



Egocentric Recognition of Handled Objects: Benchmark and Analysis

Xiaofeng Ren, Matthai Philipose
Intel Labs Seattle

@ CVPR 2009 Workshop on Egocentric Vision

June 20th, 2009

Everyday Sensing and Perception



- **ESP**: Megabets at Intel Labs
- 12-researcher, 3-year Push integrating learning, sensing, HCI, distributed computing
- The **90/90** Challenge:
To build a real-time context-recognition system that is 90% accurate over 90% of your day.



Context Awareness is Here Today...



Wii /
Guitar Hero



iPhone



Android



Nike+

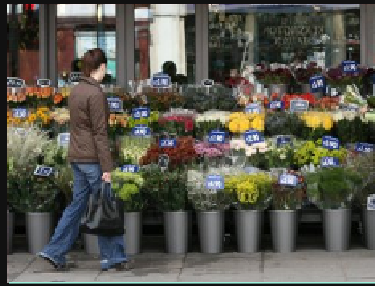


Nokia Sport

location, activity (walking, running, sitting), carbon footprint, gestures, task grading, pace, distance...

...but there is more to see and more to do

Context Awareness → Life Assistance



stay healthy

stay in touch

keep learning

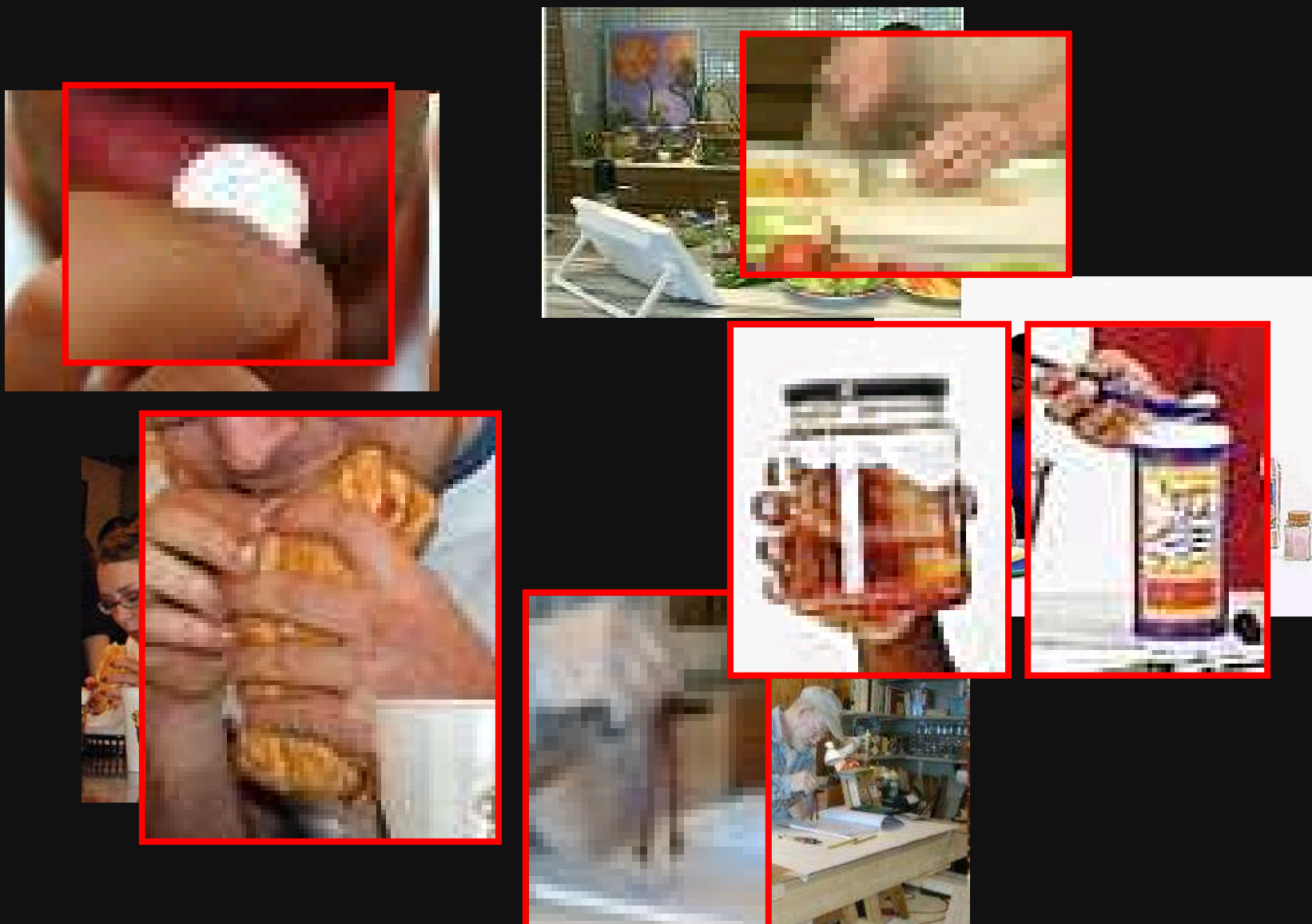
Context-aware devices will be constantly useful in achieving key life goals

