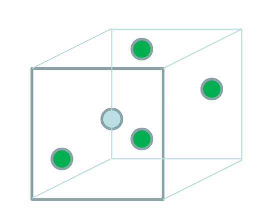
Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition

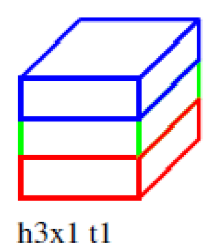
Adriana Kovashka and Kristen Grauman
University of Texas at Austin

Problem

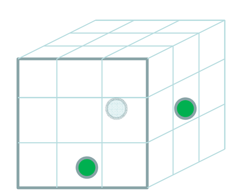
- Individually, spatio-temporal features may be “too local” for action recognition.
- How to describe relative spatio-temporal information flexibly?
- Existing neighborhoods / approaches:



global (bag of words)



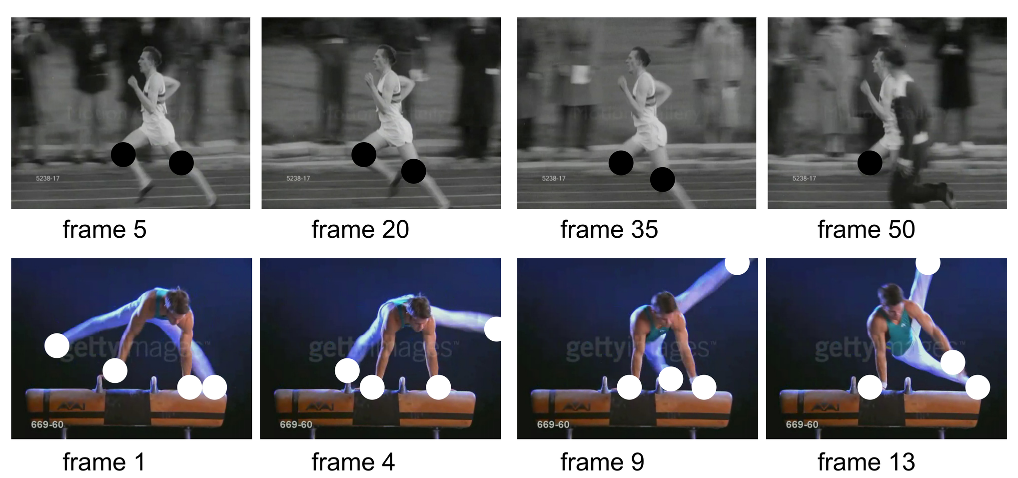
clip bins [Laptev et al. 08, Choi et al. 08, Sun et al. 09]



grids per feature [Gilbert et al. 09]

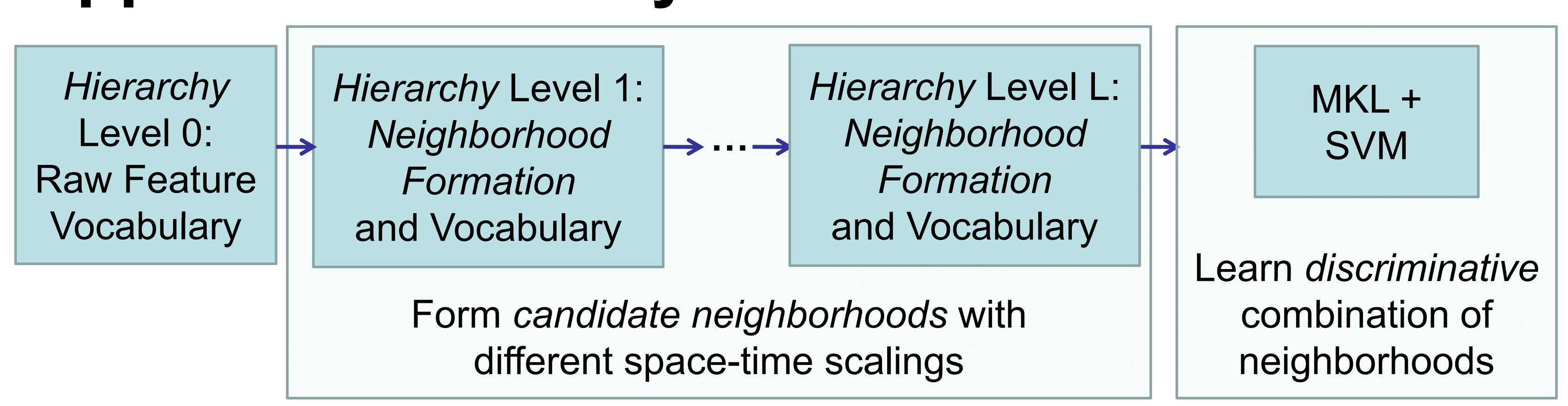
- Fixed-sized grids disregard that action classes have different sparsity of features, and clip-level binning is sensitive to segmentation.

