Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds

Sudheendra Vijayanarasimhan and Kristen Grauman
Presenter: Yangzihao Wang

University of California, Davis

yzhwang@ucdavis.edu

October 30, 2014

UCDAVIS COLLEGE OF ENGINEERING

One Challenge in Object Detection

Labeling Bottleneck: How to provide a large number of cleanly labeled images (e.g.: thousands of bounding box annotations per category) when

One Challenge in Object Detection

Labeling Bottleneck: How to provide a large number of cleanly labeled images (e.g.: thousands of bounding box annotations per category) when

- Unlabeled image is abundant;

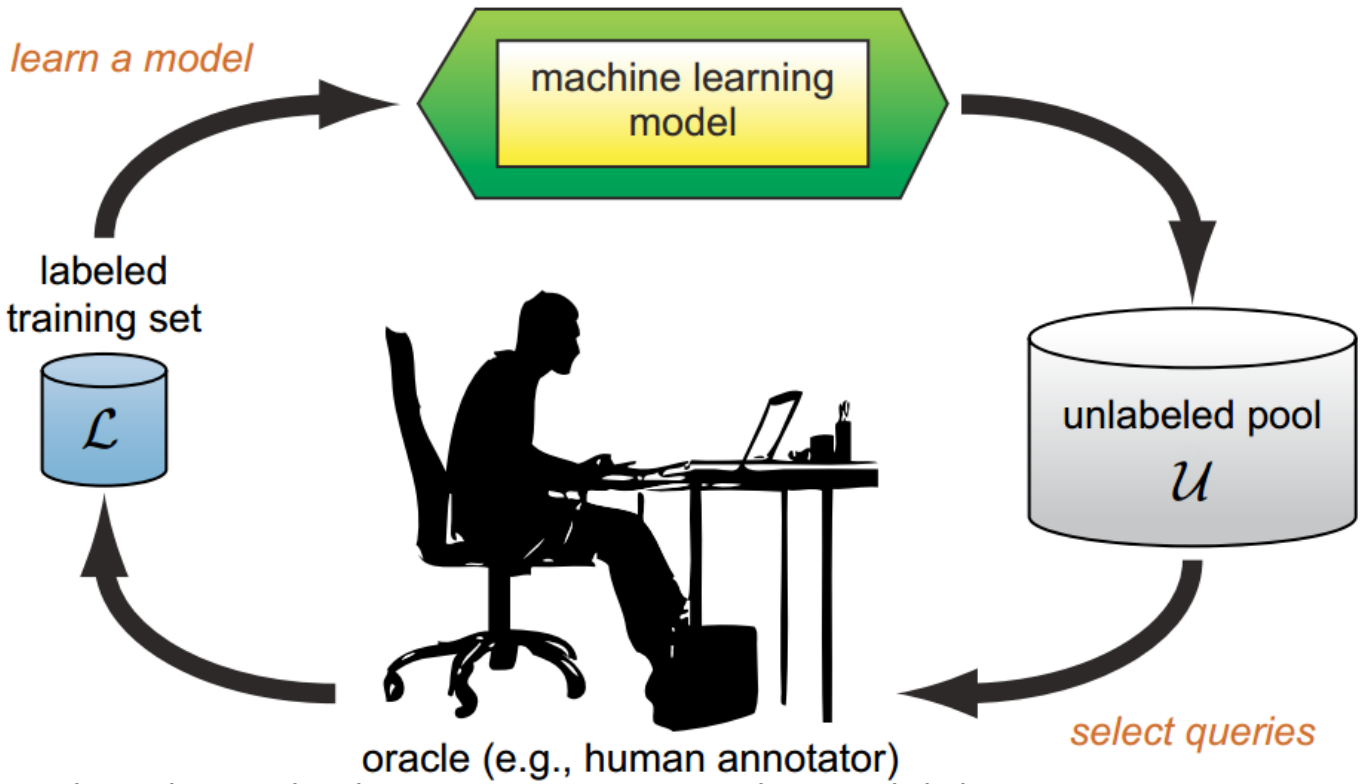
One Challenge in Object Detection

Labeling Bottleneck: How to provide a large number of cleanly labeled images (e.g.: thousands of bounding box annotations per category) when

- Unlabeled image is abundant;
- But labeling image is expensive.

Active Learning and Crowd-sourced Labeling To the Rescue!

Active learning: minimize human effort by focusing label requests on those that appear most informative to the classifier.



Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison. 2009.

UCDAVIS COLLEGE OF ENGINEERING

Active Learning and Crowd-sourced Labeling To the Rescue!

Crowd-sourced Labeling: How to package annotation tasks such that they can be dispersed effectively online.