Prior Work on Context and Wearable Sensors



- [Aloimonos, Weiss, Bandyopadhyay 1988] Active Vision
- [Starner, Schiele, Pentland 1998] Visual context awareness using wearable cameras
- [Schiele, Oliver, Jebara, Pentland 1999] Dynamic personal enhanced reality
- [Pentland, Choudhury 2000] Face recognition in smart spaces
- [Mayol, Tardoff, Murray 2000] Wearable visual robots
- [Clarkson, Mase, Pentland 2000] User context using wearable cameras
- [Starner, Weaver, Pentland 2000] Real-time american sign language recognition
- [Lee and Mase 2000] Activity and location recognition using wearable sensors
- [Kern, Schiele, Schmidt 2003] Multi-sensor activity context detection
- [Philipose, Fishkin, Perkowitz 2005] Inferring activities from objects using RFID
- [Mayol and Murray, Wearable Computers 2005] Wearable hand activity recognition
- [Intille, Larson, Tapia, 2006] Live-in laboratory for ubiquitous computing
- [Davison, Reid, Molton, Stasse 2007] MonoSLAM
- [Kerler, Galleguillos, Belongie 2007] Recognizing groceries in situ
- [Wu, Osuntogun, Choudhury, Philipose, Rehg 2007] Activity based on object use
-

Handled Objects as Context



Recognizing Handled Objects



- Scalable activity recognition based on object use
[Wu, Osuntogun, Choudhury, Philipose, Rehg, ICCV 2007]
 - 33 objects in 16 activities, 3 videos, overhead camera, good lighting
 - 74% accuracy using bags of SIFT words and background subtraction



Large-scale, real-world empirical studies of handled-object recognition using wearable video cameras



- Wearable hand activity recognition
[Mayol and Murray, Wearable Computers 2005]
Identifying objects-in-hand from a wearable camera
5 objects, 600 frames, using color histograms

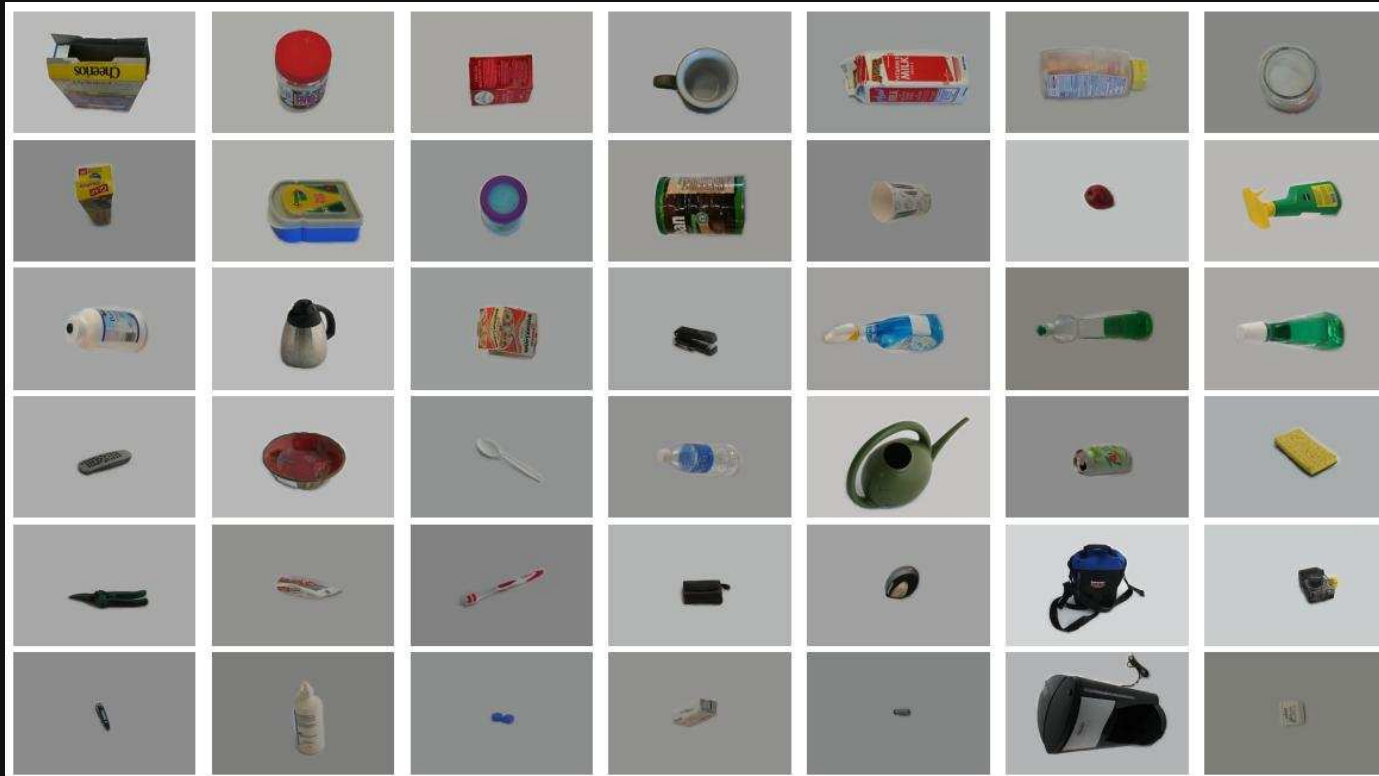
Wearable vs Environmental Cameras



- Low resolution
- Poor optics/electronics
- Limited Field-of-View
- Motion Blur

- It's **Egocentric!!**
- No need to spam the world with cameras
- Egocentric viewpoint
see the world as we see it, foveated,
better views, less occlusion, canonical
positions and scales, ...

