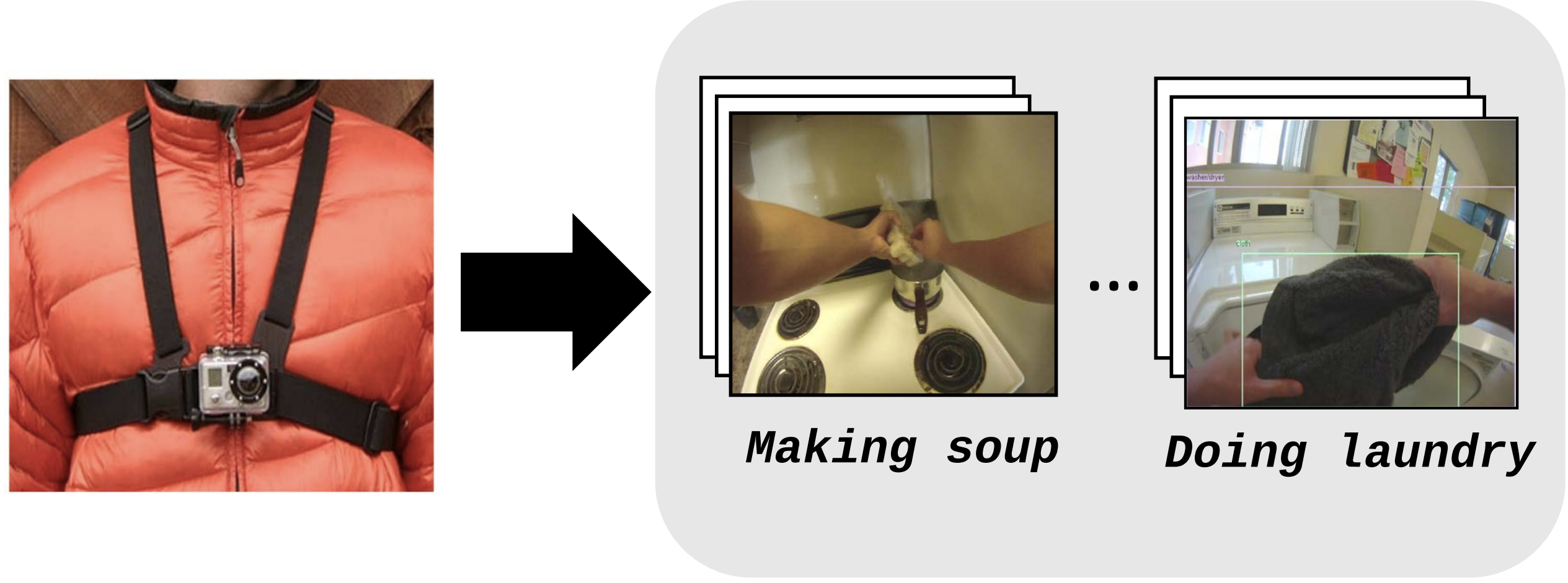
Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition

Tomas McCandless, Kristen Grauman
University of Texas at Austin

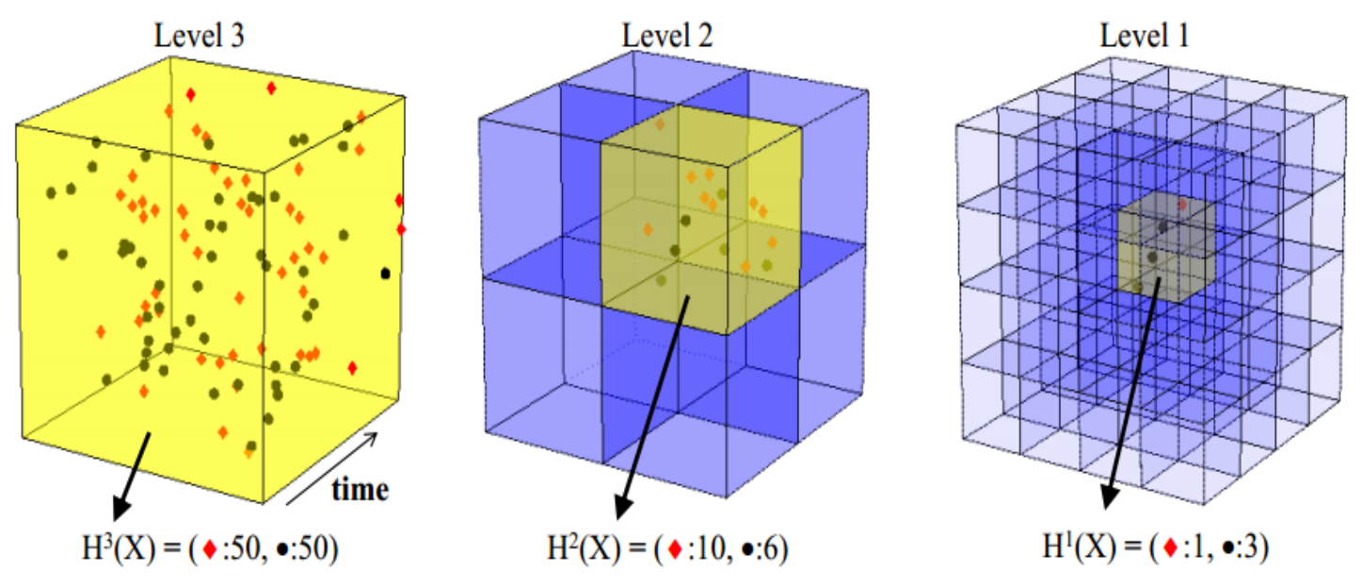
Goal

Recognize activities from first person point of view



Problem

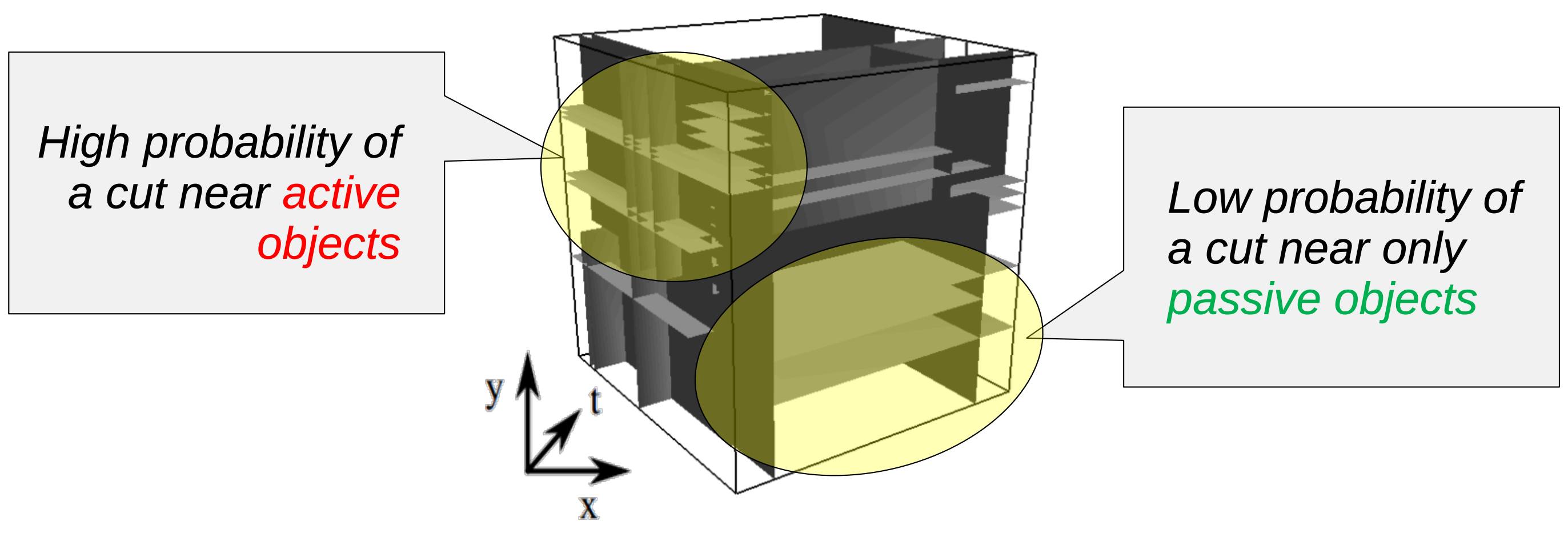
Histogram of space-time features is useful video representation [Choi et al. 08, Laptev et al. 08, Pirsiavash & Ramanan 12] …



…but hand-crafted (e.g., uniformly split) bin structures need not be most discriminative for target recognition task.

