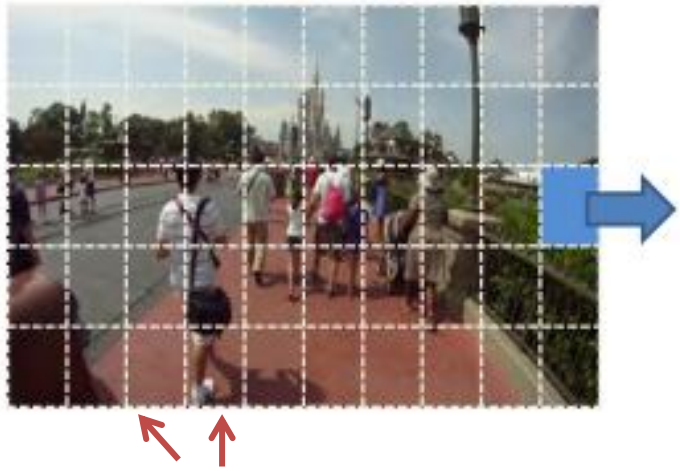
Optical Flow

Image credit: Voodoo Camera Tracker (top)

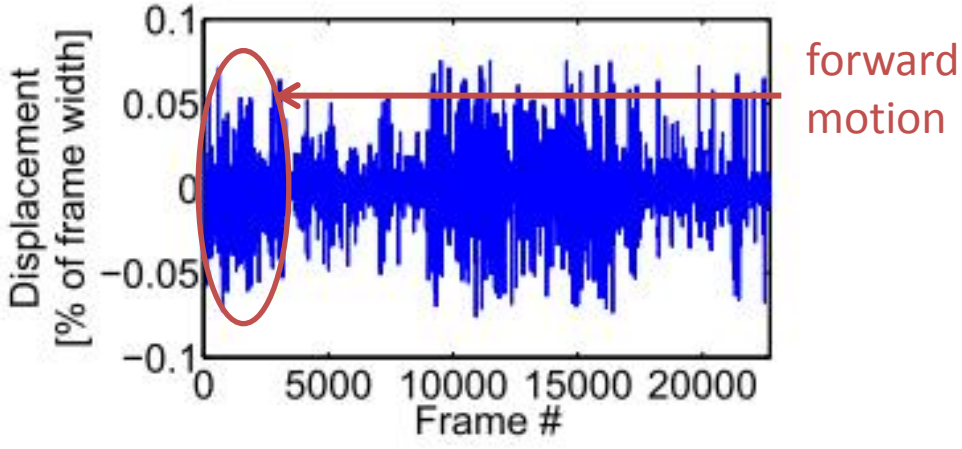
Classification of Wearer’s Motion

Instantaneous Displacement (ID)