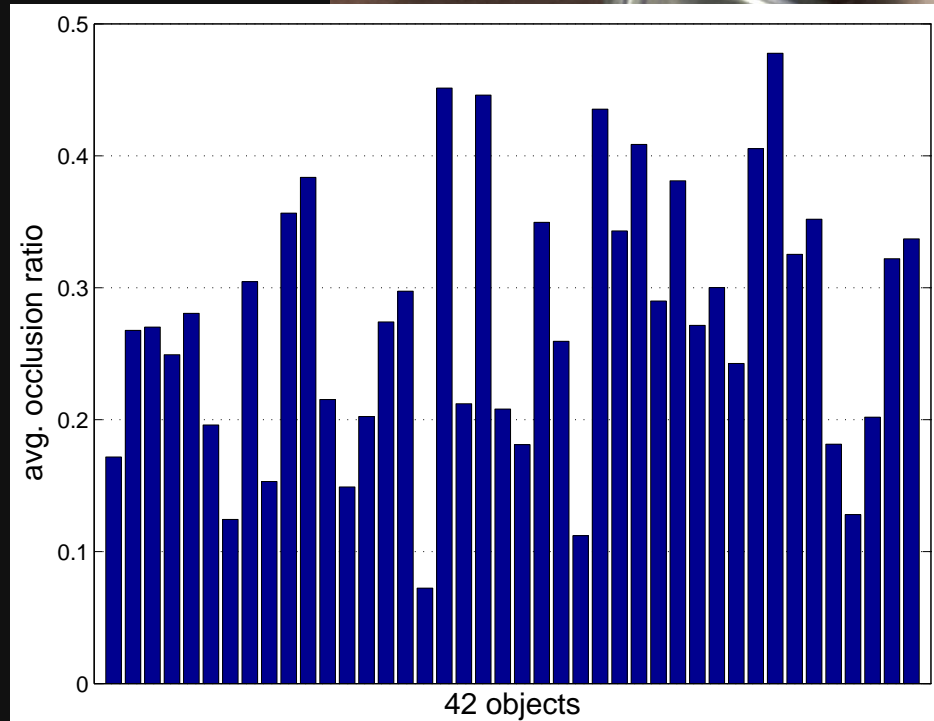
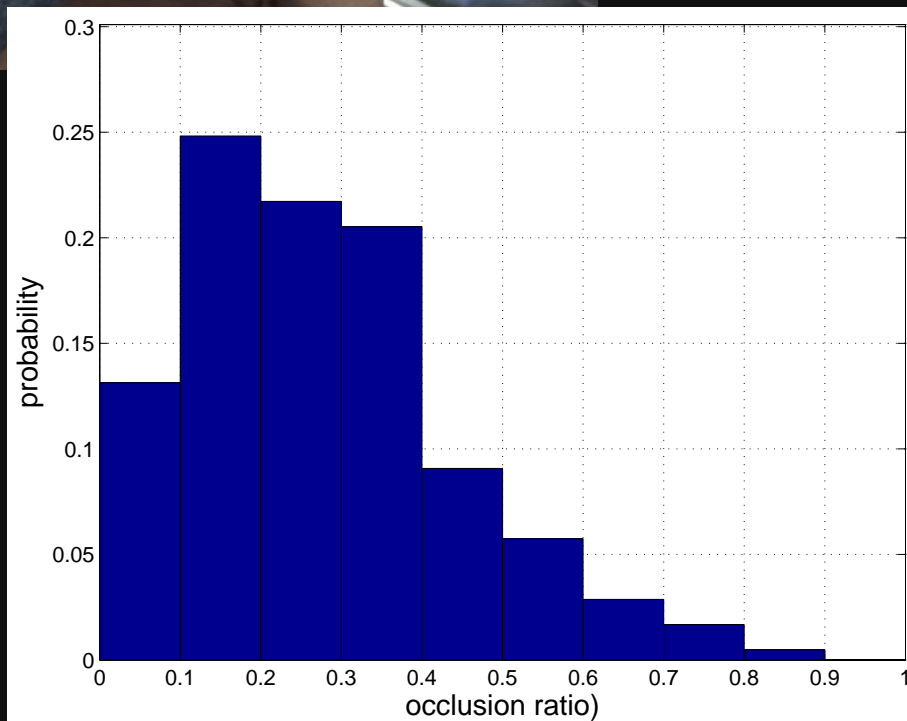
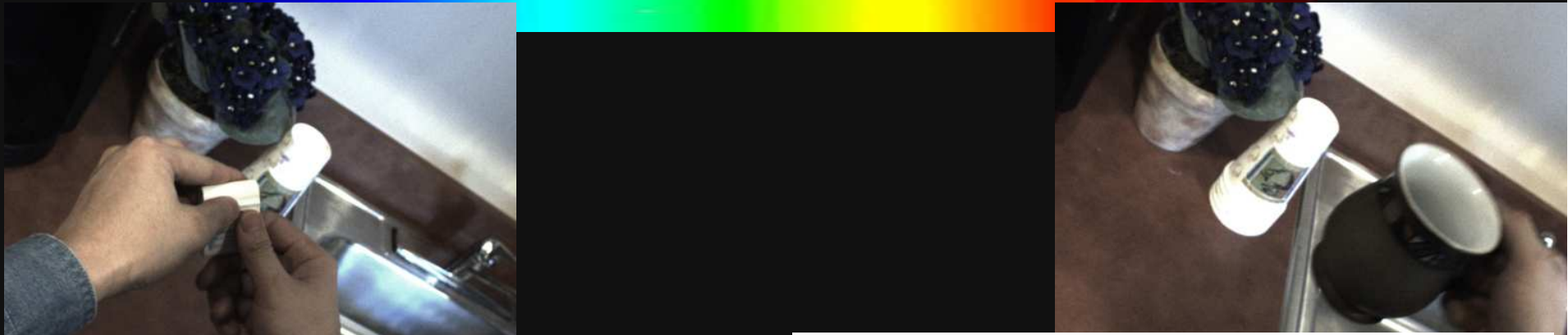
Dataset for Egocentric Object Recognition