Main idea

- Bag-of-objects histogram pyramids to summarize ego-activity
- Boosting to learn discriminative spatio-temporal partitions
- "Object-centric" cutting scheme to focus pool of randomized partitions near active objects with which camera wearer interacts
- State-of-the-art results recognizing Activities of Daily Living

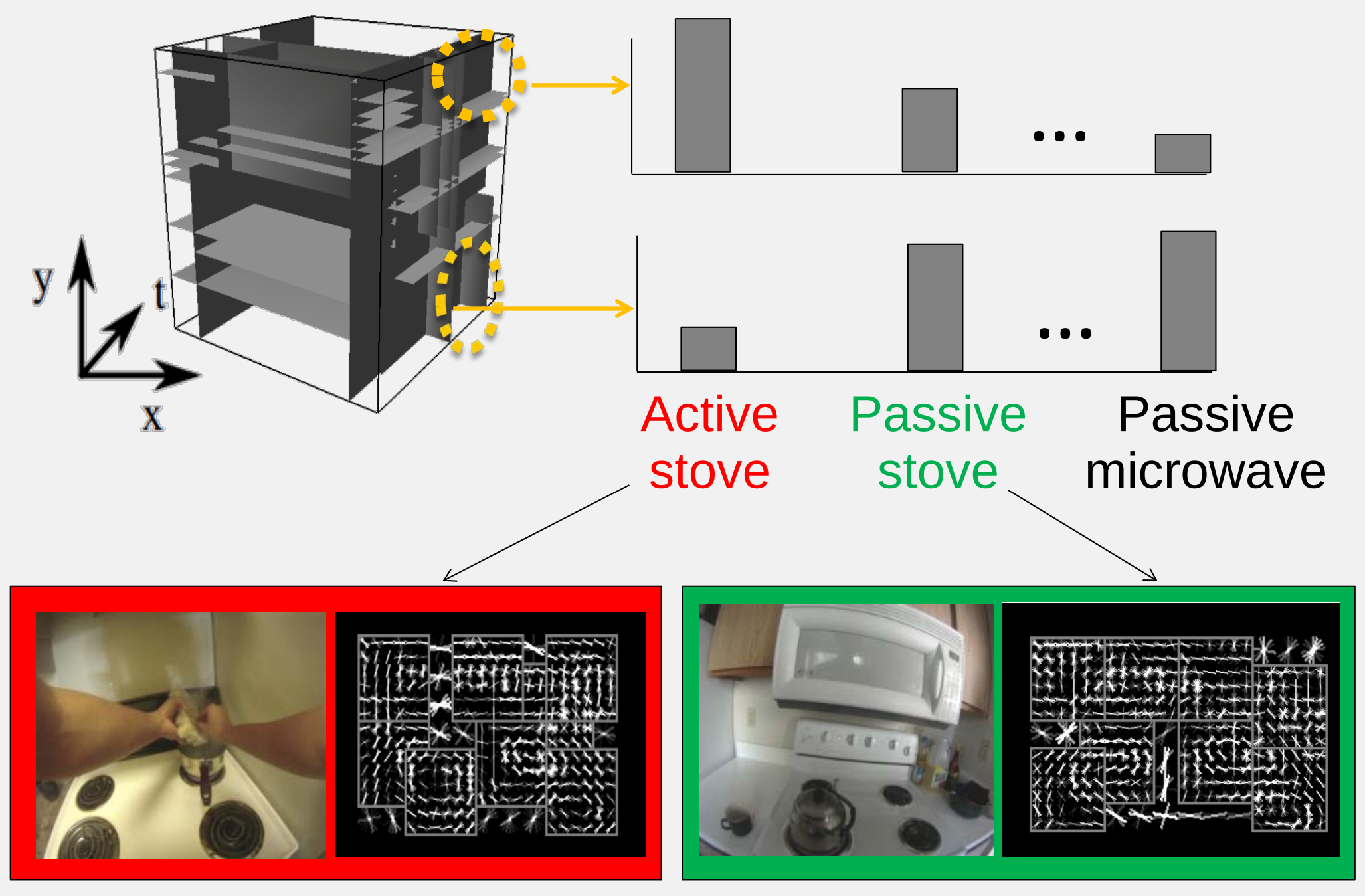

High probability of a cut near active objects
Low probability of a cut near only passive objects
y
t
x

