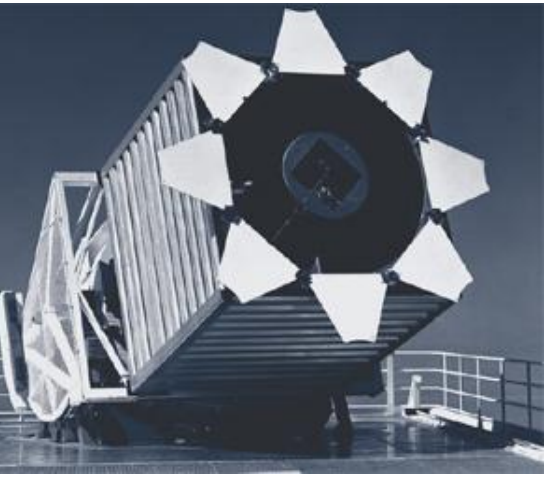
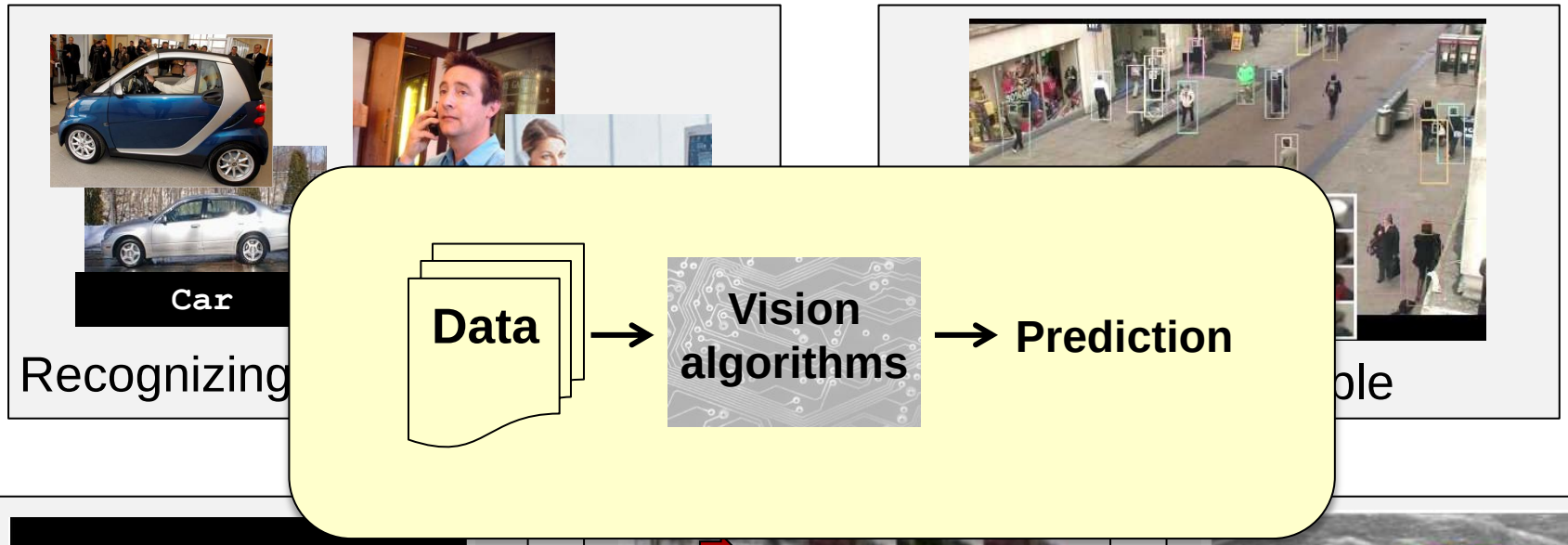
Focusing Human Attention on the "Right" Visual Data

Kristen Grauman
Department of Computer Science
University of Texas at Austin

Work with Yong Jae Lee, Sudheendra Vijayanarasimhan, and Prateek Jain



Automating visual processing



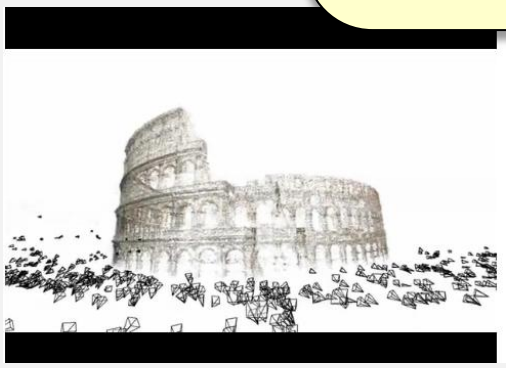
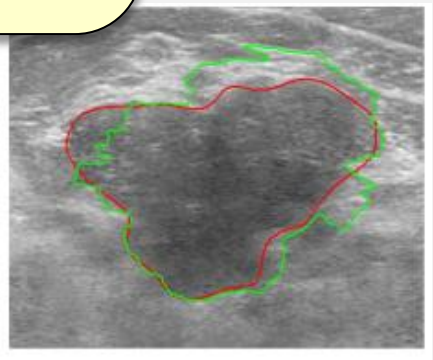
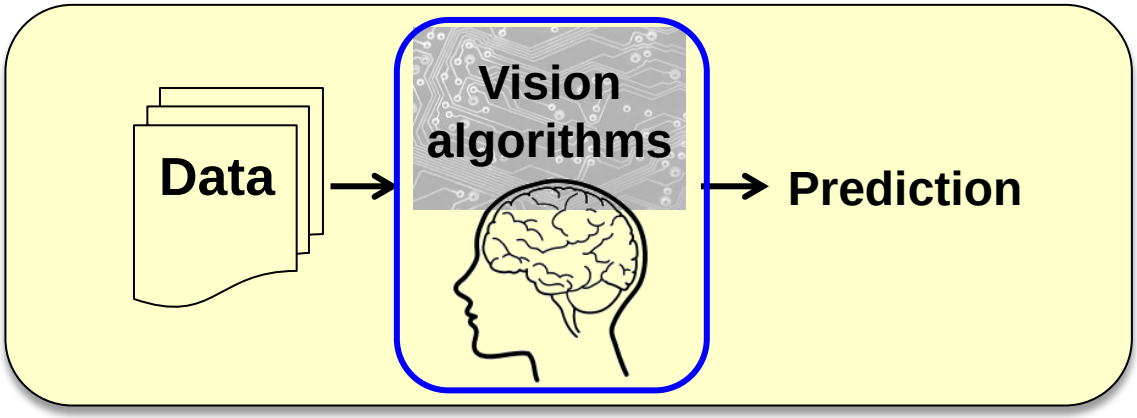

3D Reconstruction


Image retrieval


Detection

[Rome image from Agarwal et al., Tracking example from Benfold & Reid, Image retrieval example from Kulis & Grauman, Tumor image from Zhang et al.]

“Semi-automating” visual processing



Key question:

- Which visual data deserves human attention?

“Semi-automating” visual processing

We’ll consider two settings:

1. Supervised learning of object categories
2. Unsupervised video summarization

Key challenges:

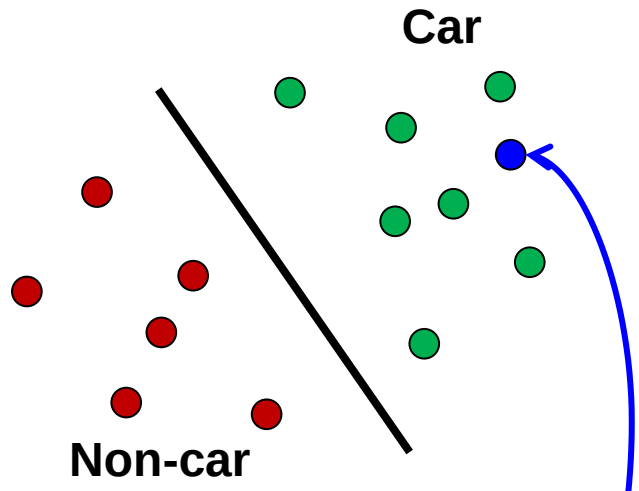
- Predicting what is important
- Scaling to large-scale data collections

The importance of data in recognition

Best approaches today rely on discriminative learning

Training images

Annotator



Novel test image

The importance of data in recognition

- Dataset creation

  [LabelMe - Russell et al. 2005, Caltech - Griffin et al. 2007, Image-Net – Deng et al. 2010, PASCAL VOC – Everingham et al.,…]

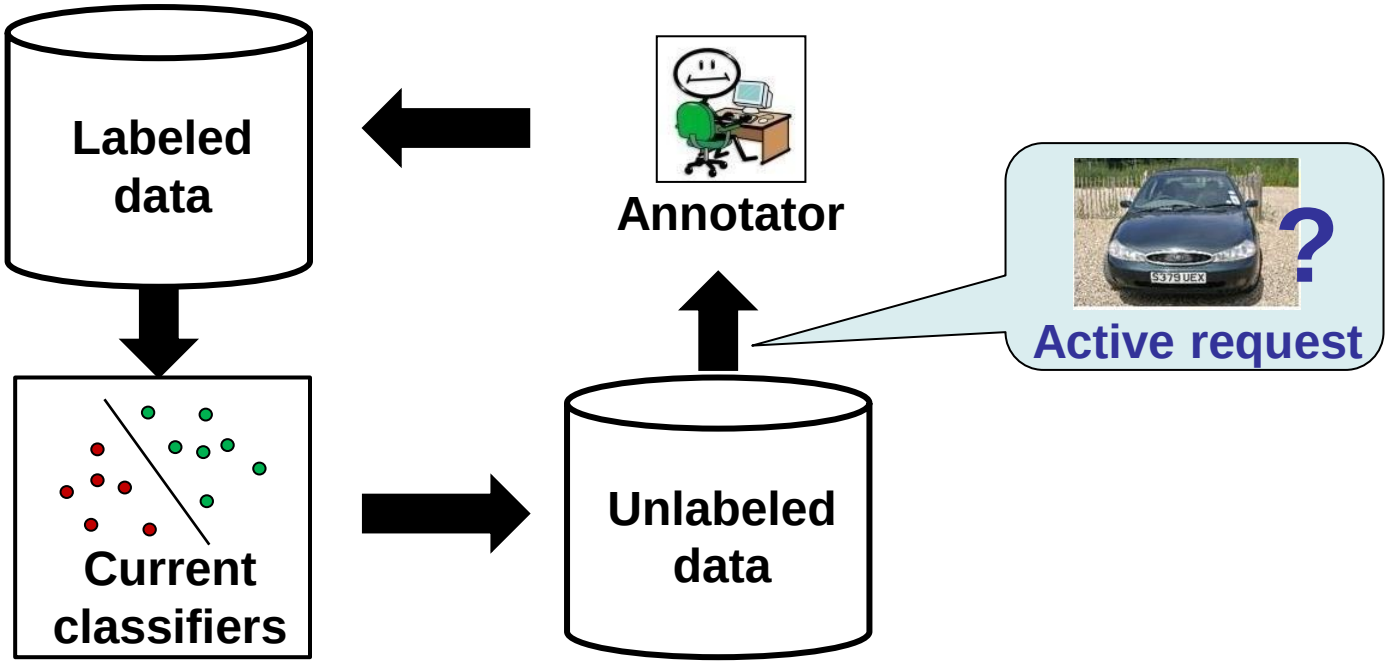
- Gathering annotations from “crowds”