- Compute the **ID** at patches



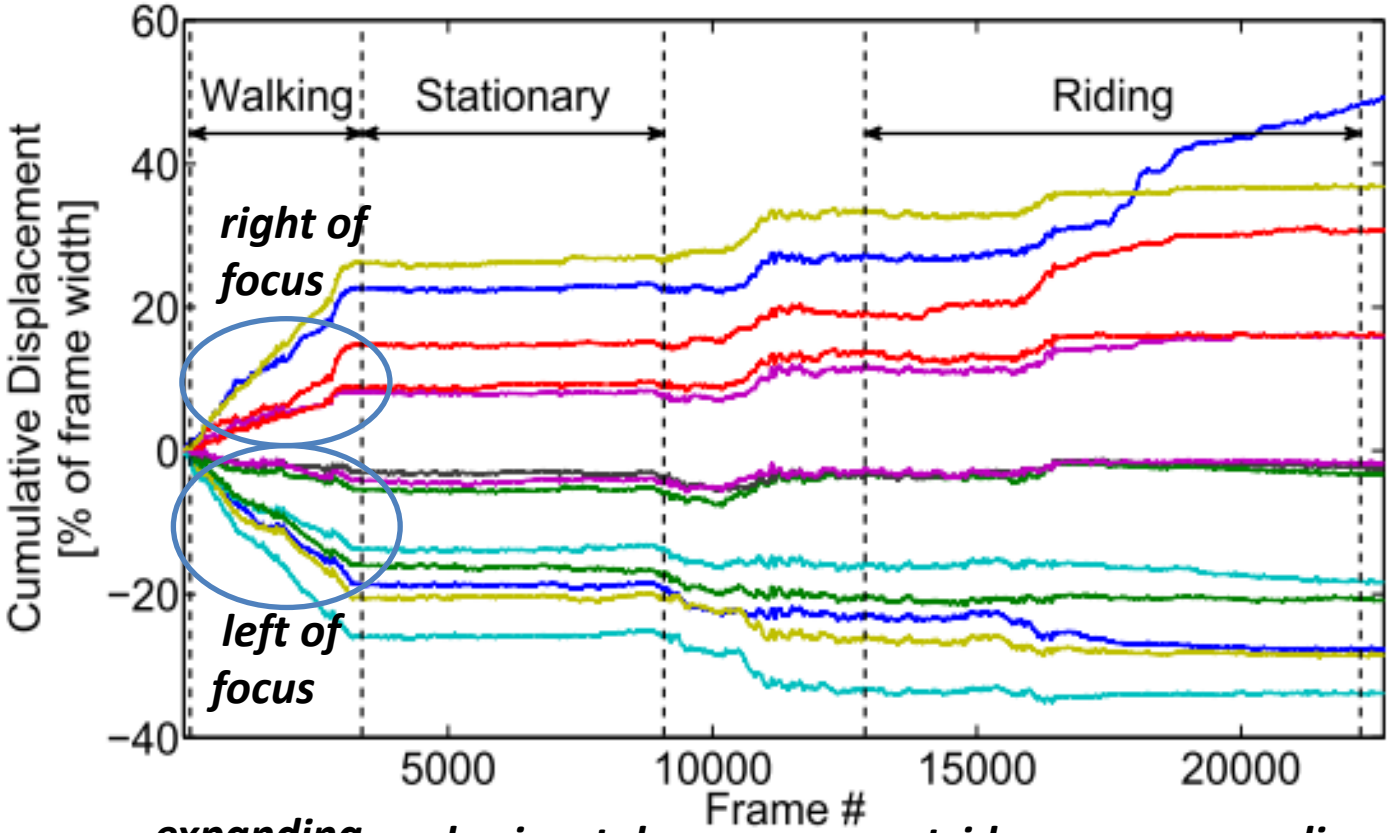
Motion Detector



Instantaneous Displacement of One Patch

Cumulative Displacement (CD)