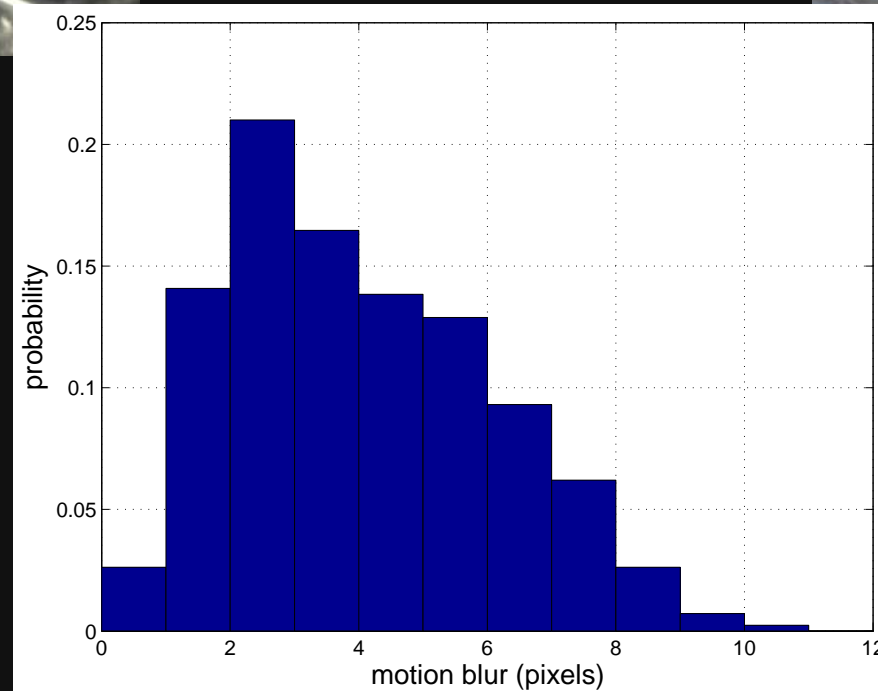
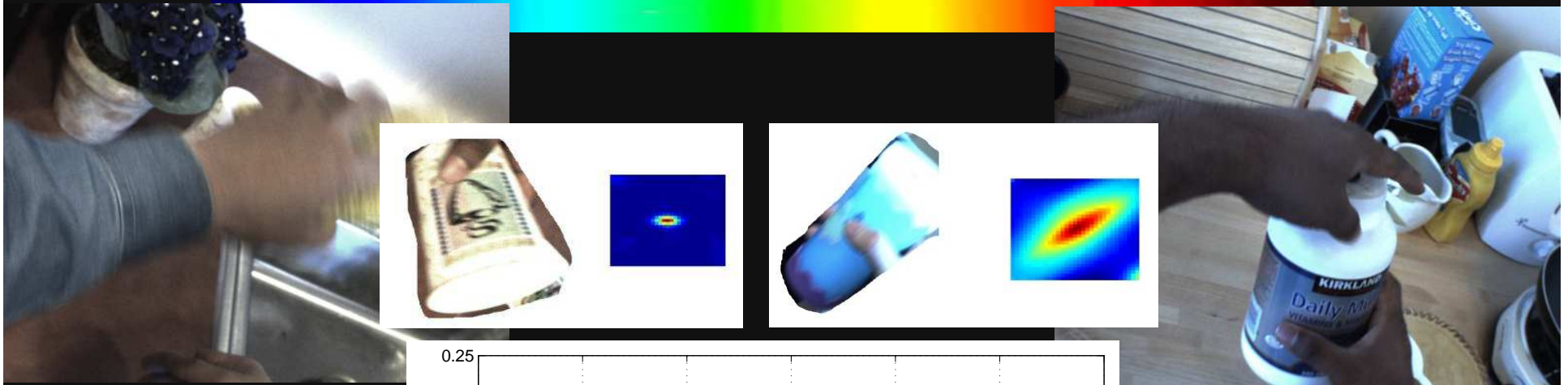
Dataset for Egocentric Object Recognition

(video)