(LabelMe gallery picture cited online)

UCDAVIS COLLEGE OF ENGINEERING

Problem with current (2011) active learning systems

Problem with current (2011) active learning systems

- **Pre-determined dataset source and scope** Meaning possible biased dataset and inflated performance;

Problem with current (2011) active learning systems

- **Pre-determined dataset source and scope** Meaning possible biased dataset and inflated performance;
- **Detection problem is ignored** Unlabeled pool are artificially assumed to contain only one prominent object;

Problem with current (2011) active learning systems

- **Pre-determined dataset source and scope** Meaning possible biased dataset and inflated performance;
- **Detection problem is ignored** Unlabeled pool are artificially assumed to contain only one prominent object;
- **Iterative fine-tuning of the crowd-sourced collection**

Problem with current (2011) active learning systems

- **Pre-determined dataset source and scope** Meaning possible biased dataset and inflated performance;
- **Detection problem is ignored** Unlabeled pool are artificially assumed to contain only one prominent object;
- **Iterative fine-tuning of the crowd-sourced collection**
- **Computational complexity of the active selection process is ignored** Critical when running a large-scale live system.

Goal of the Paper

Take crowd-sourced active annotation out of the "sandbox".

Goal of the Paper

Take crowd-sourced active annotation out of the "sandbox".

- No canned dataset, all relevant images acquired via keyword search on Flickr.

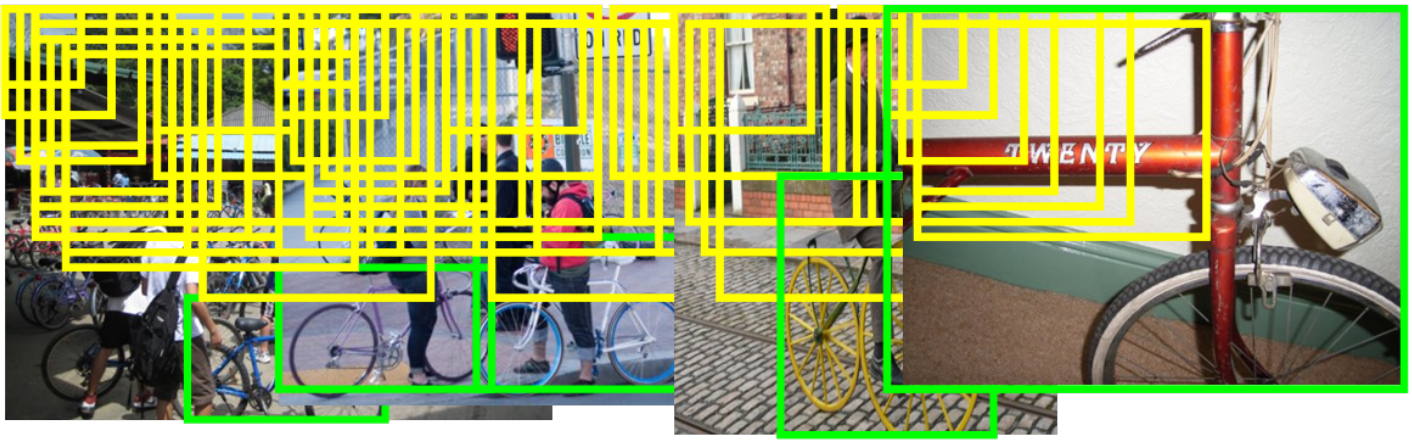
Goal of the Paper

Take crowd-sourced active annotation out of the "sandbox".

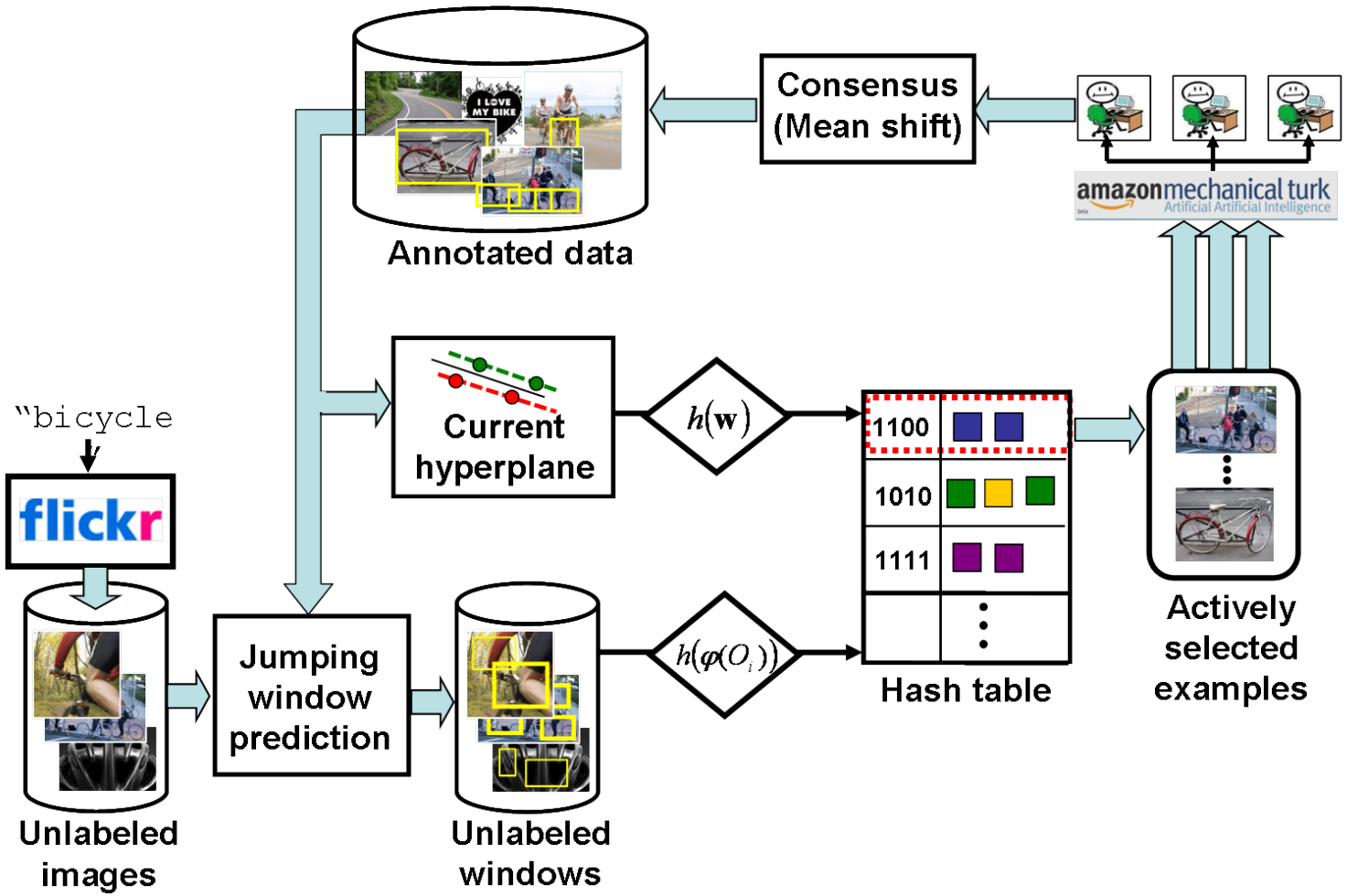
- No canned dataset, all relevant images acquired via keyword search on Flickr.
- No intervene with selection and annotation quality.