  [Sorokin et al. 2009, Vijayanarasimhan et al. 2009, Deng et al 2009, Endres et al. 2010, Branson et al. 2010, Welinder et al. 2010, …]

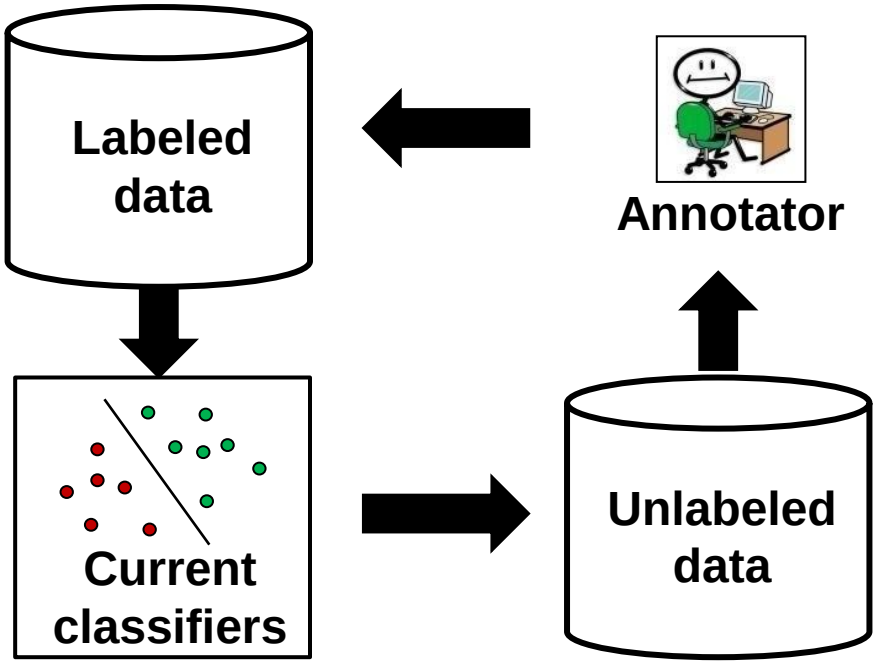
- Active learning to focus human effort

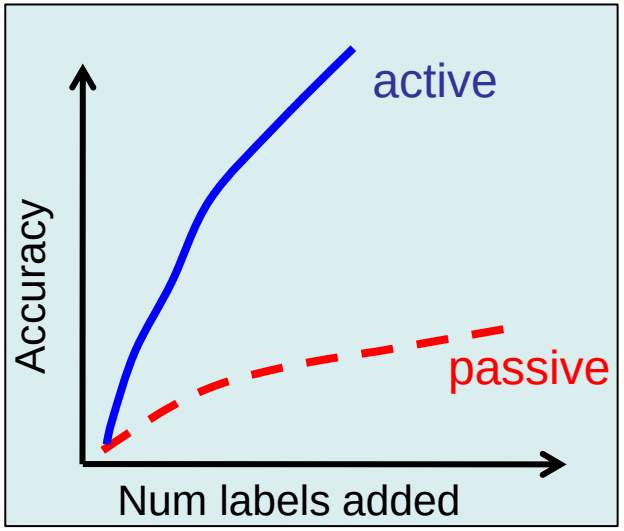
  [Kapoor et al. 2007, Qi et al. 2008, Vijayanarasimhan et al. 2008, 2009, Joshi et al. 2009, Jain et al. 2010, Siddique et al. 2010]

Active learning for image annotation



Active learning for image annotation





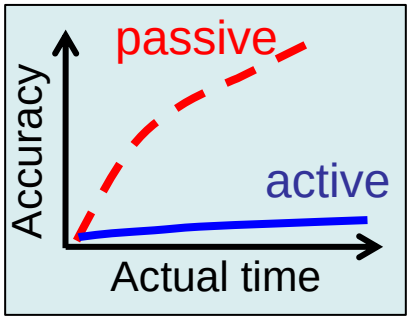
Intent: better models, faster/cheaper

Problem: “Sandbox” learning

Thus far, tested only in artificial settings:

- Unlabeled data already fixed, small scale, biased

- Computational cost ignored



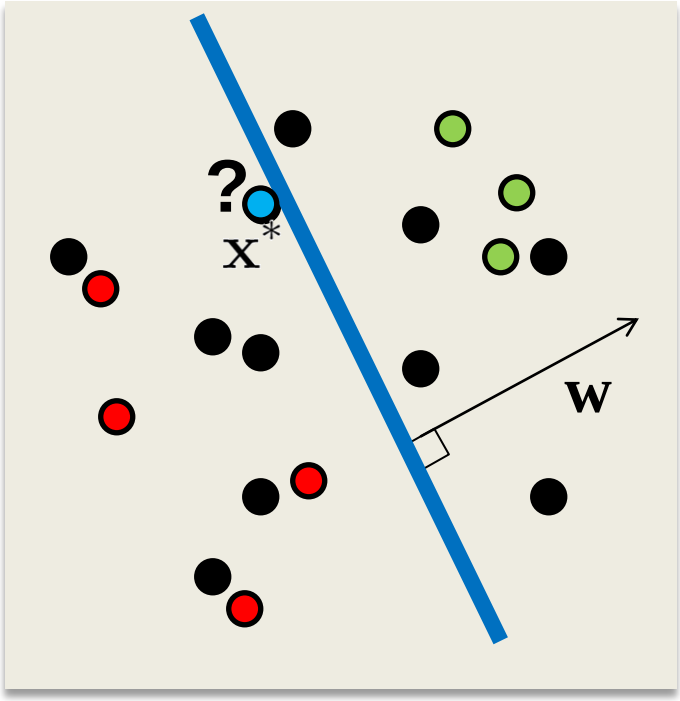
- Really, “researcher in the loop”

Our idea: Live active learning

Large-scale active learning of object detectors with crawled data and crowdsourced labels.

But how to scale active learning to massive unlabeled pools of data?

SVM margin criterion for active selection



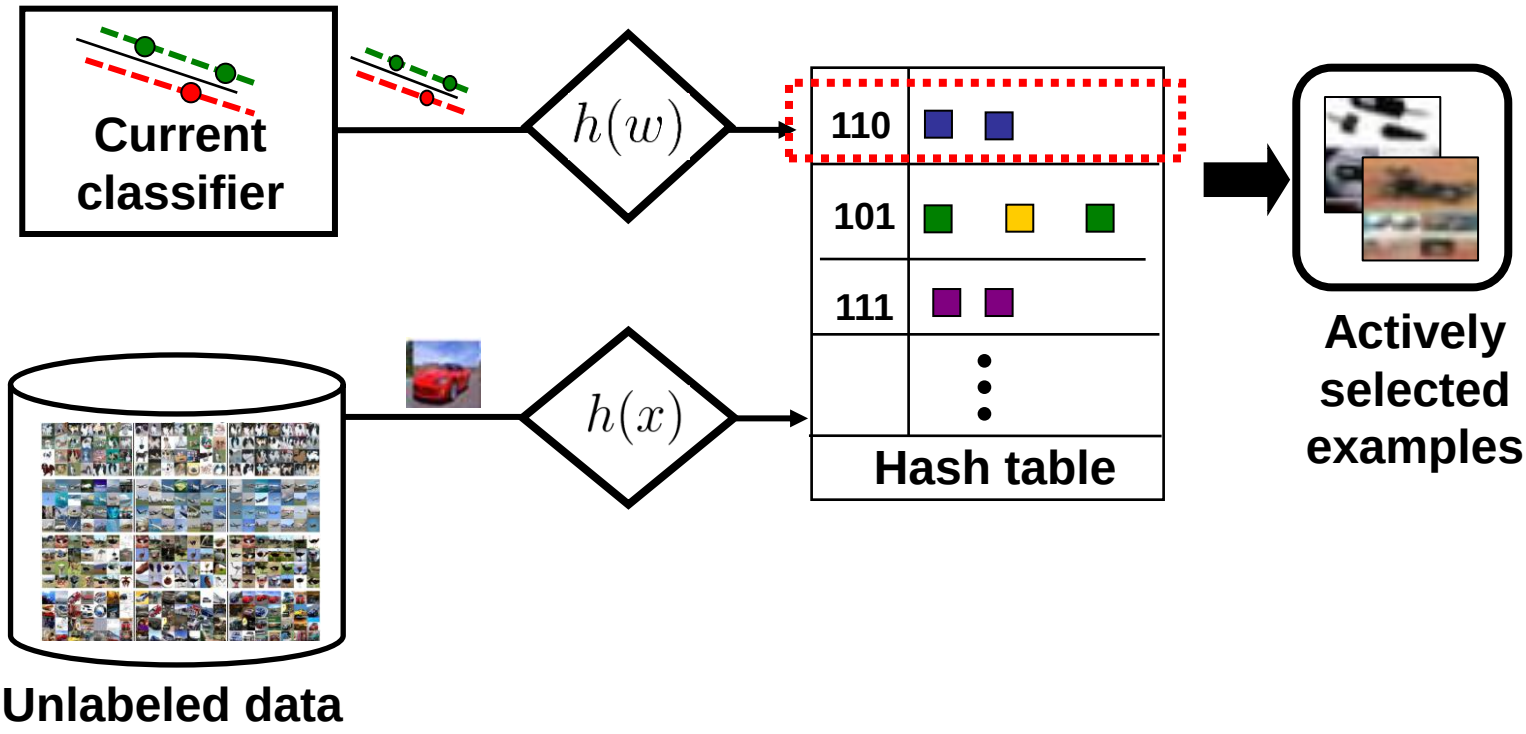
Select point nearest to hyperplane decision boundary for labeling.