- Compute the **CD** by integrating the **ID**



expanding curve

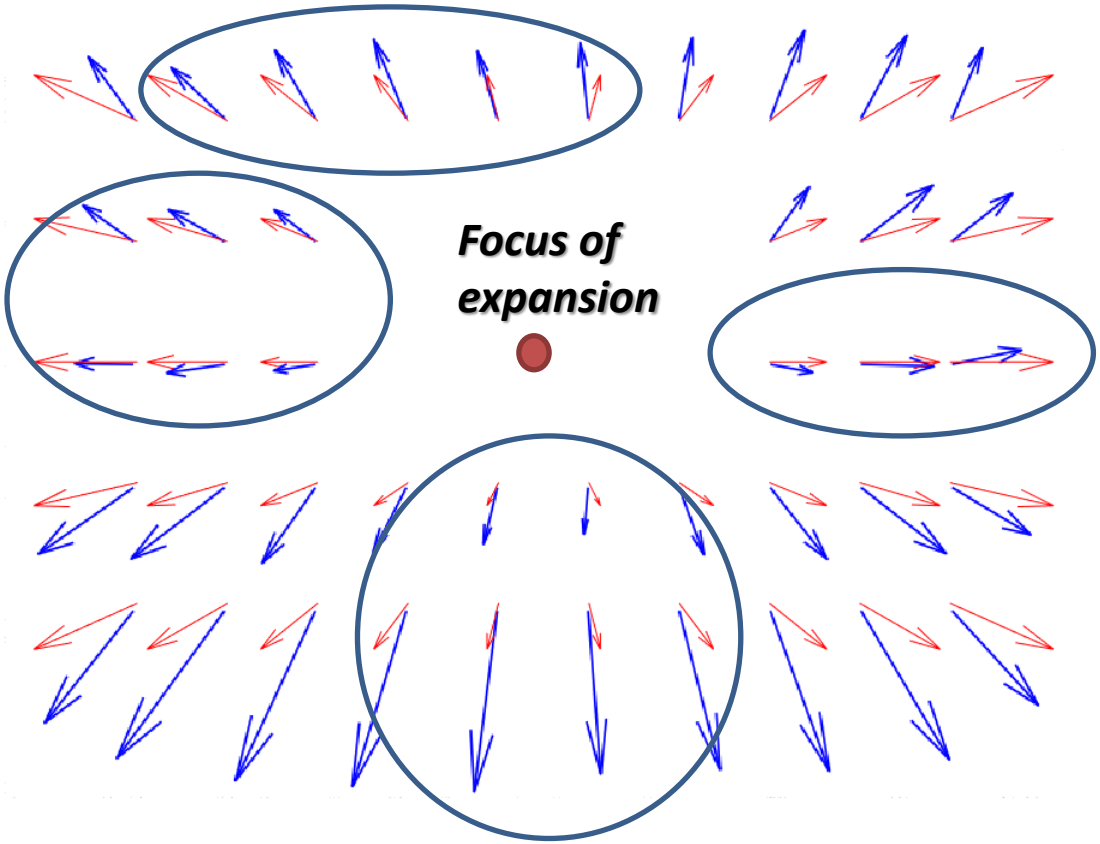
horizontal

outside scene: expanding curve
inside scene: horizontal

Motion Vector and Radial Projection Response

- Compute **motion vectors** as the slopes of smoothed CDs
- Compute **radial projection response**

< ϕ ?



- Video

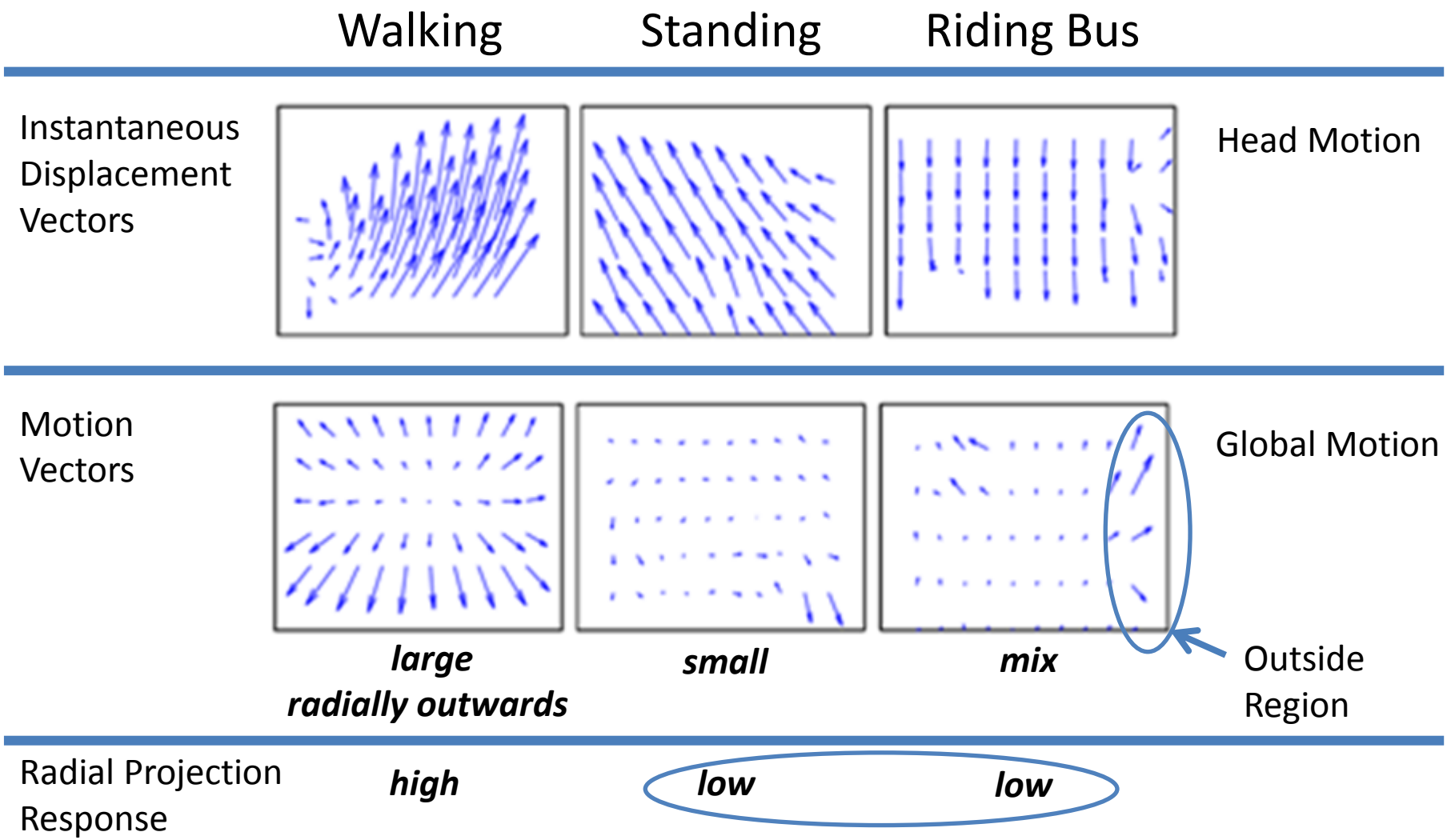
Moving Forward: Expansion from FOE is visible only on smoothed flow