Challenge: Occlusion



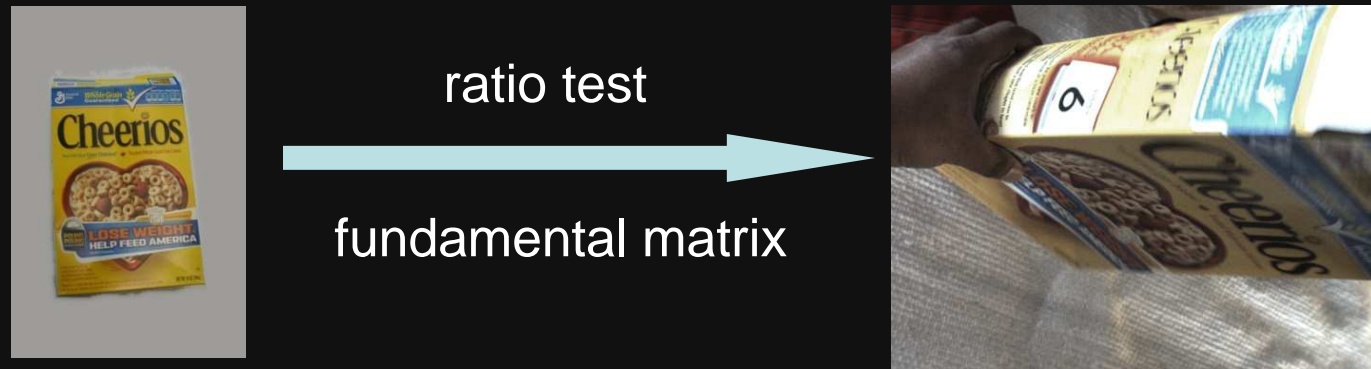
Challenge: Motion Blur



Baseline Approach

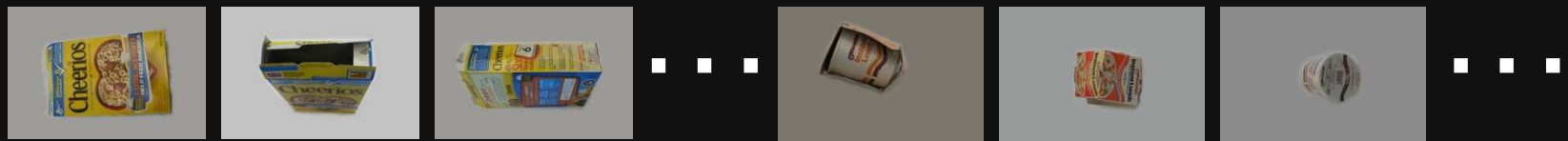


- SIFT matching



Similarity(clean exemplar, video frame)