$$\mathbf{x}^* = \text{argmin}_{\mathbf{x}_i \in \mathcal{U}} |\mathbf{w}^T \mathbf{x}_i|$$

[Tong & Koller, 2000; Schohn & Cohn, 2000; Campbell et al. 2000]

Sub-linear time active selection

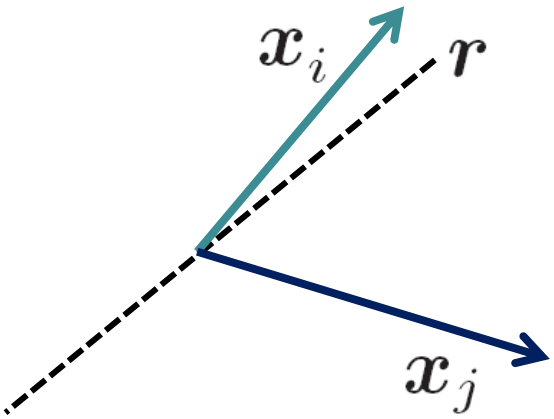
We propose a novel hashing approach to identify these most uncertain examples in sub-linear time.



[Jain, Vijayanarasimhan, Grauman, NIPS 2010]

Background: Locality-Sensitive Hashing

Probability a *random hyperplane* separates two unit vectors depends on the angle between them:



Bigger angle: likely to split

Corresponding hash function:

$$h_{\boldsymbol{r}}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{r}^T \boldsymbol{x} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$r_i \sim \mathcal{N}(0,1)$$

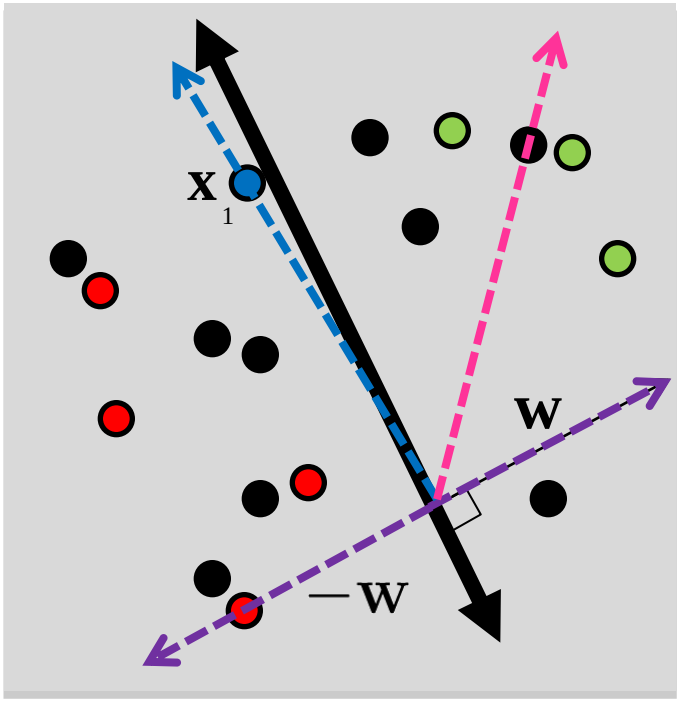
Probability of collision:

$$\Pr(h_r(x_i) = h_r(x_j)) = 1 - \frac{1}{\pi} \cos^{-1}(x_i^T x_j)$$

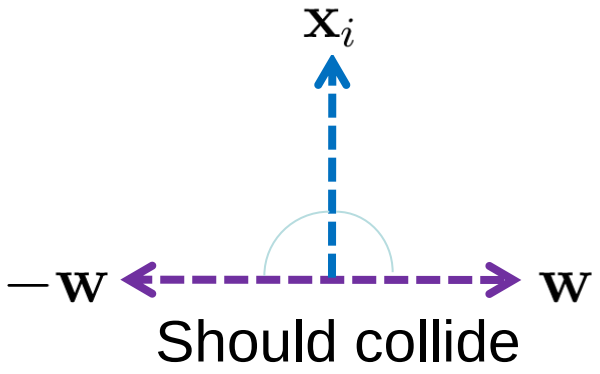
[Goemans and Williamson 1995, Charikar 2004]

Hashing a hyperplane query

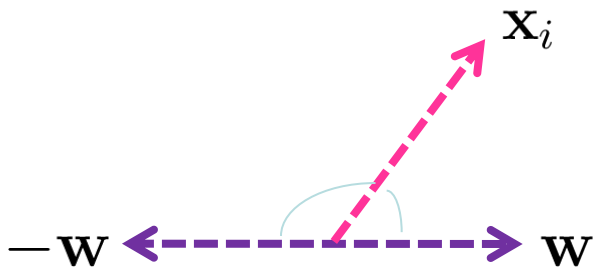
To retrieve those points for which $\left|\mathbf{w}^T \mathbf{x}_i\right|$ is small, want probable collision for **perpendicular** vectors:



Assuming normalized data.



Should collide



Should not collide

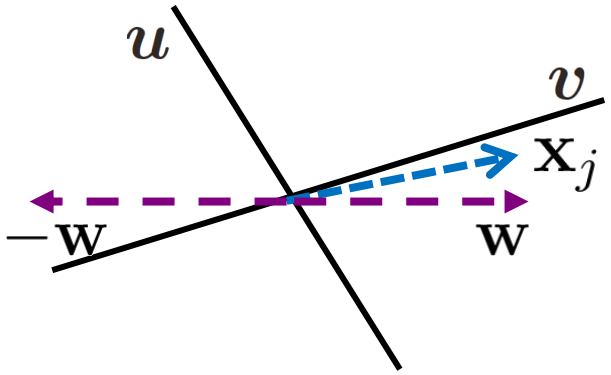
Hashing a hyperplane query

We generate two independent random vectors **u** and **v:**

- one to constrain angle between **x** and **w**
- one to constrain angle between **x** and **–w**

Collision likely only if neither vector splits

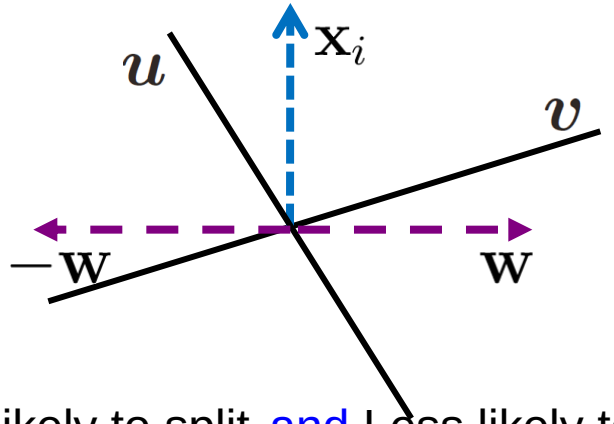
For parallel vectors



Unlikely to split and Likely to split

= Likely to split

For perpendicular vectors



Less likely to split and Less likely to split

= Unlikely to split

Hashing a hyperplane query

- We define an asymmetric 2-bit hash function:

H-Hash family:

$$h_{\mathcal{H}}(\boldsymbol{z}) = \begin{cases} h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{z},\boldsymbol{z}), & \text{if } \boldsymbol{z} \text{ is a database point vector,} \\ h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{z},-\boldsymbol{z}), & \text{if } \boldsymbol{z} \text{ is a query hyperplane vector.} \end{cases}$$

where $h_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{a},\boldsymbol{b}) = [h_{\boldsymbol{u}}(\boldsymbol{a}), h_{\boldsymbol{v}}(\boldsymbol{b})] = [\text{sign}(\boldsymbol{u}^T\boldsymbol{a}), \text{sign}(\boldsymbol{v}^T\boldsymbol{b})]$