Our Idea

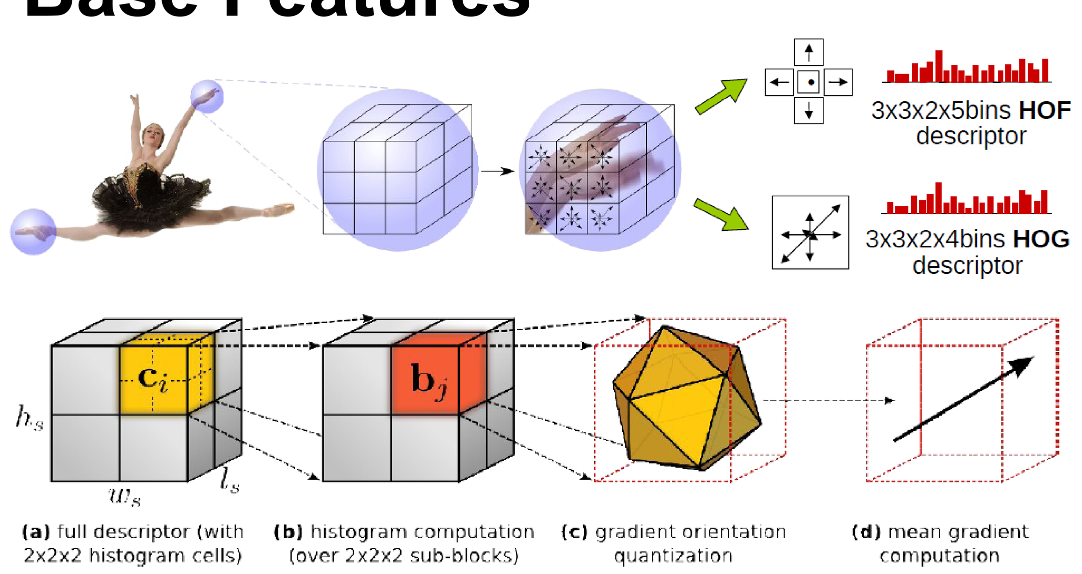


- Compute a visual vocabulary preserving *relative* spatio-temporal relationships.
- Form *variable-shaped* neighborhoods of interest points.
- *Learn* a hierarchy of discriminative neighborhoods for different action classes.

Approach Summary



Base Features

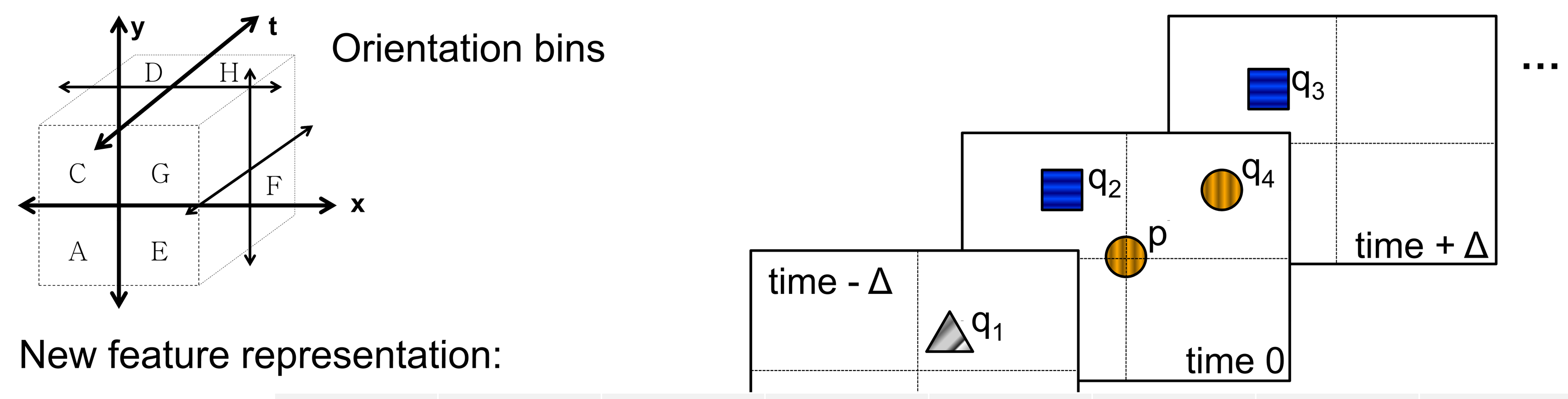


HoF/HoG features [Laptev et al., CVPR 2008] for sparse interest points [image: Wang et al. 09]

HoG3D features [Kläser et al., BMVC 2008] for dense interest points

Neighborhood Formation

- Form neighborhoods of interest points around each point as a center.
- For each of the *N* nearest neighbors, record its orientation with respect to the central point and its level-0 visual word (computed on the raw interest point level).