- SVM on the similarity vector

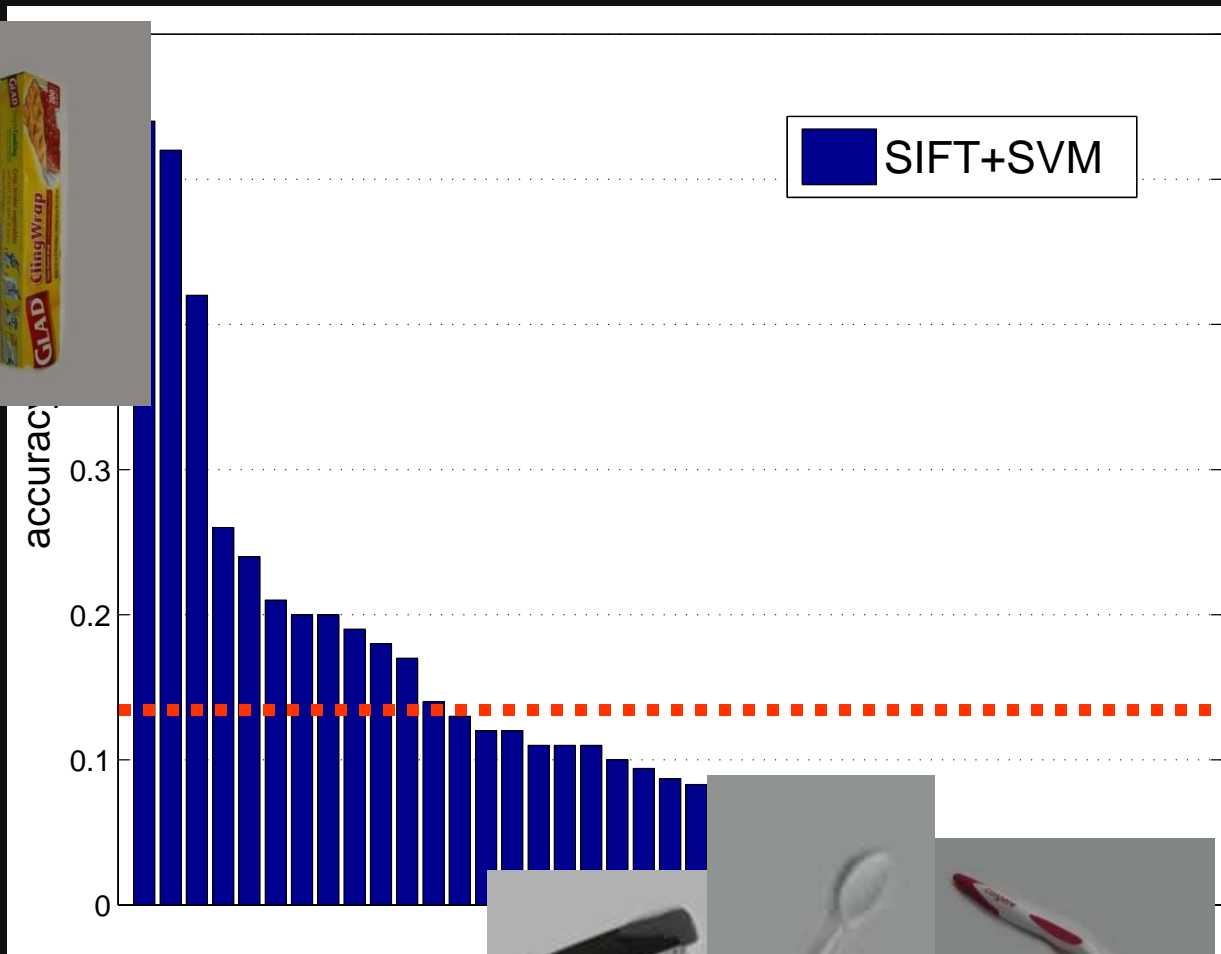


$[S_1, S_2, S_3, \dots, S_k, S_{k+1}, S_{k+2}, \dots \dots]$

Baseline Results



Accuracy: 12.0%



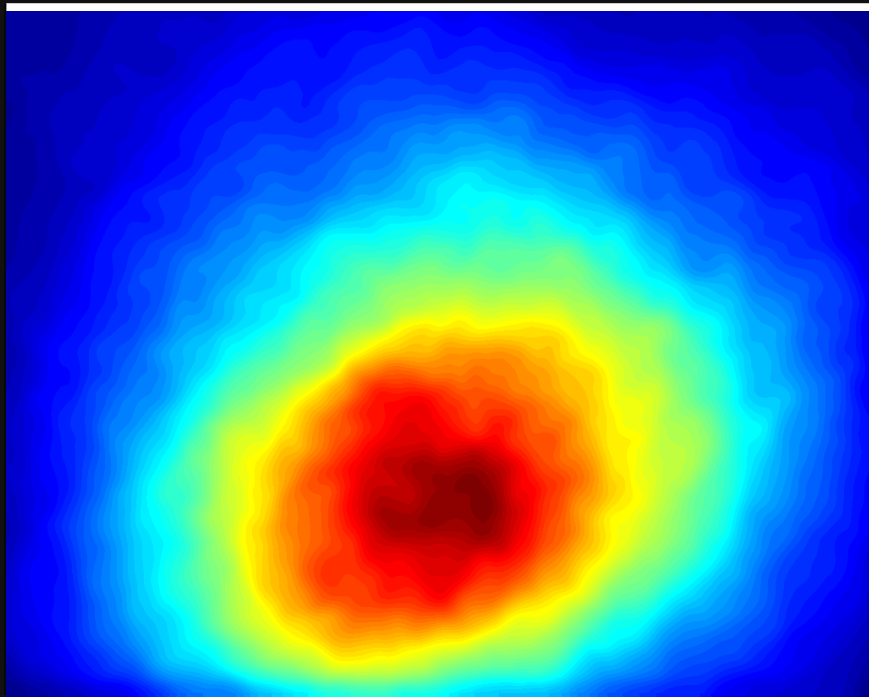
Further Evaluations



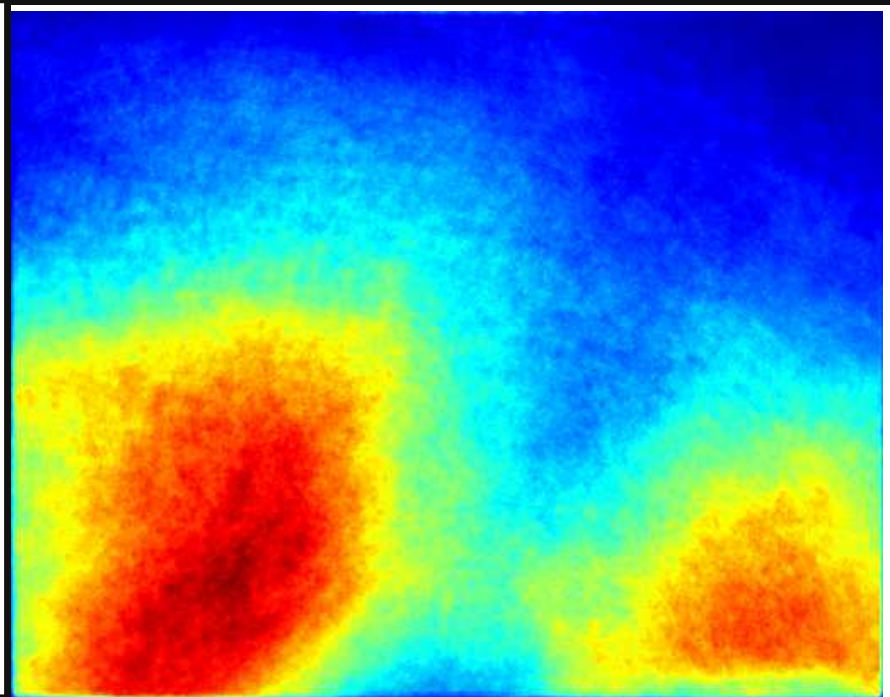
- [Grauman and Darrel ICCV 2005]
 - LIBPMK [Lee et al 2008]
 - Hierarchical clustering + Pyramid matching kernel
 - 1/5 of the data (~200 training per object), 42 objects
 - Average accuracy: **11%** vs **12%** baseline, 42 objects

- [Felzenszwalb, McAllester, Ramanan CVPR 2008]
 - HOG templates + Deformable parts
 - Manual bounding box input, search over orientation
 - 1/5 of the data, 11 objects
 - Average accuracy: 54% vs 31% baseline, 11 objects
 - 21% ??** vs **12%** baseline, 42 objects