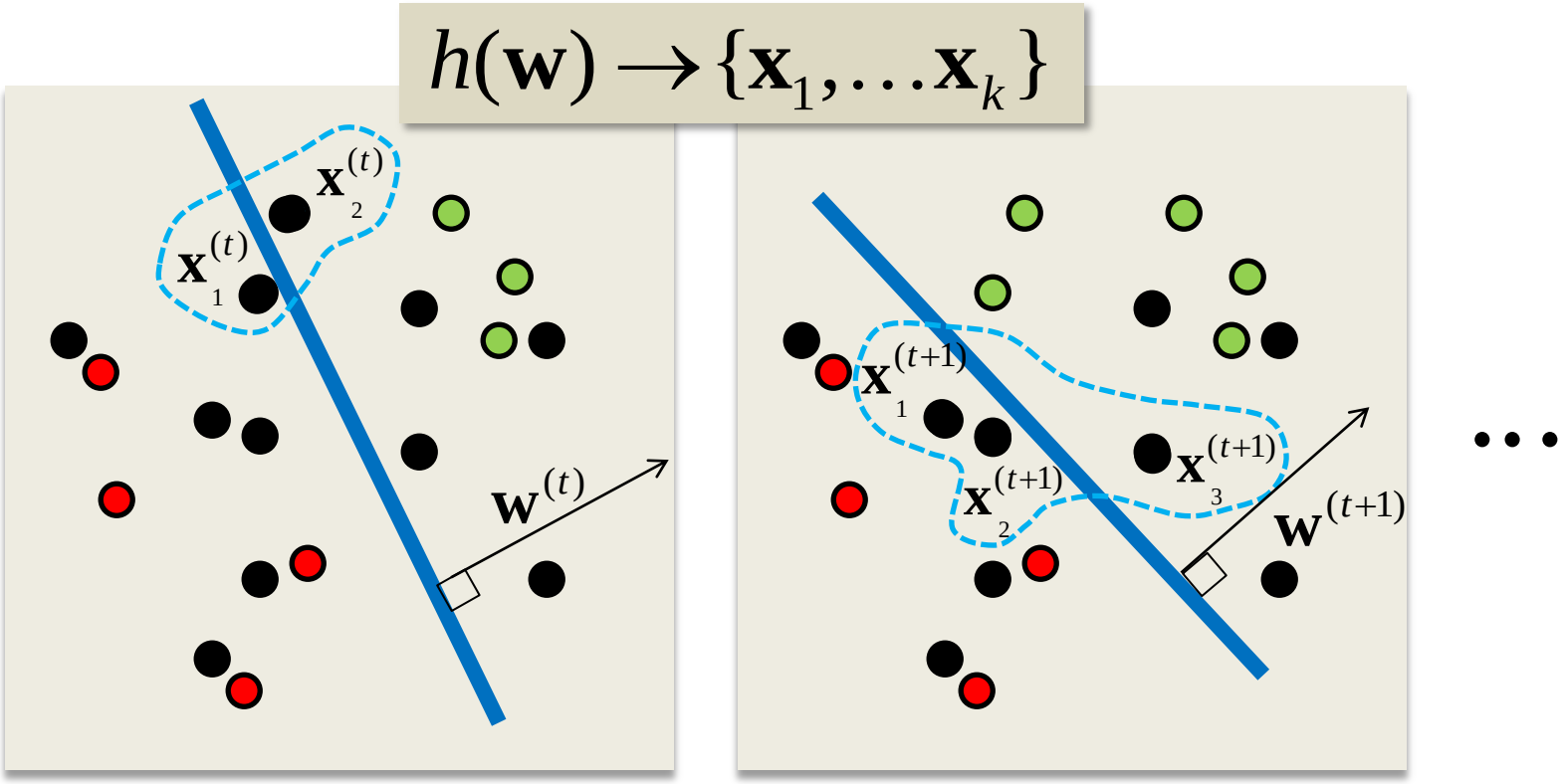
- We prove necessary conditions for locality sensitivity:

$$\Pr[h_{\mathcal{H}}(\boldsymbol{w}) = h_{\mathcal{H}}(\boldsymbol{x})] = \Pr[h_{\boldsymbol{u}}(\boldsymbol{w}) = h_{\boldsymbol{u}}(\boldsymbol{x})]\ \Pr[h_{\boldsymbol{v}}(-\boldsymbol{w}) = h_{\boldsymbol{v}}(\boldsymbol{x})]$$

$$= \frac{1}{4} - \frac{1}{\pi^2}\left(\theta_{\boldsymbol{x},\boldsymbol{w}} - \frac{\pi}{2}\right)^2$$

[Jain, Vijayanarasimhan & Grauman, NIPS 2010]

Hashing a hyperplane query



At each iteration of the learning loop, our hash functions map the current hyperplane directly to its nearest unlabeled points.

H-Hash result on Tiny Images

Images actively selected in first 9 iterations

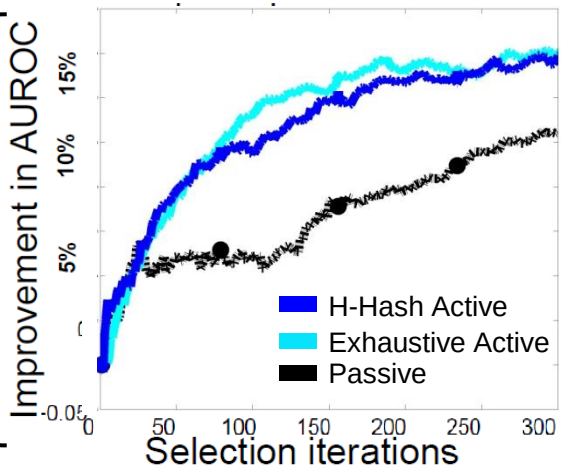
Learning "airplane"

Learning "automobile"

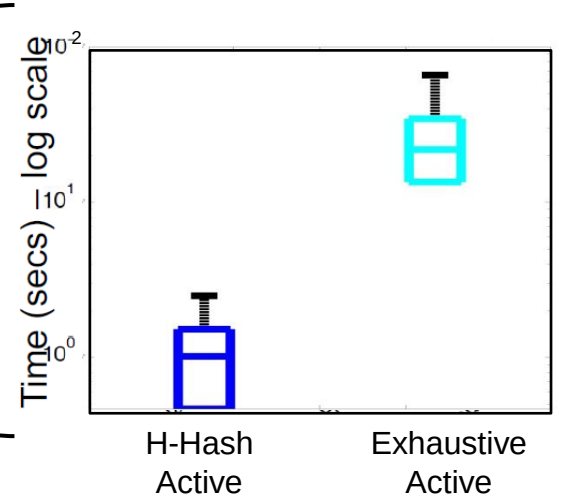
Efficient active selection with pool of 1 Million unlabeled examples and 1000s of categories!

H-Hash result on Tiny Images

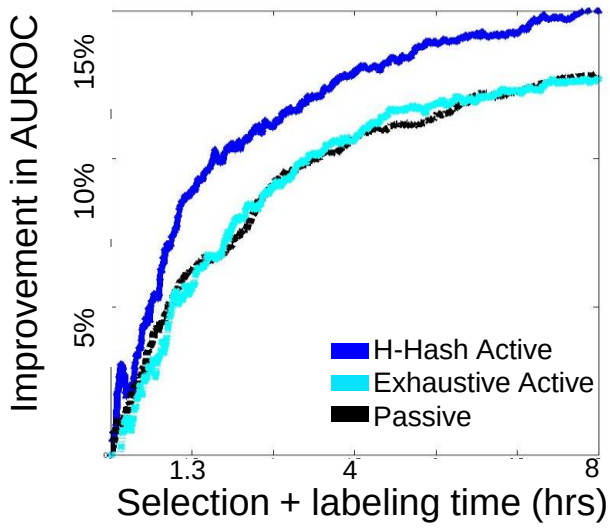
Accuracy improvements as more data labeled