One randomized candidate

Approach

Bag-of-objects

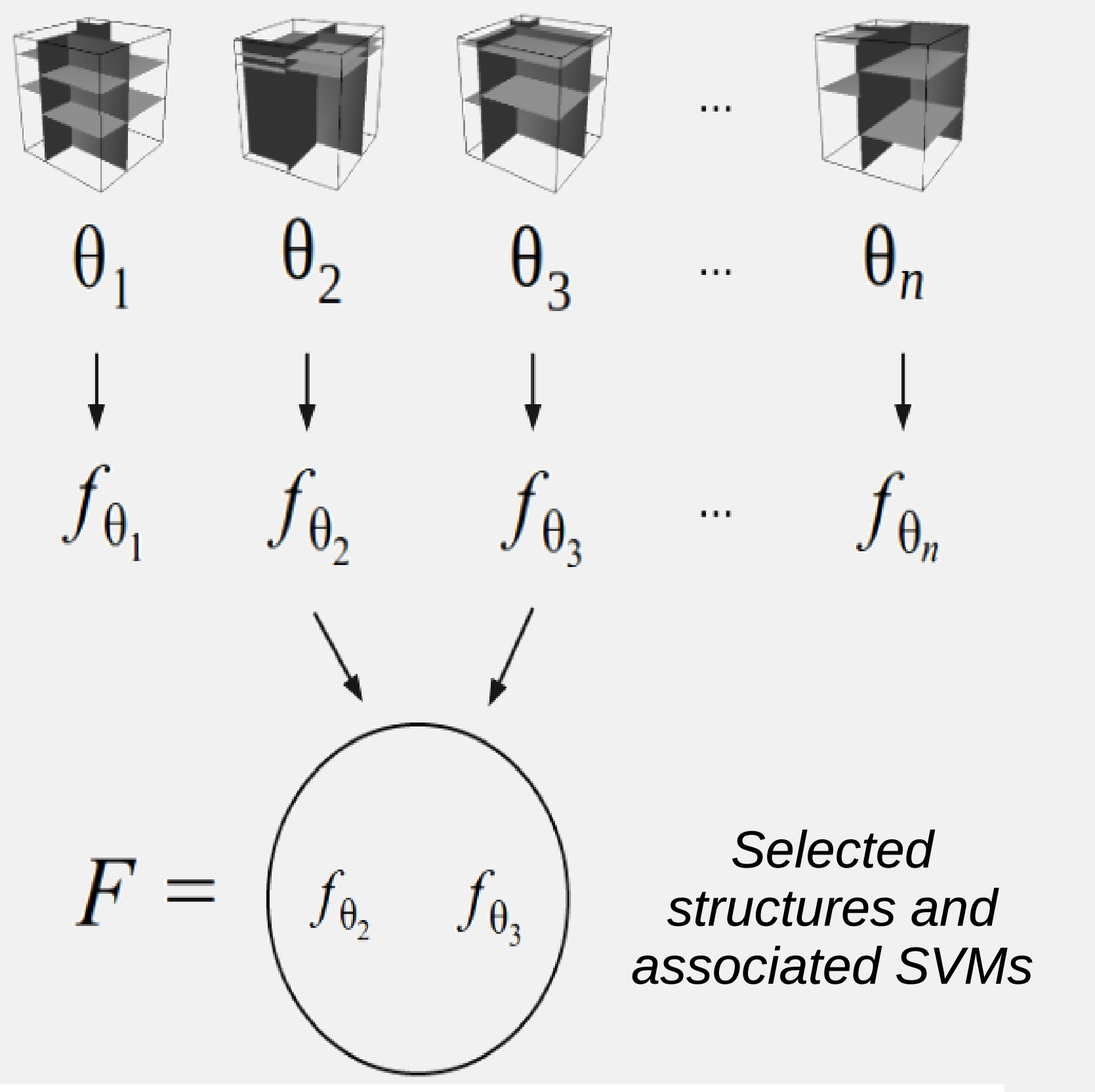
Histograms count detected object occurrences in series of space-time bins



Following Pirsiavash & Ramanan, we use separate detectors for active and passive versions of an object.