Opportunity: Location Prior



Object location

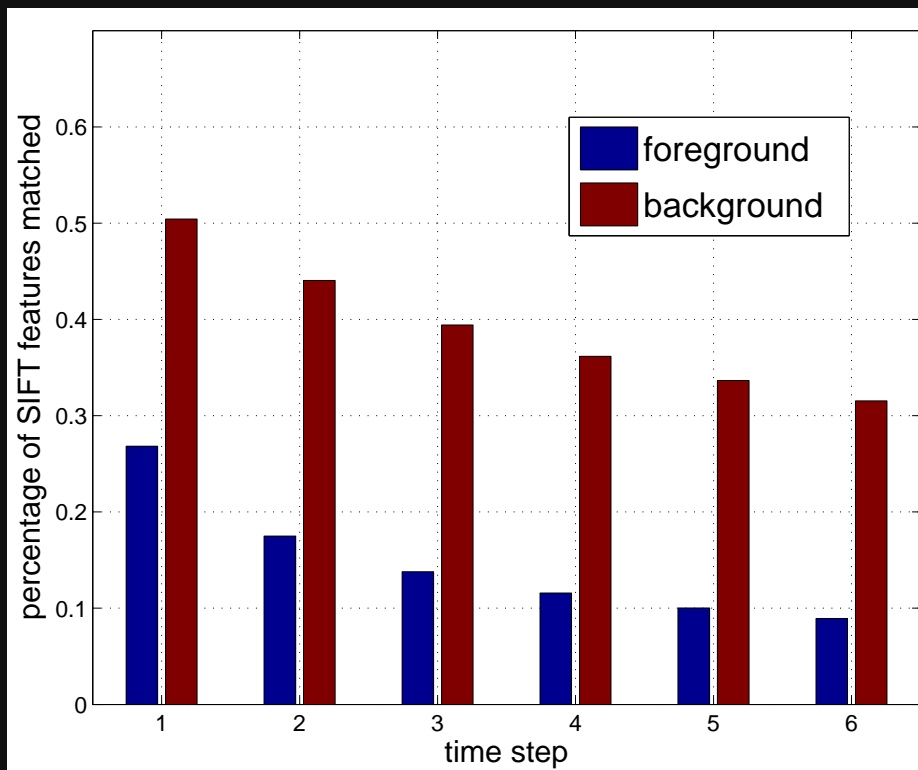


Hand location

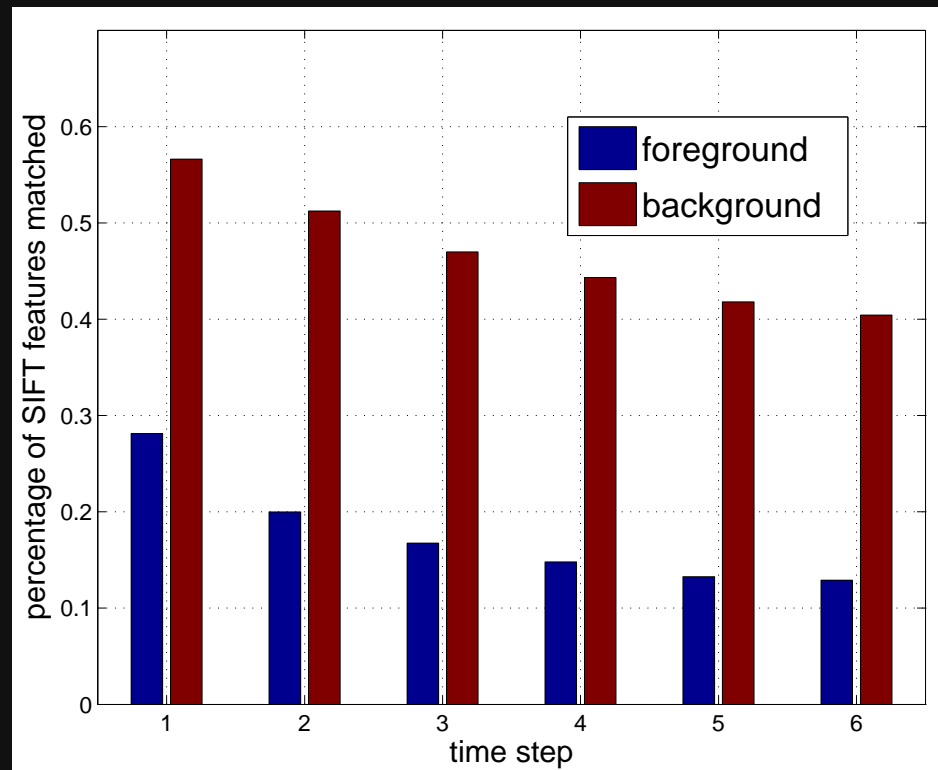
Opportunity: Feature Tracking



SIFT matching across frames, w/ or w/o enforcing epipolar constraint

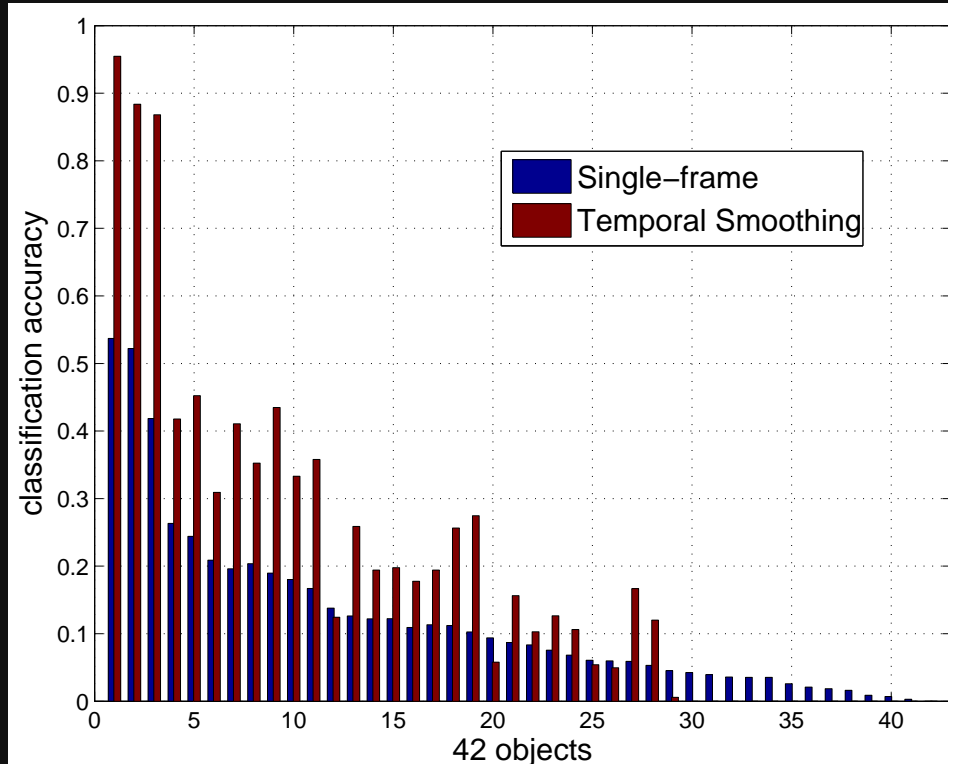
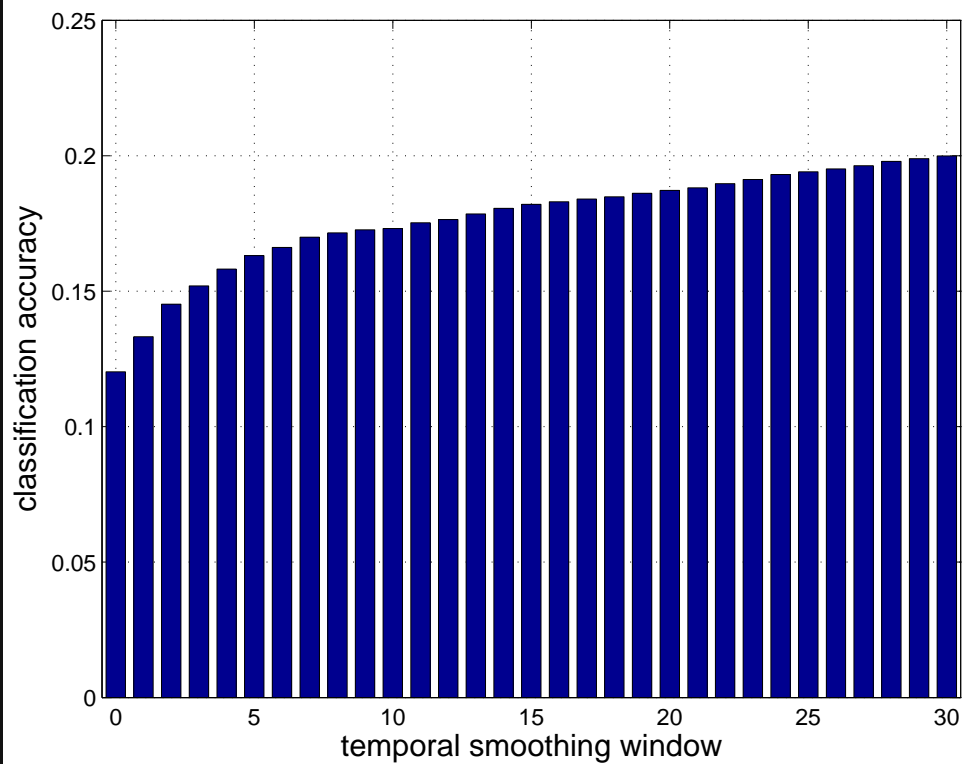