Time spent searching for selection

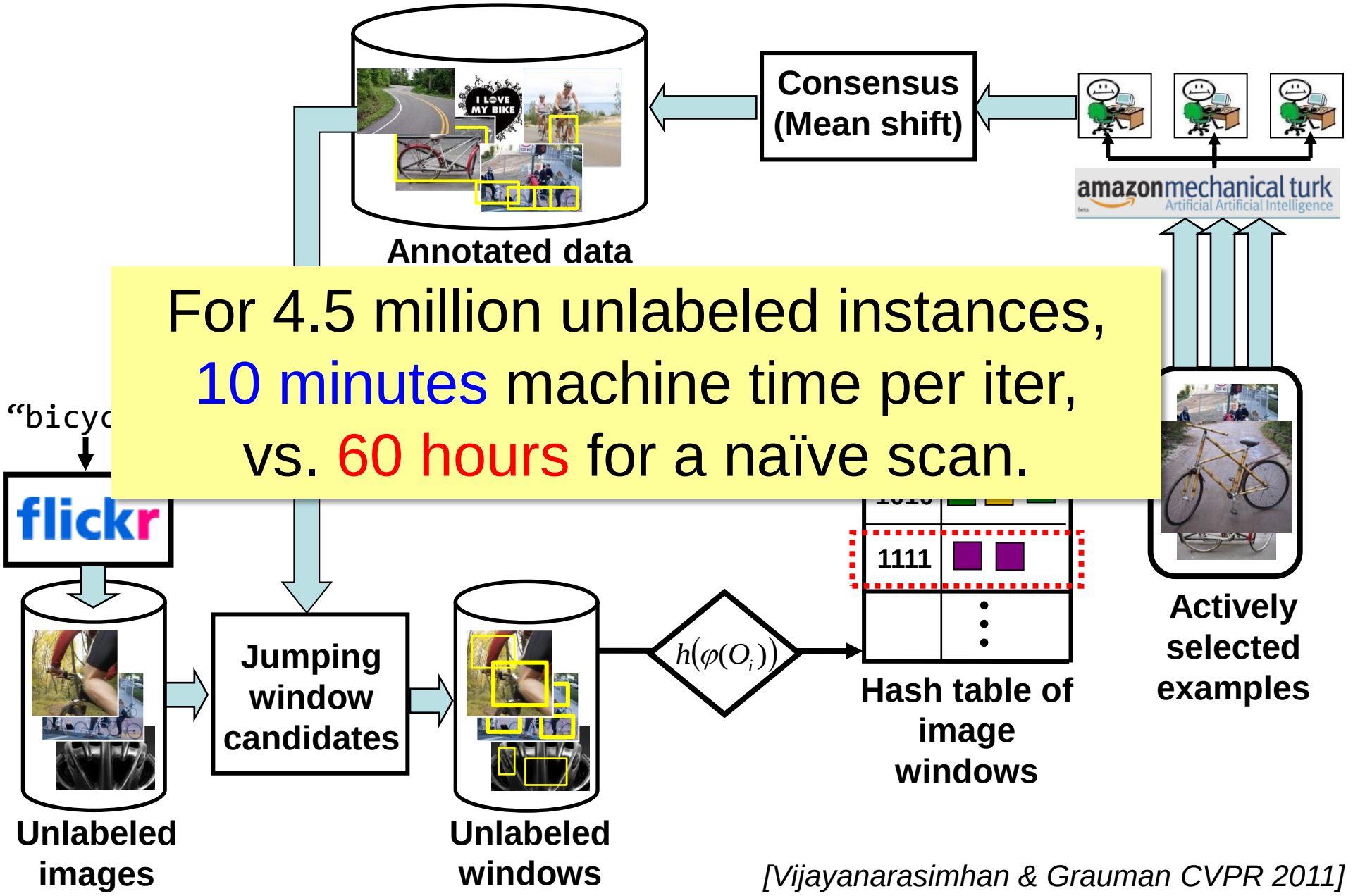


Accounting for all costs



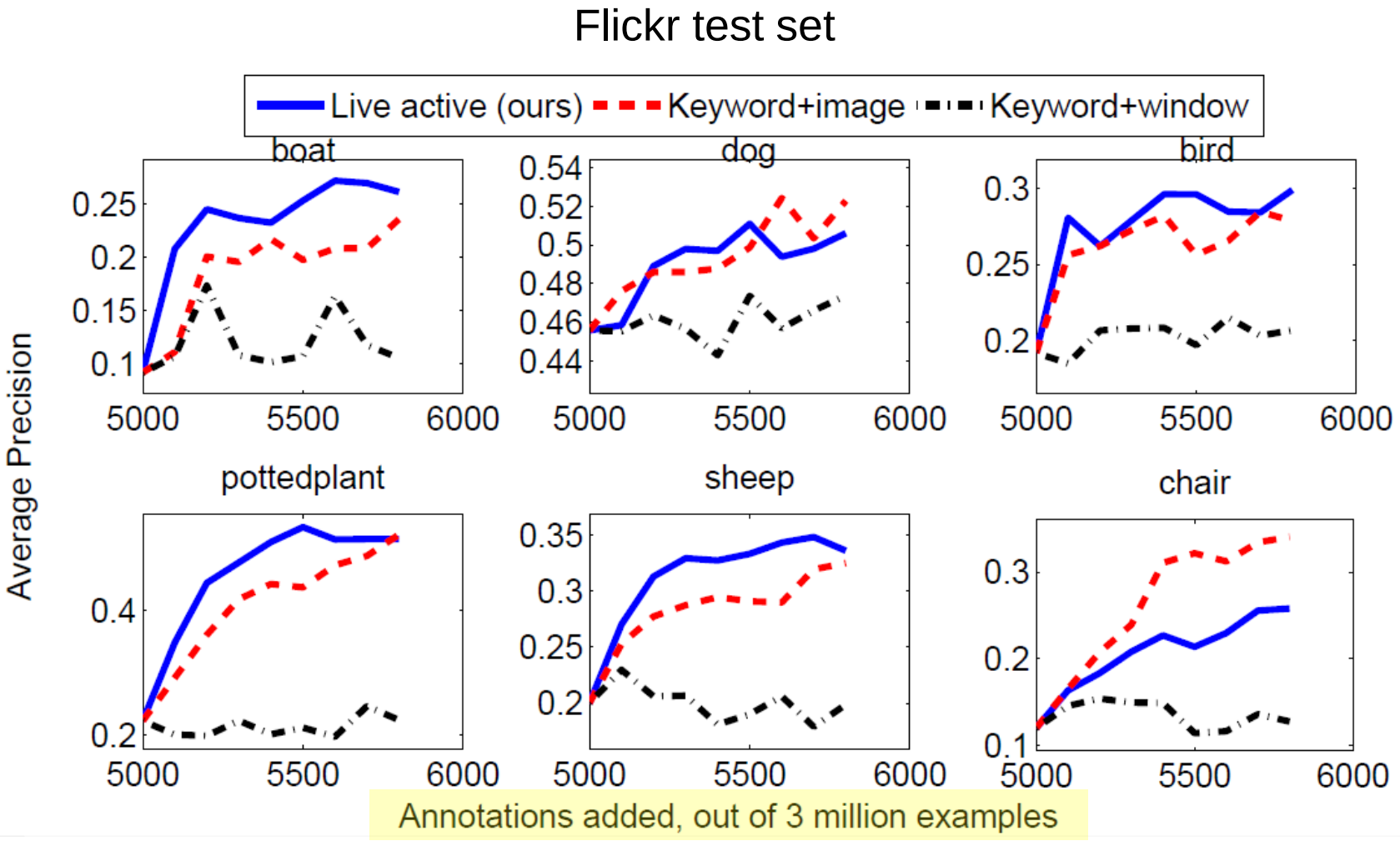
By minimizing **both** selection and labeling time, obtain the best accuracy per unit time.

Live active learning



[Vijayanarasimhan & Grauman CVPR 2011]

Live active learning results



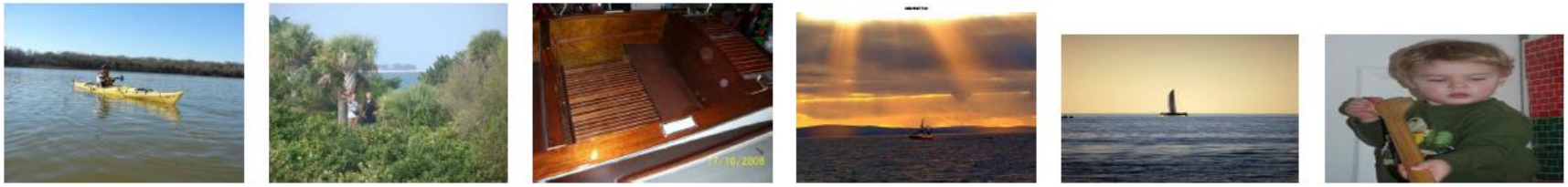
Outperforms status quo data collection approach

Live active learning results

First selections made when learning “boat”:

Ours: live active learning

Keyword+image baseline



PASCAL Visual Object Classes (VOC)

- “The” object detection benchmark
- Train/test data from Flickr

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

PASCAL Live active learning results

Live learning improves the state-of-the-art for some of most difficult PASCAL VOC categories:

	bird	boat	dog	potted plant	sheep	chair
Ours	**15.8***	**18.9***	**25.3***	11.6*	**28.4***	9.1*
Previous best	15.3	16.8	21.5	**14.6**	23.9	**17.9**

Our approach's efficiency makes live learning feasible

	Active selection	Training	Detection per image
Ours + active	10 mins	5 mins	150 secs
LSVM [Felzenszwalb et al. 2009]	3 hours	4 hours	2 secs
SP+MKL [Vedaldi et al. 2009]	93 hours	$>$ 2 days	67 secs

Previous best : [Vedaldi et al. ICCV 2009] or [Felzenszwalb et al. PAMI 2009]

Summary so far

- Active training for object recognition
- Breaking free from “sandbox” learning requires new large-scale learning algorithms
- Live active learning challenges the status quo in data collection

Main contributions:

- Hyperplane hashing for sub-linear time active selection
- First autonomous live learning results

“Semi-automating” visual processing

We’ll consider two settings:

1. Supervised learning of object categories
2. Unsupervised video summarization

Key challenges:

- Predicting what is important
- Scaling to large-scale data collections

Problem: Summarizing egocentric videos

Wearable camera

Input: Egocentric video of the camera wearer's day

9:00 am 10:00 am 11:00 am 12:00 pm 1:00 pm 2:00 pm

Output: Storyboard summary

Potential applications of egocentric video summarization

Steve Mann life logger; RHex Hexapedal Robot, Penn's GRASP Laboratory

Existing approaches to video summarization

- Previous methods rely only on low-level cues, assume a static camera, or require human intervention. [Wolf 1996, Zhang et al. 1997, Goldman et al., 2006, Caspi et al. 2006, Pritch et al. 2007, Simakov et al. 2008]

- They are indifferent to the impact that each object has on generating the "story" of the video.

Important person/object discovery

- **Our idea**: Discover important people and objects for egocentric video summarization
  - ***"Important"***: things with which the camera wearer has significant interaction
  - Develop novel egocentric and high-level saliency features to train a **category-independent** important person/object detector
  - Produce a concise visual summary driven by those detections

[Lee, Ghosh, Grauman, CVPR 2012]

Approach overview

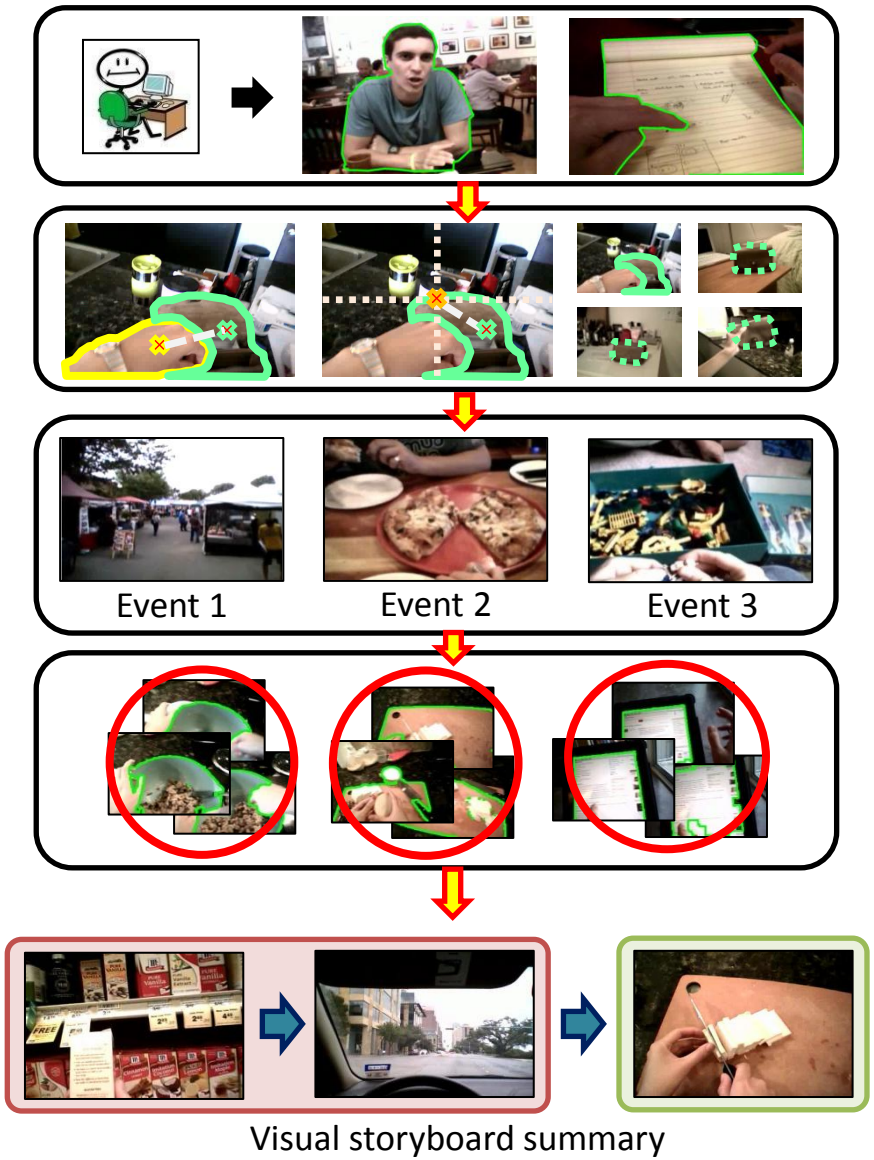
1) Crowd-source important person/object annotations