Boosting

Select discriminative combination of bin structures from randomized pool

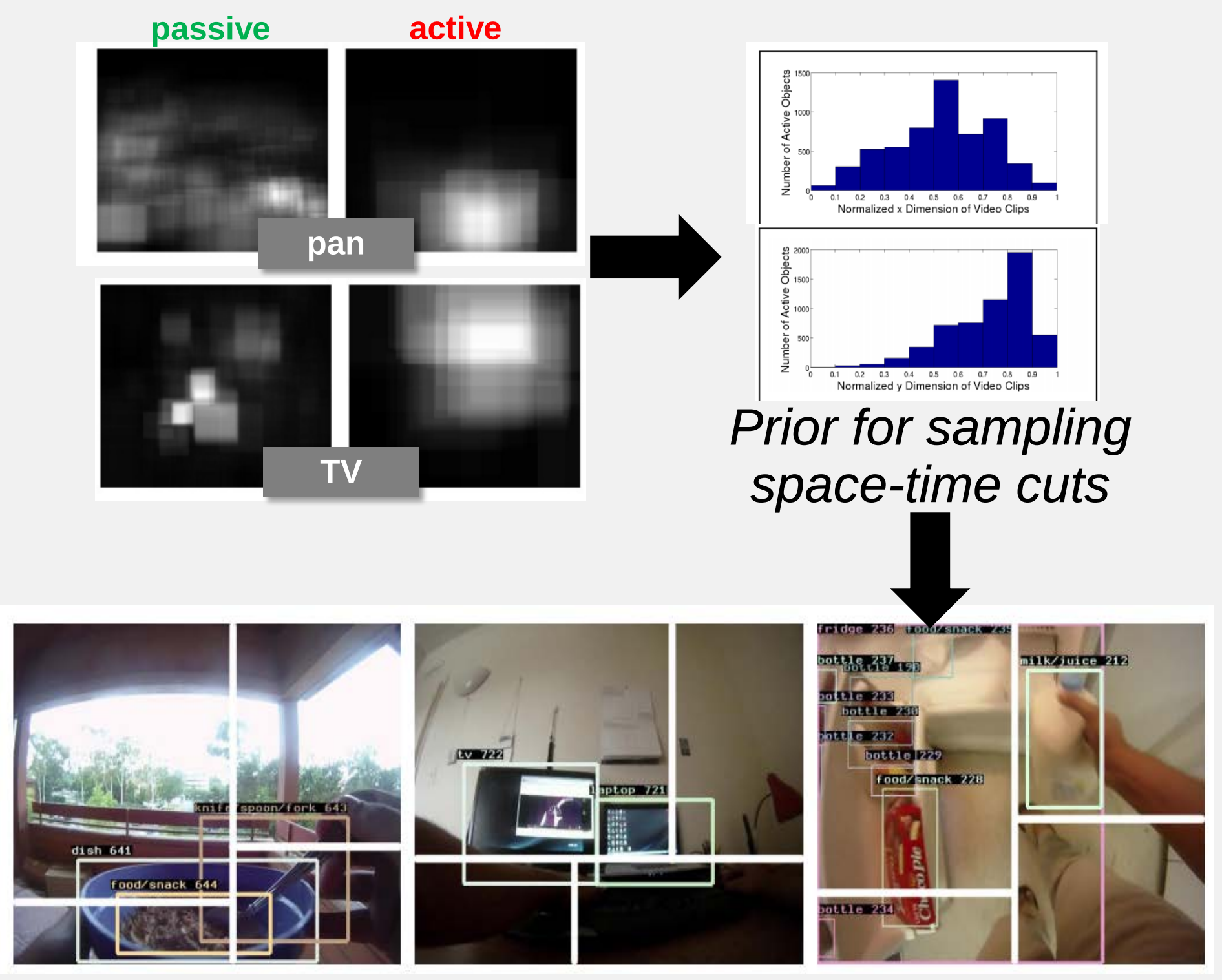


Selected structures and associated SVMs

$$F(V) = \mathrm{argmax}_c \Sigma_{m=1}^{j} \alpha_m \cdot \mathbf{I}(f_{\theta_m}(V) = c)$$