New feature representation:

counts for	bin A			bin B			bin C			bin D			bin E			bin F			bin G			bin H		
	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲
neighbors 0 to 0																						1		
neighbors 0 to 1																					1	1		
neighbors 0 to 2											1										1	1		
neighbors 0 to 3											2										1	1		
neighbors 0 to 4											2										1	2		

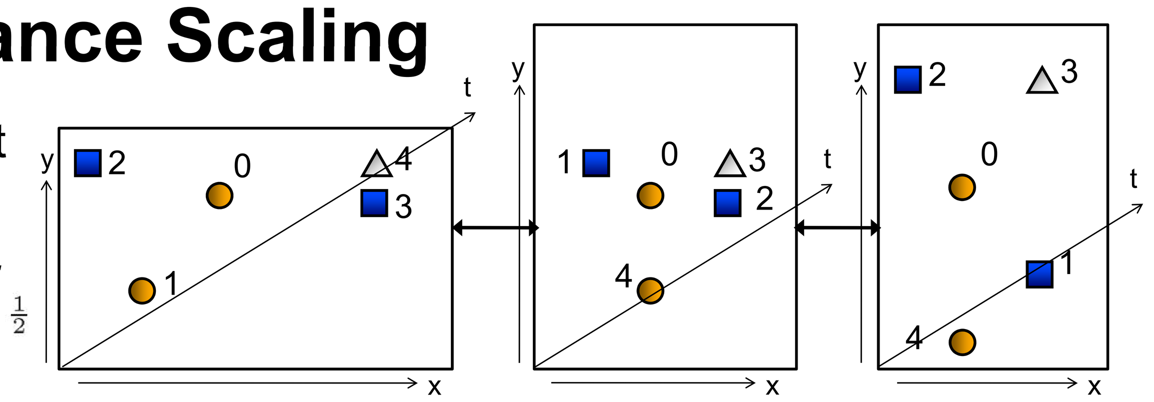
cumulative

- Reshape each histogram into a vector to obtain next level’s feature representation.
- Quantize new representations of all points to form next level’s vocabulary **H.**

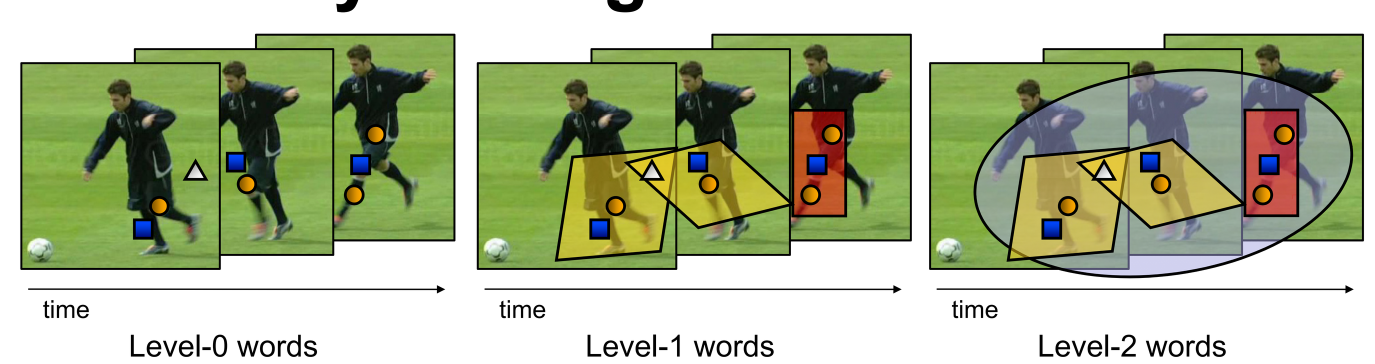
Space-Time Distance Scaling

- One pixel != one frame, must consider neighborhoods for different scalings of *x, y, t.*

$$D_\sigma(p,q) = \left(\sum_{i=1}^{3} \frac{1}{\sigma_i} (p(i) - q(i))^2 \right)^{\frac{1}{2}}$$



Hierarchy of Neighborhood Words



- Repeat neighborhood formation to generate higher-level vocabularies.

$$\mathcal{H}_\sigma(V) = \{H_0(V), H_1(V), \ldots, H_L(V)\}$$

Discriminative Space-Time Neighborhoods

- **C** = F(ML+1) χ^2 kernels (F feature types, M distance scalings, L levels).
- Given these kernels, use Multiple Kernel Learning (MKL) to learn the most discriminative combinations.

$$K(H_i, H_j) = \sum_{c \in \mathbf{C}} w_c \exp\left(-\frac{1}{A_c} \chi^2(H_i^c, H_j^c) \right)$$