2) Design features to train an importance detector

3) Given a new video, segment it into unique temporal events

4) For each event, discover important people and objects

5) Create a compact storyboard summary that encapsulates the main people and objects



[Lee, Ghosh, Grauman, CVPR 2012]

Data collection

- 15 fps, 320 x 480 resolution
- 10 videos, 3-5 hrs in length; total of 37 hrs
- Four subjects: one undergraduate, two grad students, and one office worker

Crowdsourcing training data

Man wearing a blue shirt and watch in coffee shop

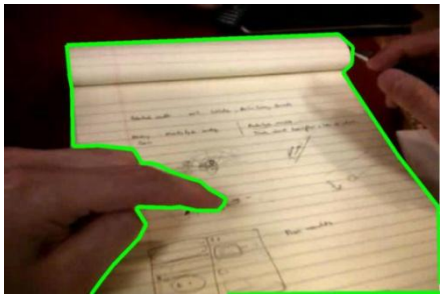
Yellow notepad on table

Coffee mug that cameraman drinks

- First task: watch a short clip, and *describe in text* the essential people or objects necessary to create a summary

Crowdsourcing training data

Man wearing a blue shirt and watch in coffee shop



Yellow notepad on table

Coffee mug that cameraman drinks

Iphone that the camera wearer holds

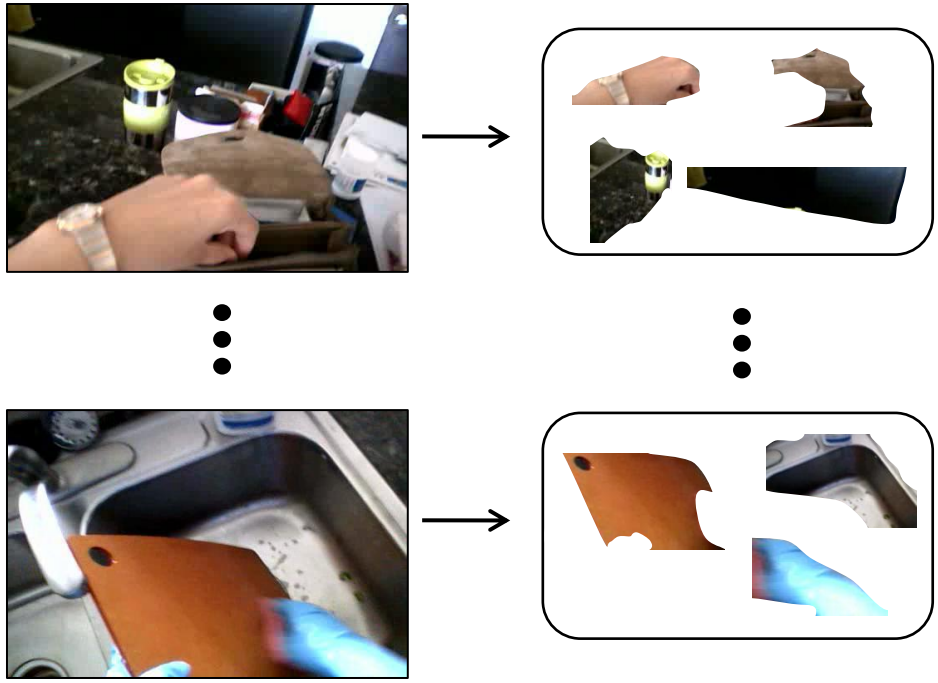
Camera wearer cleaning the plates

Soup bowl

- Second task: draw polygons around any described person or object *obtained from the first task* in sampled frames

Learning region importance

Input: egocentric video



Generate candidate object regions for uniformly sampled frames

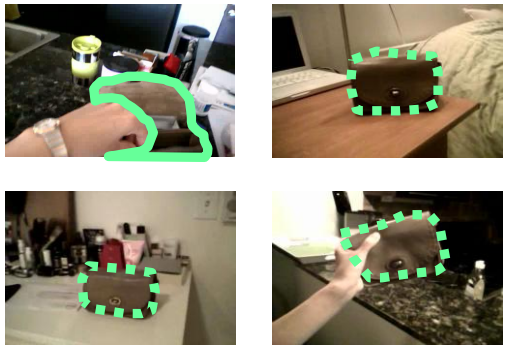
- Uncontrolled setting prohibits reliable space-time segmentation

Learning region importance

Egocentric features:

distance to hand

distance to frame center



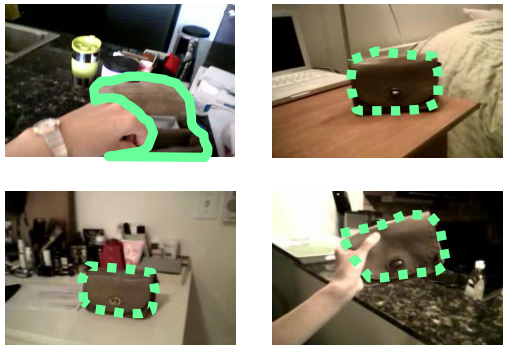
frequency

Learning region importance

Egocentric features:

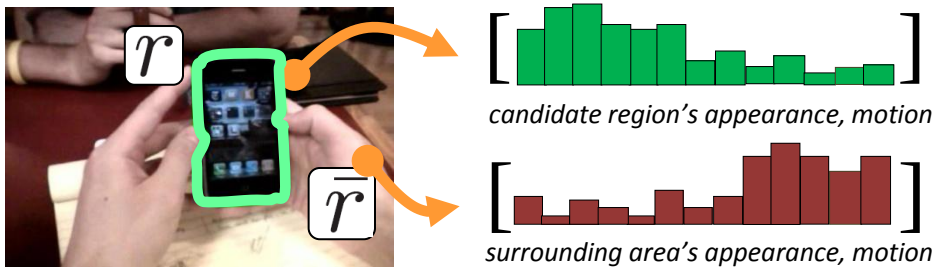
distance to hand

distance to frame center



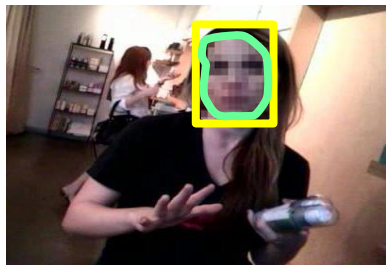
frequency

Object features:



"Object-like" appearance, motion

[Endres et al. ECCV 2010, Lee et al. ICCV 2011]



overlap w/ face detection

Region features: *size, width, height, centroid*

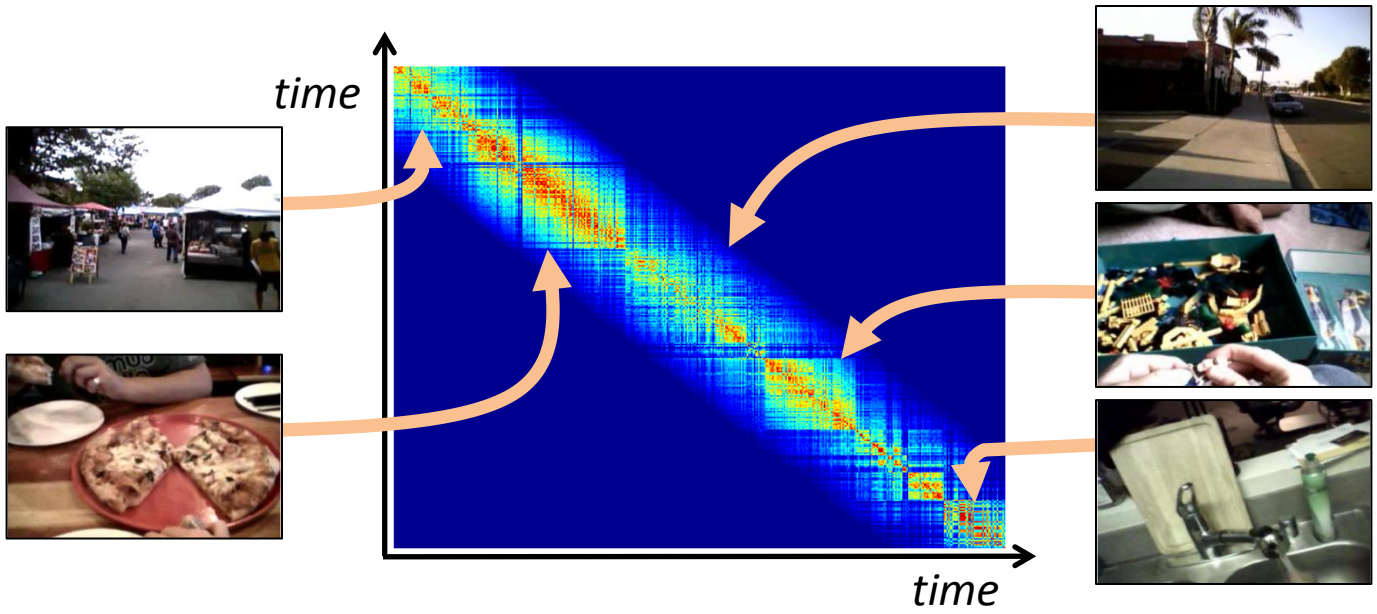
Learning region importance

$$I(r) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i(r) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \beta_{i,j} x_i(r) x_j(r)$$

importance — learned parameters — i'th feature value

- Regressor to learn and predict a region's *degree* of importance
- Expect significant interactions between the features
- For training: $I(r) = \frac{|GT \cap r|}{|GT \cup r|}$
- For testing: predict $I(r)$ given $x_i(r)$'s