Optical Flow

Smoothed Optical Flow

Video credit: Shmuel Peleg

Motion Vector and Radial Projection Response

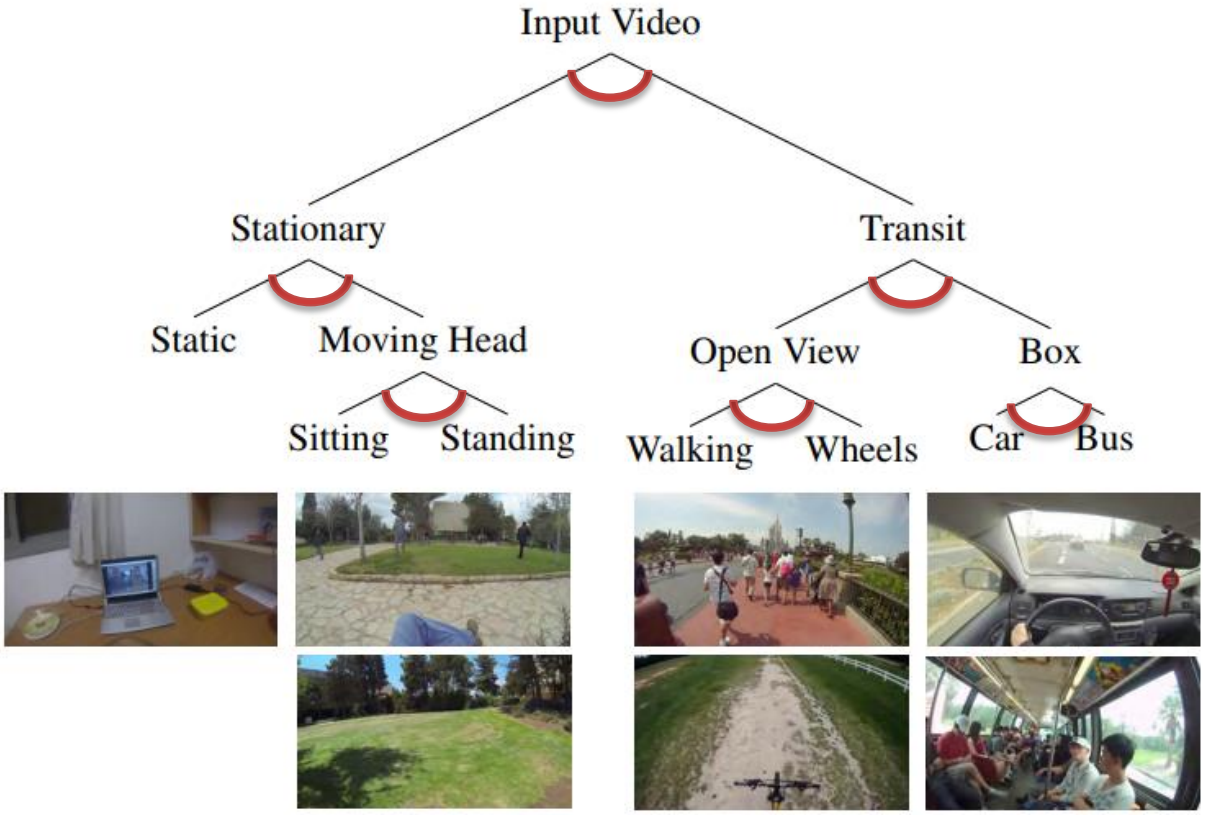


Feature

- AVG of **top/bottom 6% motion vectors**
- DIFF of **top/bottom 6% motion vectors**
- AVG of motion vectors
- Motion vectors
- # of successful flow computation
- AVG and SD of instantaneous displacements
- Radial projection response