Object-centric cuts (OCC)

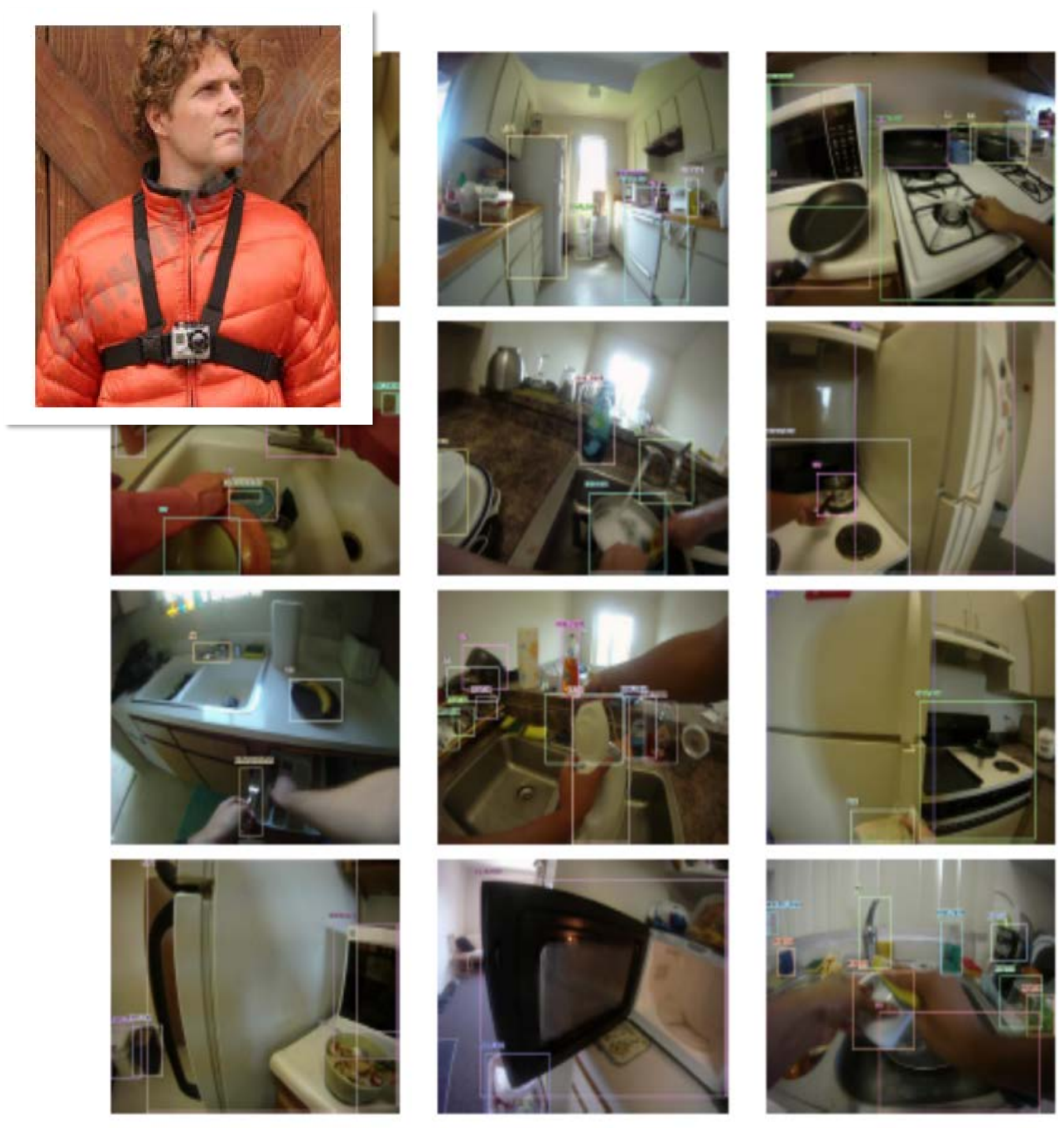
Focus sampling of bins where "active" objects are concentrated



Prior for sampling space-time cuts

Emphasize video regions likely to characterize key interactions → Control pool size for boosting

Results



Activities of Daily Living (ADL)
[Pirsiavash & Ramanan, 2012]

18 actions ~ food, hygiene, entertainment (wash hands, make tea, brush teeth, etc.)