Segmenting the video into events



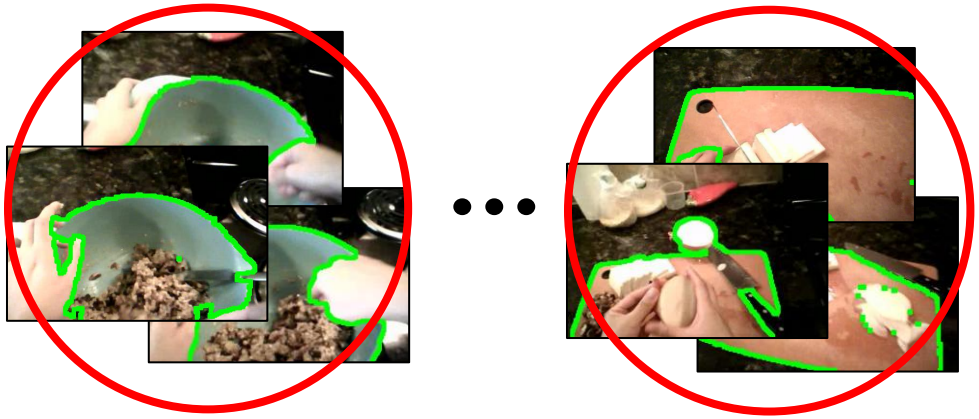
$$D(f_m, f_n) = 1 - w^t_{m,n} \exp(-\frac{1}{\Omega}\chi^2(f_m, f_n))$$

- Events allow summary to include multiple instances of a person/object that is central in multiple contexts in the video

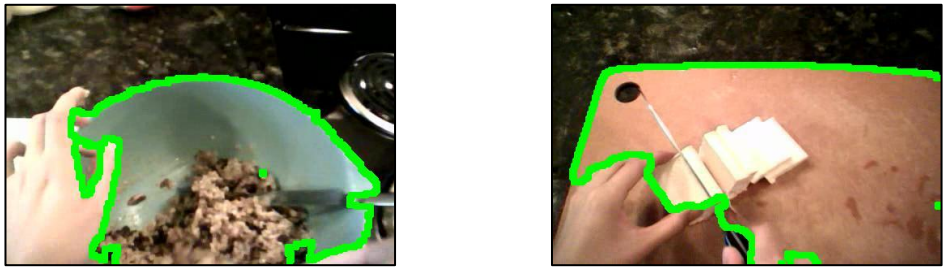
Discovering an event's key people/objects

Event A

Score and group regions in event



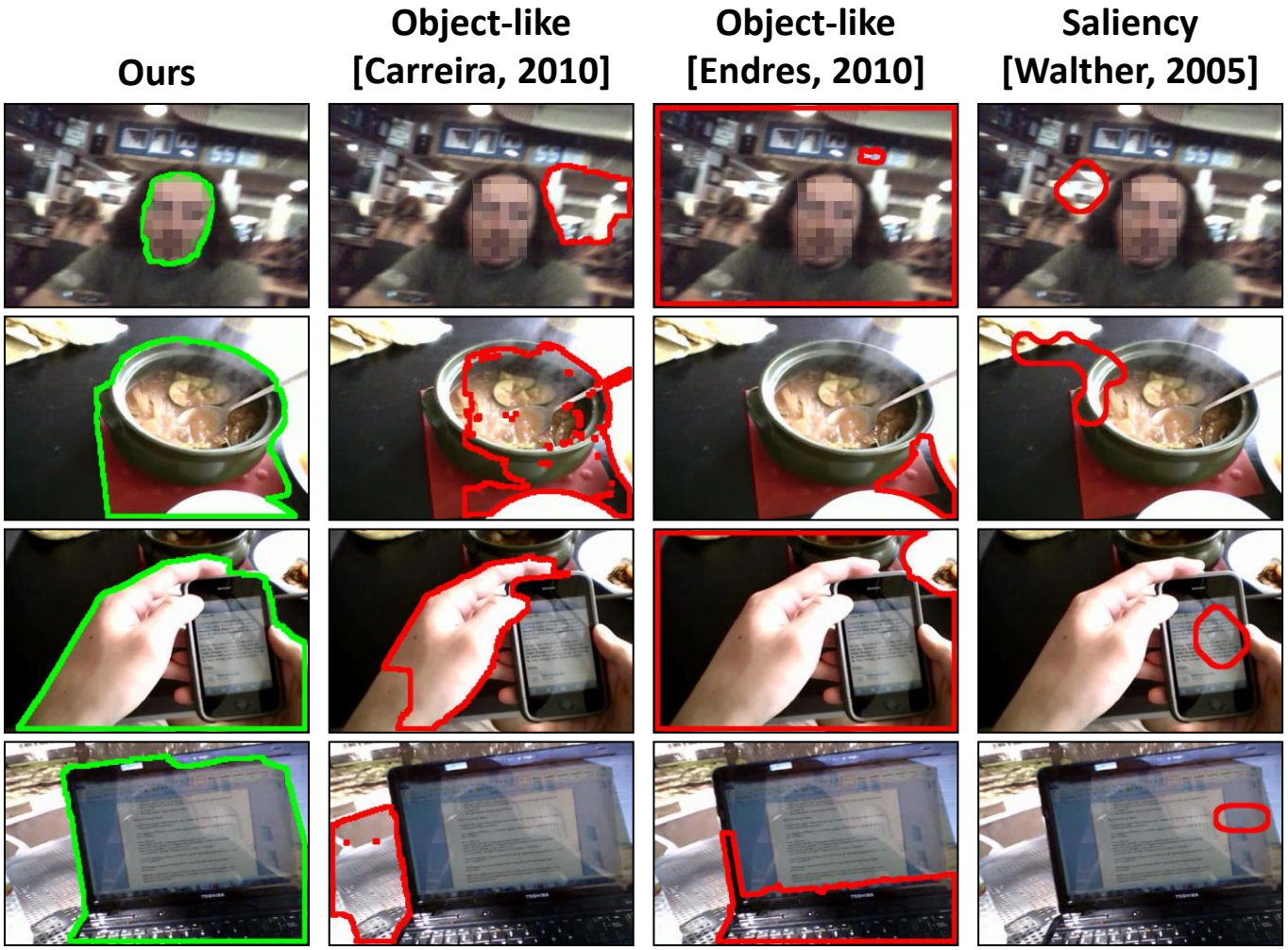
Select representative region w/ highest $I(r)$



Generating a storyboard summary

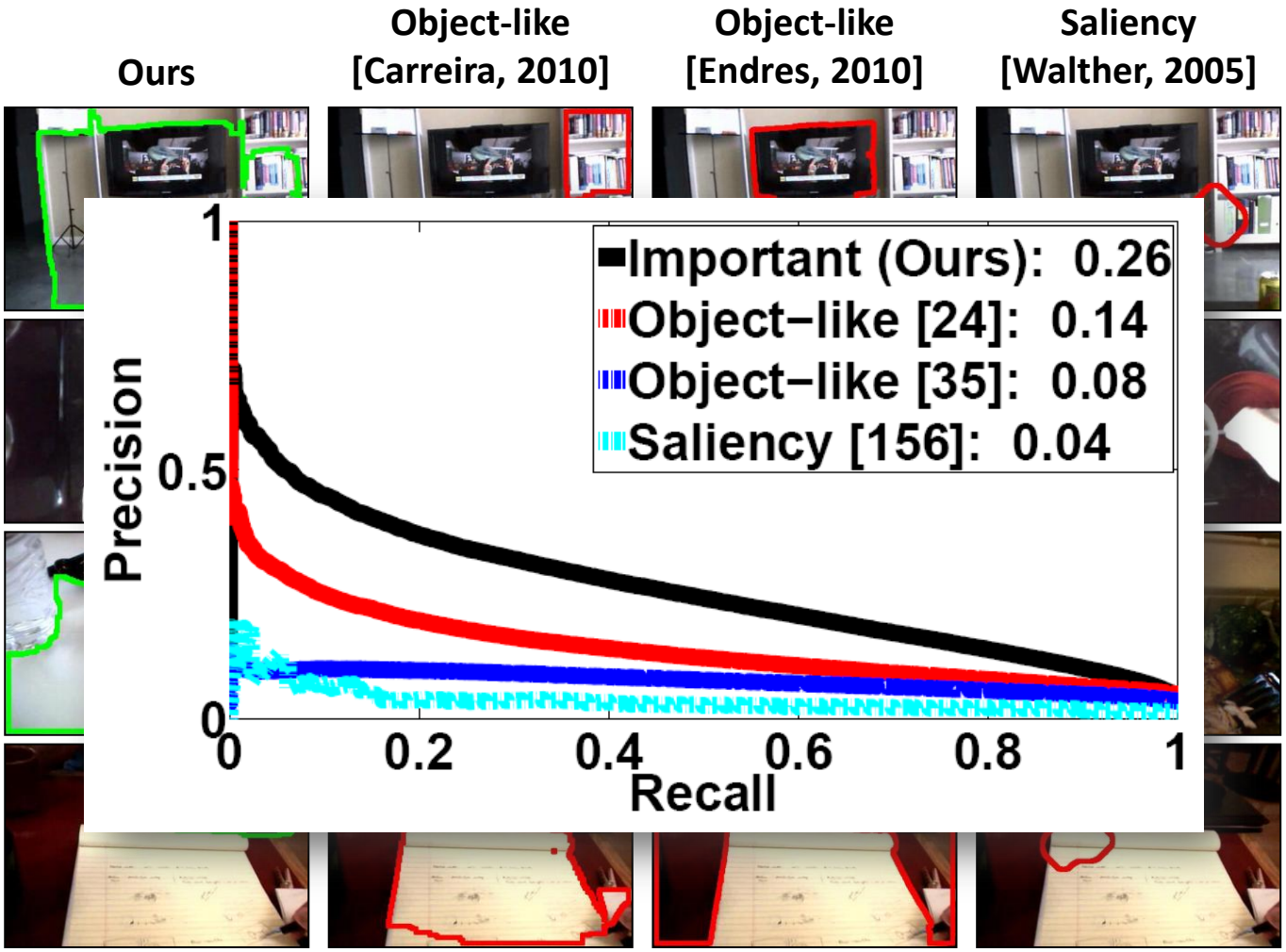
- Display event boundaries and frames of the selected important people and objects