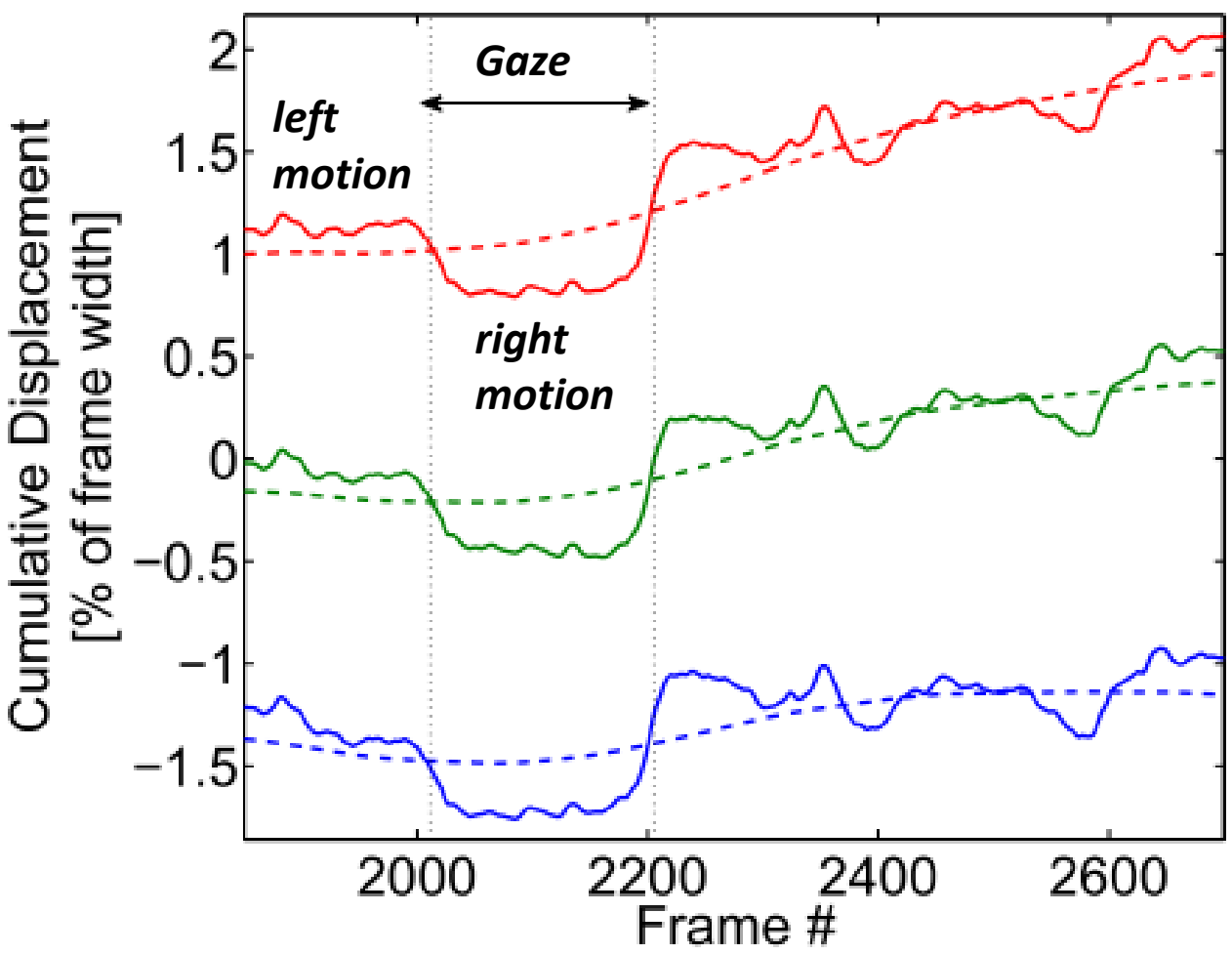
Classifier

- Train SVM classifiers for each binary classification task in the proposed class hierarchy