Individual feature matching



Motion layer matching

Opportunity: Temporal Consistency



Analysis: SIFT Upper Bound



12% accuracy



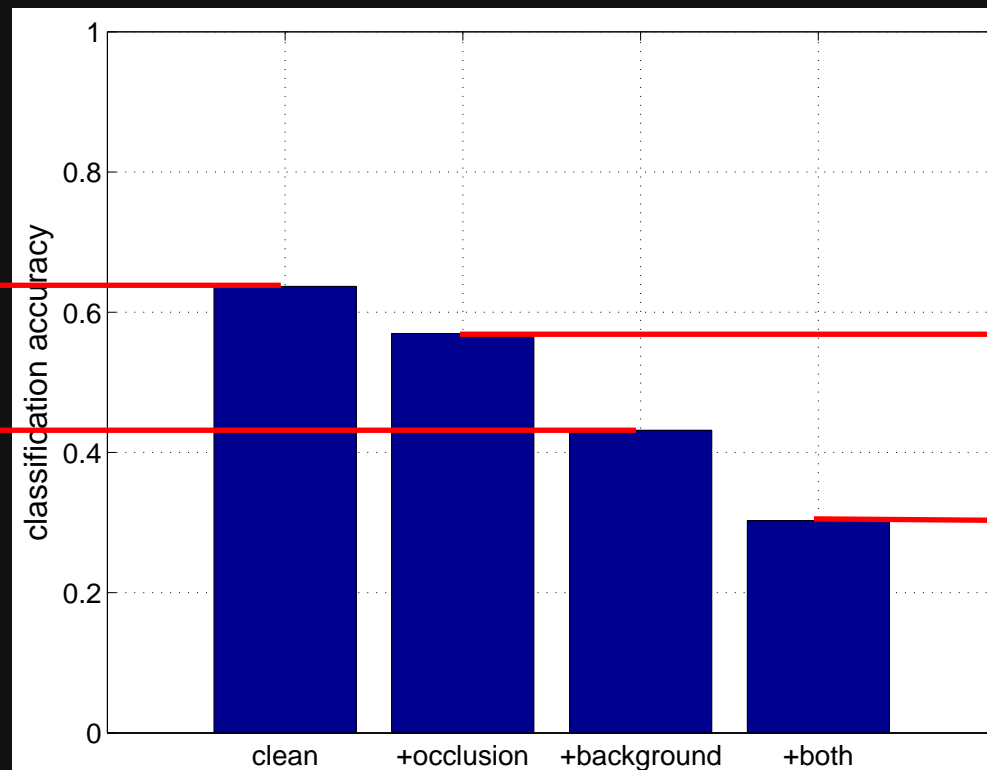
64% accuracy

Analysis: Clutter and Occlusion



64%

42%



57%

30%

Work in Progress



- Motion segmentation and figure-ground separation
- Shape matching (contour and/or HOG or ...)
- Hand and object color
- Better classification and learning framework
- From objects to activities
- Real-time implementation

... ..

Dataset available at:

<http://www.seattle.intel-research.net/~xren/egovision09/>

Everyday Sensing and Perception



- Egocentric Recognition of Handled Objects
- In-door Localization
- Face Recognition and Tracking
- Integrating Vision with Other Sensors
- High-performance / Parallel / Cloud Computing
- Power-efficient Perception
- On-line learning
- Vision meets Robotics

... ..



THANK YOU