Results: Important region prediction



Good predictions

Results: Important region prediction



Failure cases

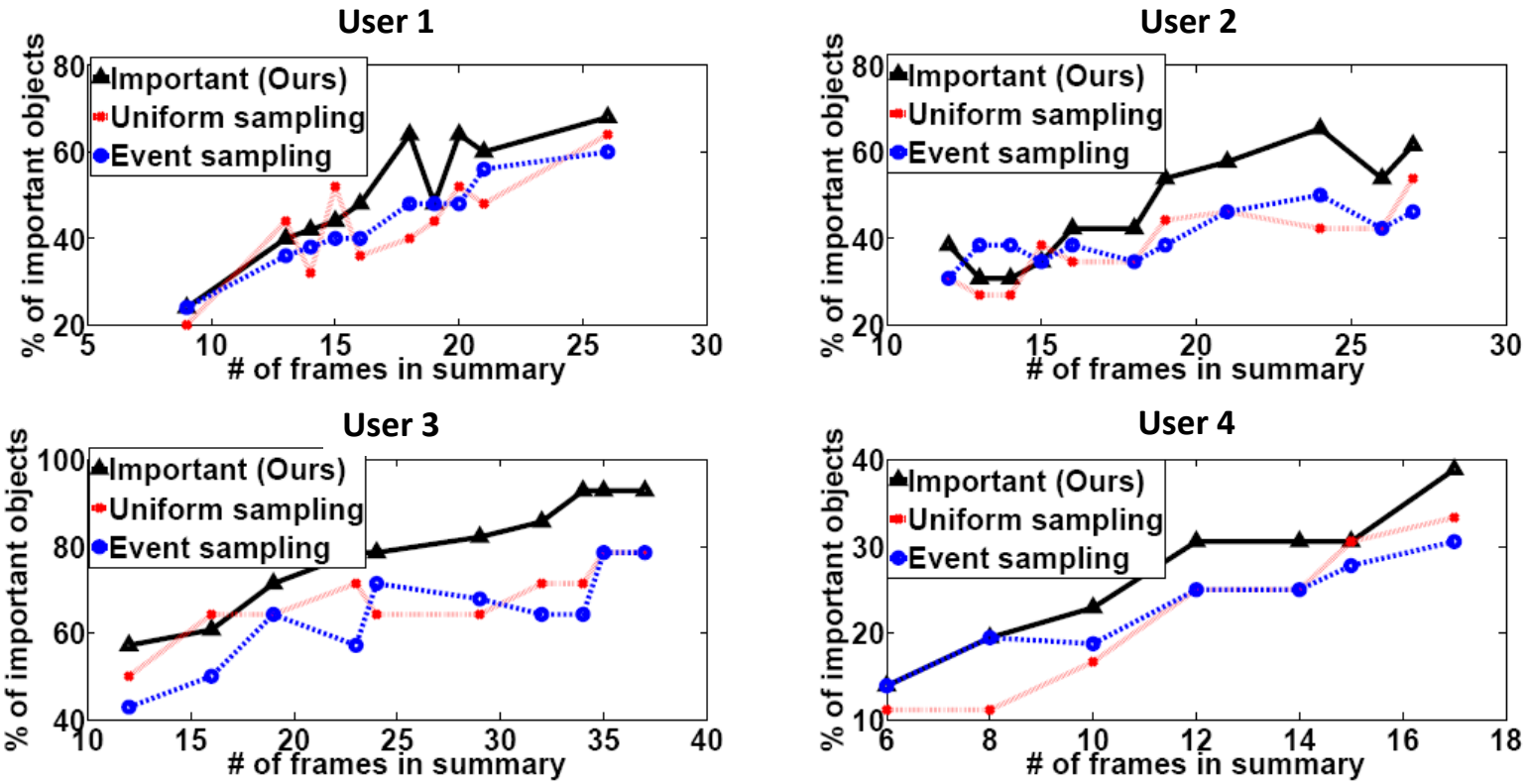
Results: Egocentric video summarization

Alternative methods for comparison

Uniform keyframe sampling (12 frames)

[Liu & Kender, 2002] (12 frames)

Results: Egocentric video summarization



- Our summaries include more important objects with fewer frames

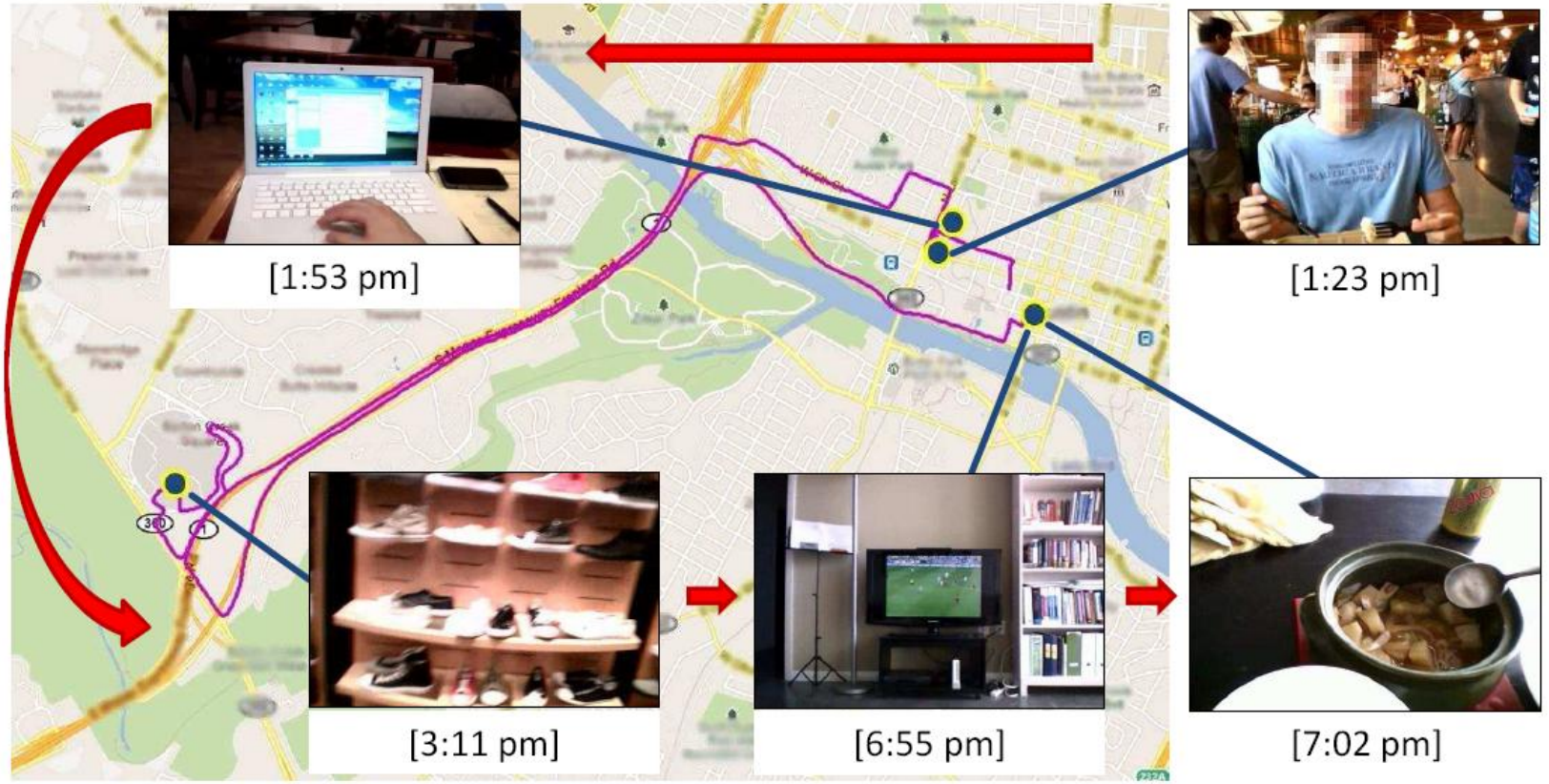
Results: Egocentric video summarization

	Much better	Better	Similar	Worse	Much worse
Imp. captured	31.25%	37.5%	18.75%	12.5%	0%
Overall quality	25%	43.75%	18.75%	12.5%	0%

- User studies to compare ours vs. uniform sampling
  (1) Which summary captures important objects/people better?
  (2) Which provides a better overall summary?

Results: Egocentric video summarization

Generating a storyboard map



Summary

- Learn to focus human attention on the right data
  - Actively train object detector with human in the loop
  - Summarize videos for fast human consumption

- Semi-automating computer vision tasks → new applications in large-scale visual analysis