Detecting Period of Gaze Fixation

Cumulative Displacement

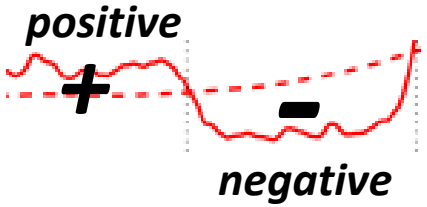


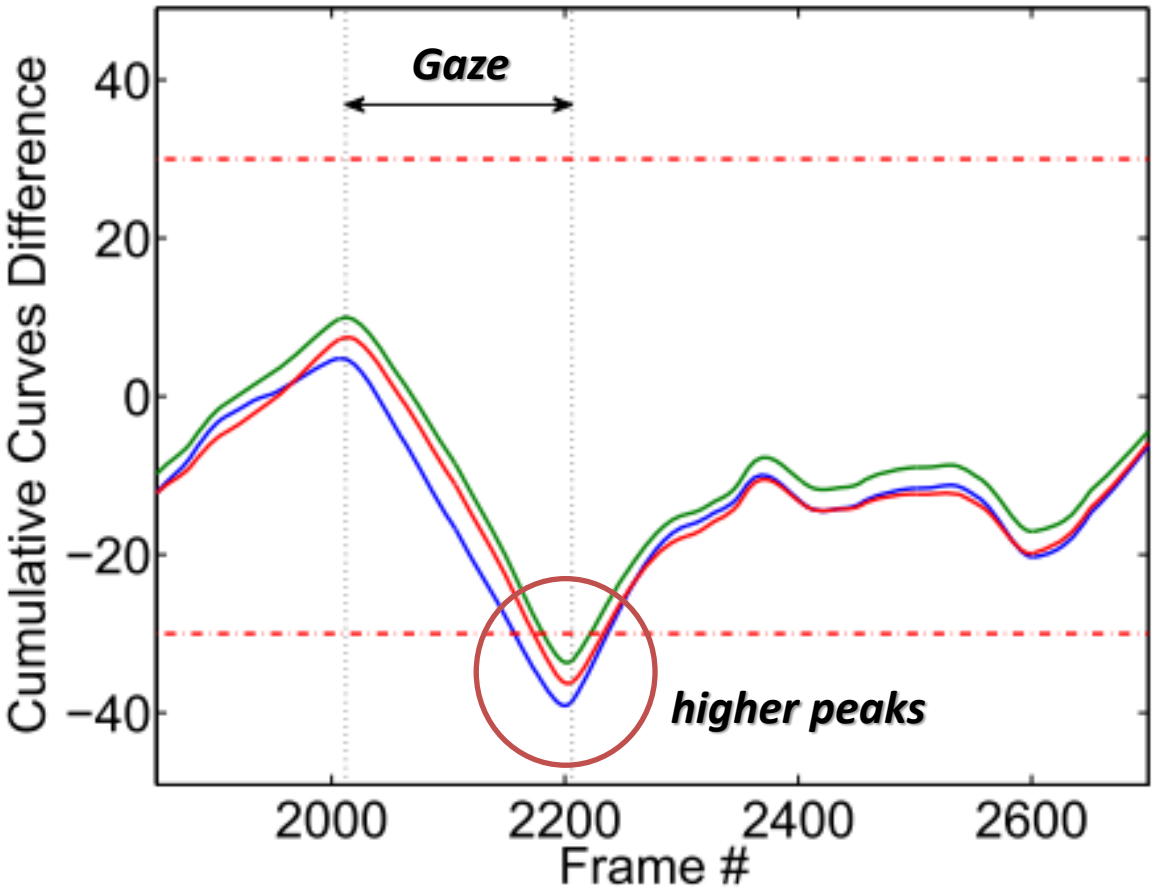
Original CD Curve

Smoothed CD Curve

Cumulative Difference