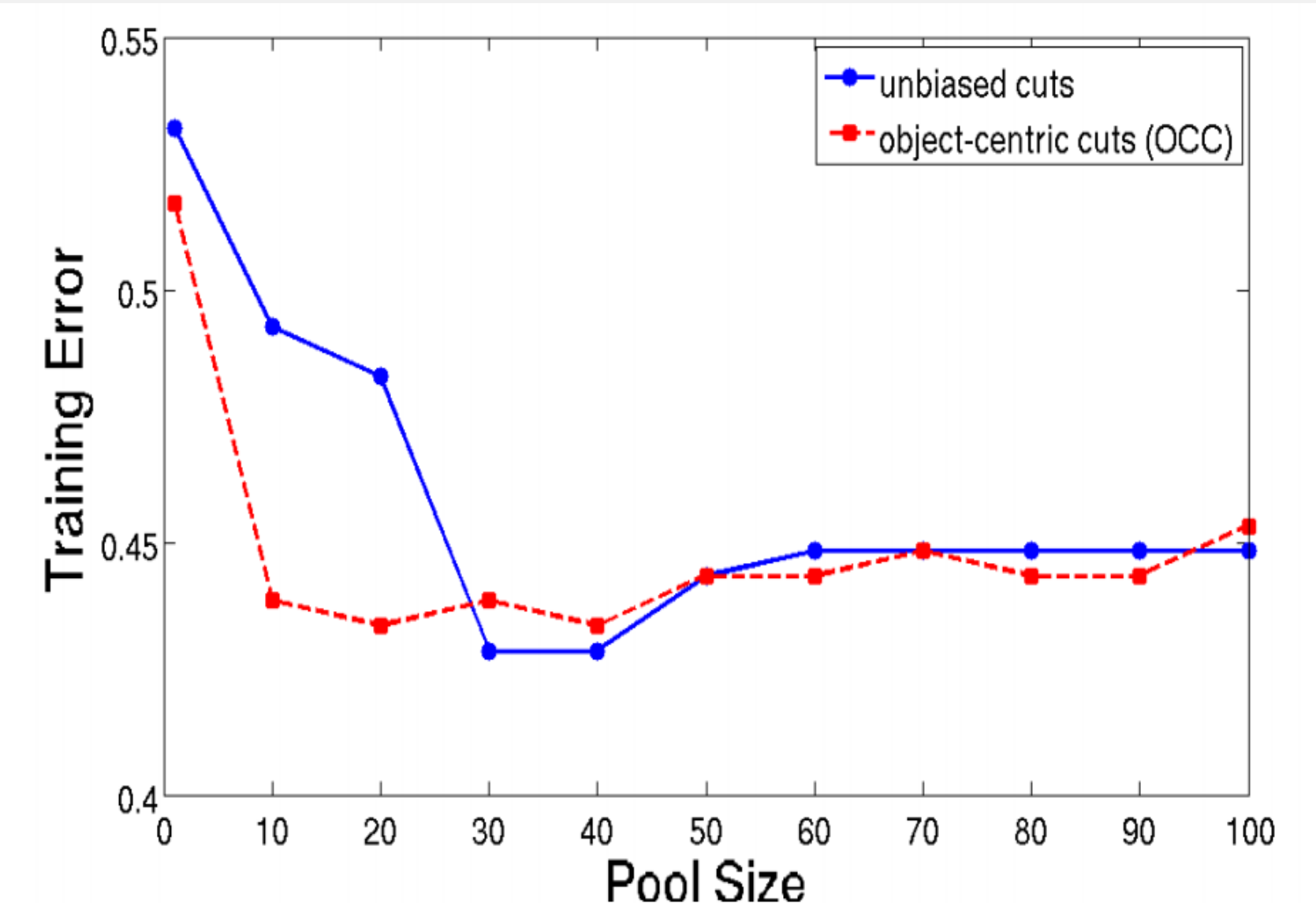
20 people, 10 hours of video

We improve the state-of-the-art accuracy on this challenging dataset.