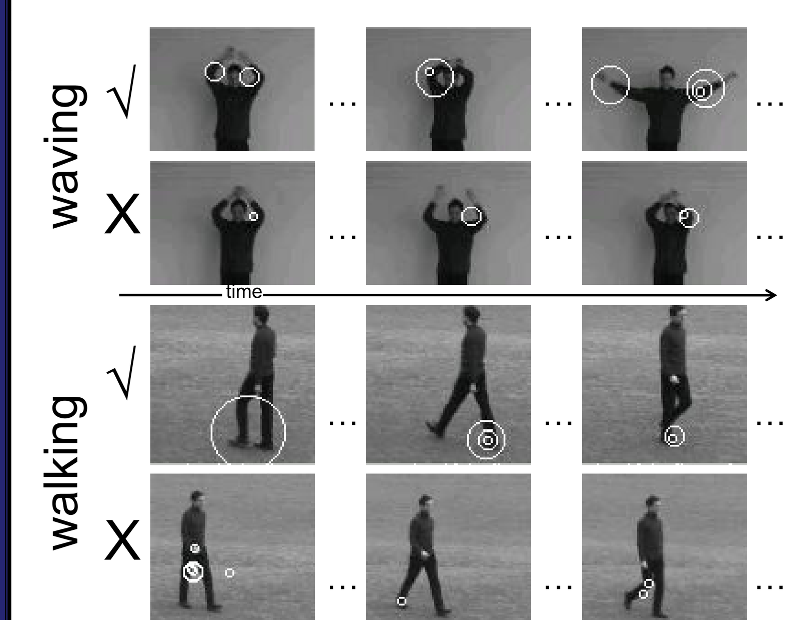
Results

KTH Dataset

[Schuldt et al., ICPR 2004]
6 classes, 600 videos

Approach	Accuracy
Laptev et al. 2008	91.80%
Gilbert et al. 2009	94.50%
Our method	94.53%

Accuracy equal to best known for KTH.



Examples of neighborhoods learned to be *useful / not useful.*

UCF Sports Dataset

[Rodriguez et al., CVPR 2008]
10 classes, 150 videos + 150 flipped
Leave-one-out, flip of test *not* in train

Approach	Accuracy/Class
Our method	87.27%
Average of all kernels	84.43%
Level-0 baseline	85.49%

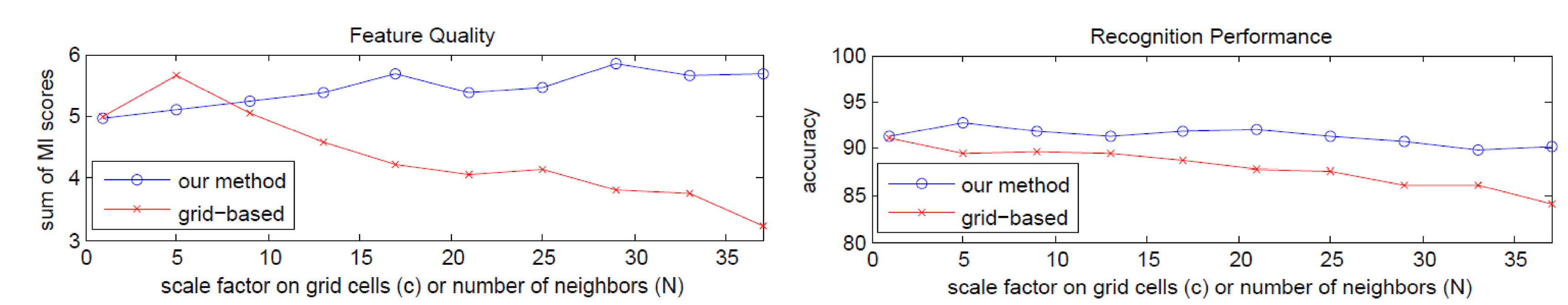
State-of-the-art results for UCF Sports.

Levels	Accuracy/Class	Average MKL Weight
0	85.49%	0.63 ($\pm$ 0.3)
1	82.16%	0.10 ($\pm$ 0.2)
2	73.30%	0.10 ($\pm$ 0.2)

All vocabulary levels for one feature distance contribute to the accuracy on UCF Sports.

Implementation details: *k* = 300 (4k for UCF level-0); *N* = 5; *L* = 2

Fixed-Size vs Variable-Shaped Neighborhoods



Feature distinctiveness and recognition accuracy of our level-1 neighborhood words (one distance scaling) less sensitive to neighborhood size parameter than grid-based baseline.

Conclusions

- Hierarchies capture feature relationships at multiple granularities.
- Showed importance of translation-invariant and discriminative variable-shaped neighborhoods.