- Compute the **cumulative difference**





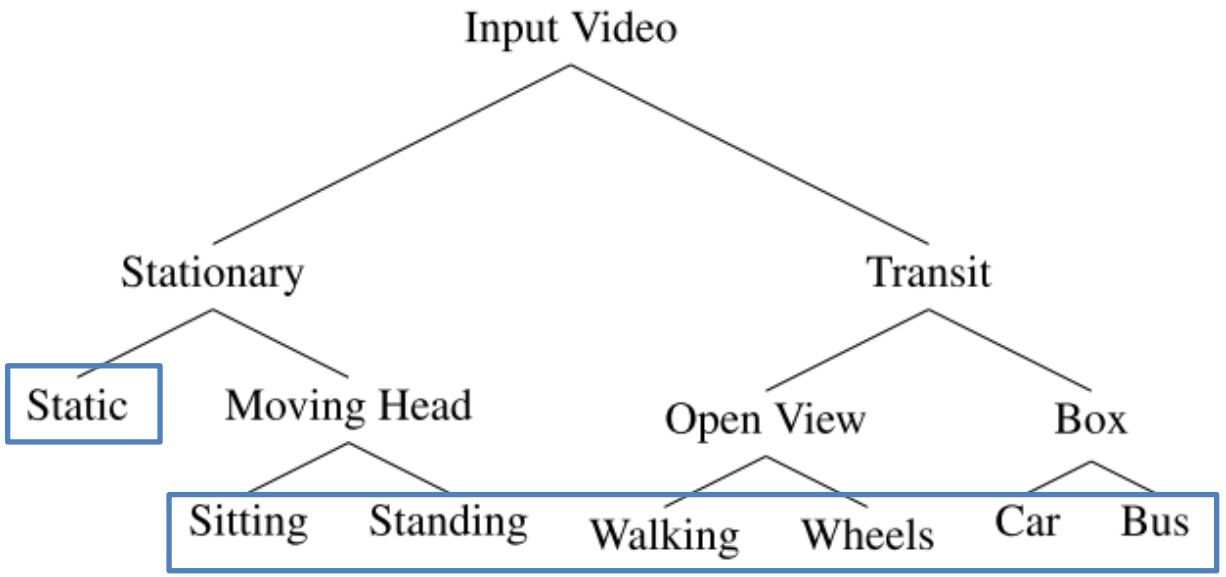
Gaze Hypothesis Threshold
> 80%

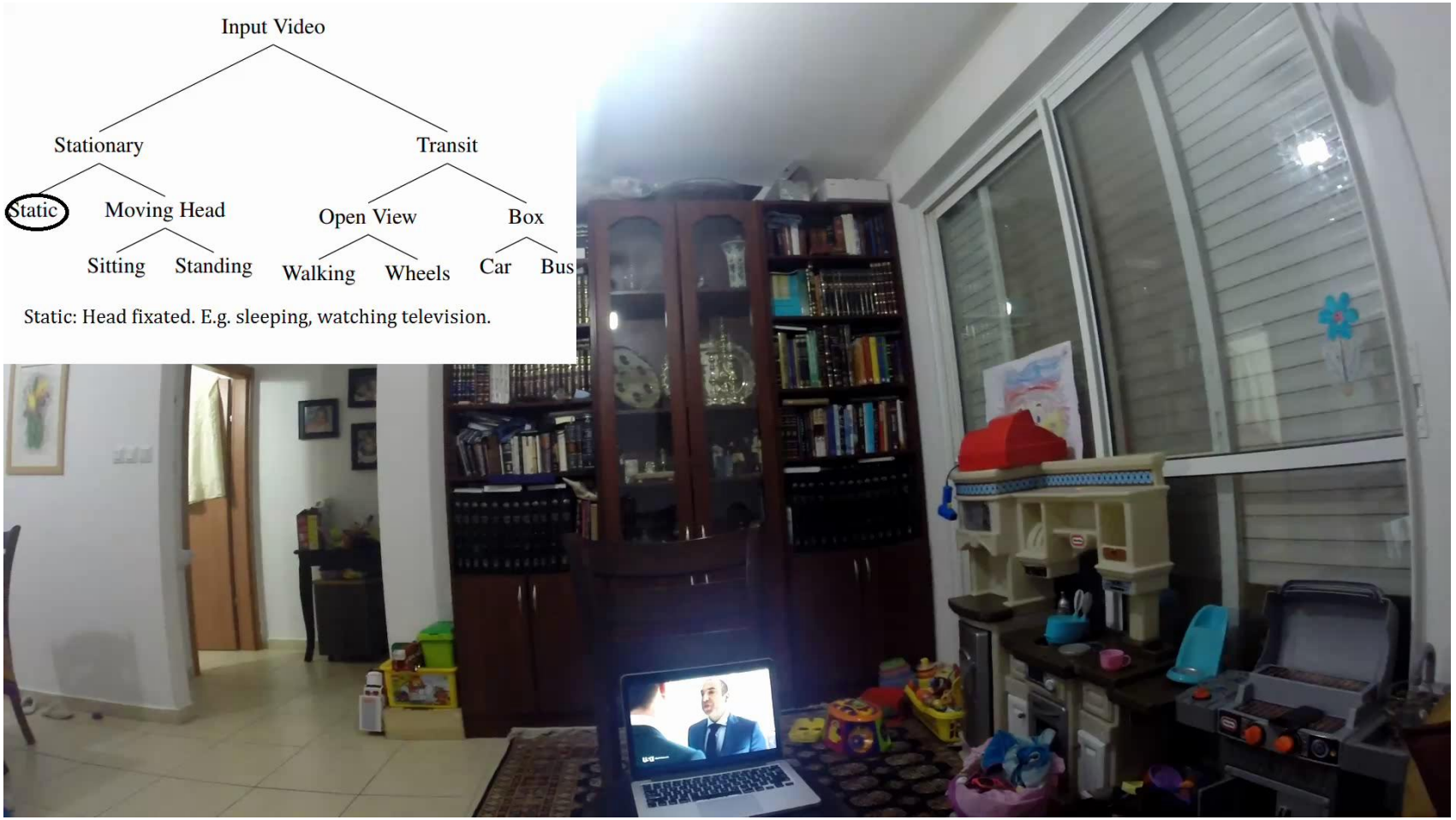
Motion Detector Threshold
> 1 standard deviation