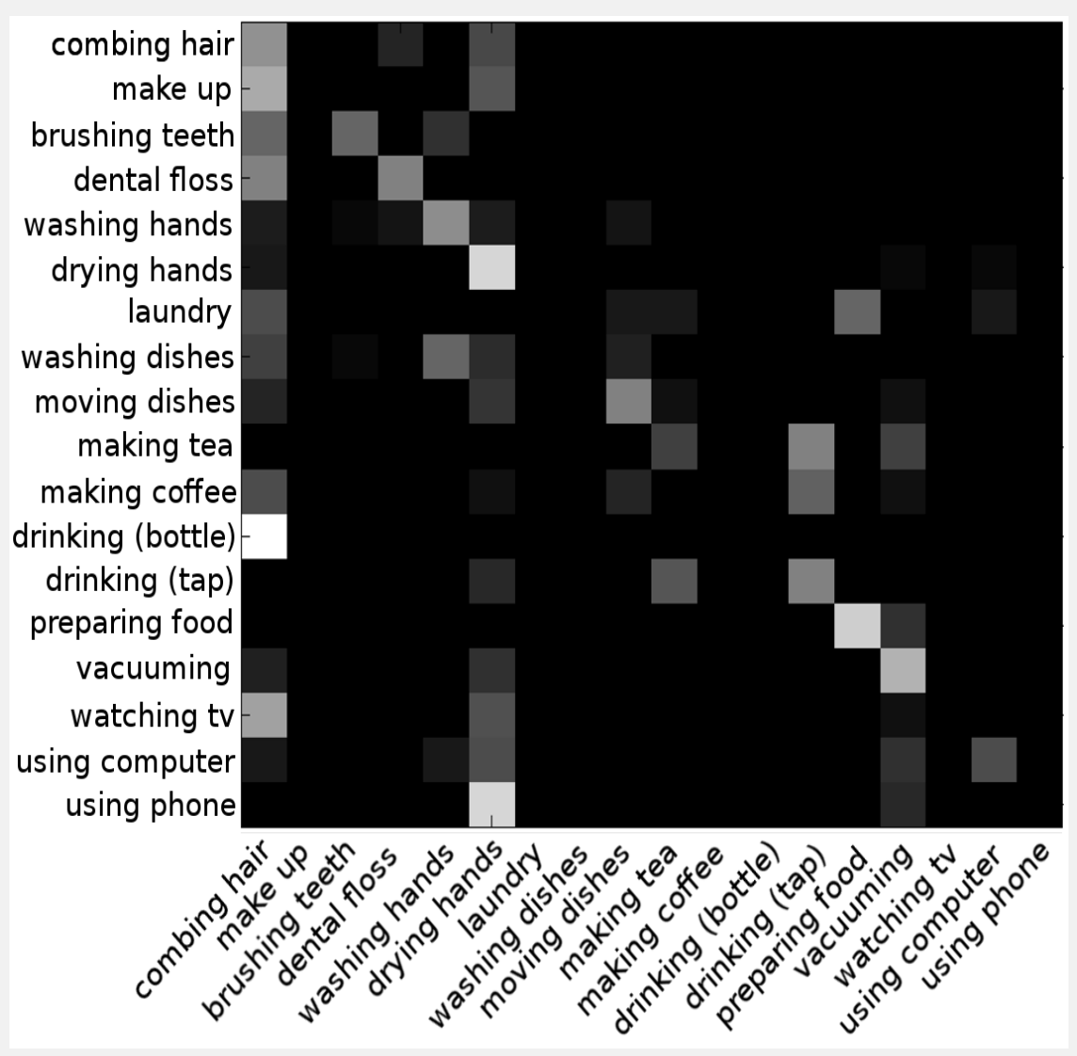
BoW	Bag-of-objects	TempPyr [21]	Boost-RSTP	Boost-RSTP+OCC (ours)
16.5%	34.9%	36.9%	33.7%	**38.7%**

Methods compared:

•Bag-of-words (BoW): space-time interest points and HoG/HoF visual words
•Bag-of-objects: global histogram of detected objects
•Temporal Pyramid: hand-crafted, one cut in time [Pirsiavash & Ramanan, CVPR12]
•Boost-RSTP: randomized spatio-temporal pyramids *without* object-centric cuts



Object-centric cuts achieve lower error with smaller pool of candidates → More efficient training for boosting.



Best accuracy: actions with regular space-time structure (e.g., comb hair, dry hands)

Most confusions: same active objects involved (e.g., making tea vs. making coffee)