Experiment

Dataset

- > 65 hours egocentric videos
- Manually annotated as one of the leaf classes
- Video





Video credit: HUJI EgoSeg Dataset

Classification of Wearer's Motion

leaf node accuracy

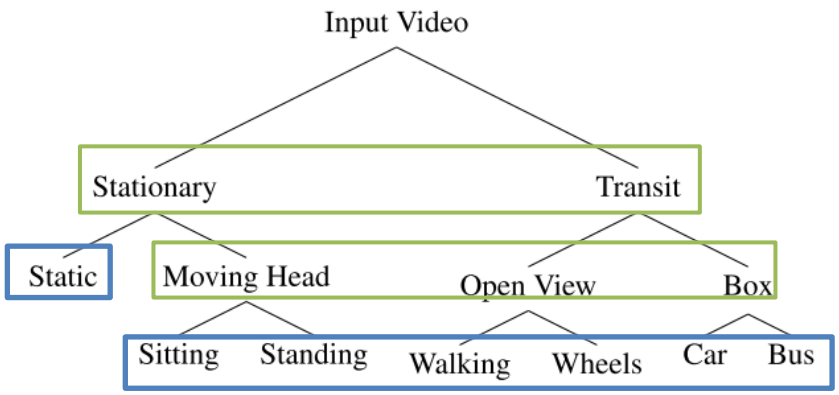
	Walking	Car	Standing	Bus	Wheels	Sitting	Static
Walking	**83%**	0%	6%	6%	4%	1%	0%
Car	1%	**74%**	3%	15%	0%	3%	4%
Standing	14%	2%	**47%**	4%	0%	31%	2%
Bus	3%	19%	27%	**43%**	0%	7%	1%
Wheels	9%	0%	0%	6%	**86%**	0%	0%
Sitting	3%	1%	22%	1%	0%	**62%**	10%
Static	0%	1%	1%	0%	0%	1%	**97%**

Average: 70%
Best: 97%

Sitting vs Standing
Bus vs Standing

inner node accuracy

Class Label	Accuracy	# Samples
Static-Moving	91%	1083115
Sitting-Standing	82%	1036217
Box-Open	87%	1197623
Car-Bus	76%	228108
Walking-Wheels	82%	969515



Detecting Period of Gaze Fixation

- Valid gaze fixation: a head fixation > 5 seconds

Seq.	Frames	# Fixation Detected	# True Positives	Accuracy
C1-C2	32017	47	39	82.97%
Y1-Y8	121208	219	163	74.43%
Total	153225	266	202	75.93%

Conclusion

Weakness

- Mixed features from adjacent activities
  - Short-term sitting when riding

	Walking	Car	Standing	Bus	Wheels	Sitting	Static
Walking	**83%**	0%	6%	6%	4%	1%	0%
Car	1%	**74%**	3%	15%	0%	3%	4%
Standing	14%	2%	**47%**	4%	0%	31%	2%
Bus	3%	19%	27%	**43%**	0%	7%	1%
Wheels	9%	0%	0%	6%	**86%**	0%	0%
Sitting	3%	1%	22%	1%	0%	**62%**	10%
Static	0%	1%	1%	0%	0%	1%	**97%**

Weakness

- Mixed activities

Waiting in line =
Standing + Walking

Riding an open train =
Open or Riding ?

Standing while coming into the station =
Static or Box ?

- Ambiguity in gaze fixation
  - A left and right turn in quick succession
  - A person turns in place

Strength

- Simple, efficient and robust
- Use only the recorded video
- Make no assumptions on the scene structure
- Focus on long-term activities to prevent over-segmentation of the video

Extension

- Use bilateral filter to find long-term trends
- Use a regularization framework like MRF on the classification results
- Handle the ambiguity in gaze fixation
- Combine with external sources such as GPS and inertial sensors
- Generalize to detect short-term activities
- Aid video summarization