Technical Challenge

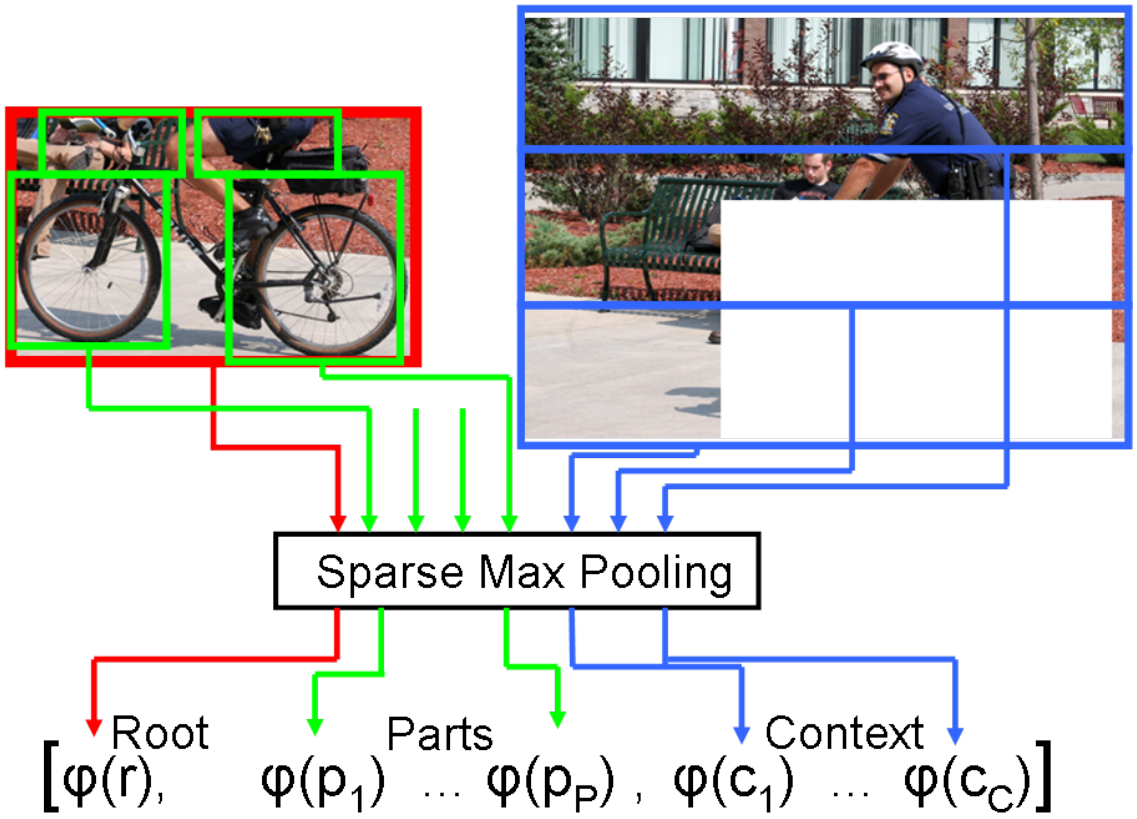
Large-scale active selection



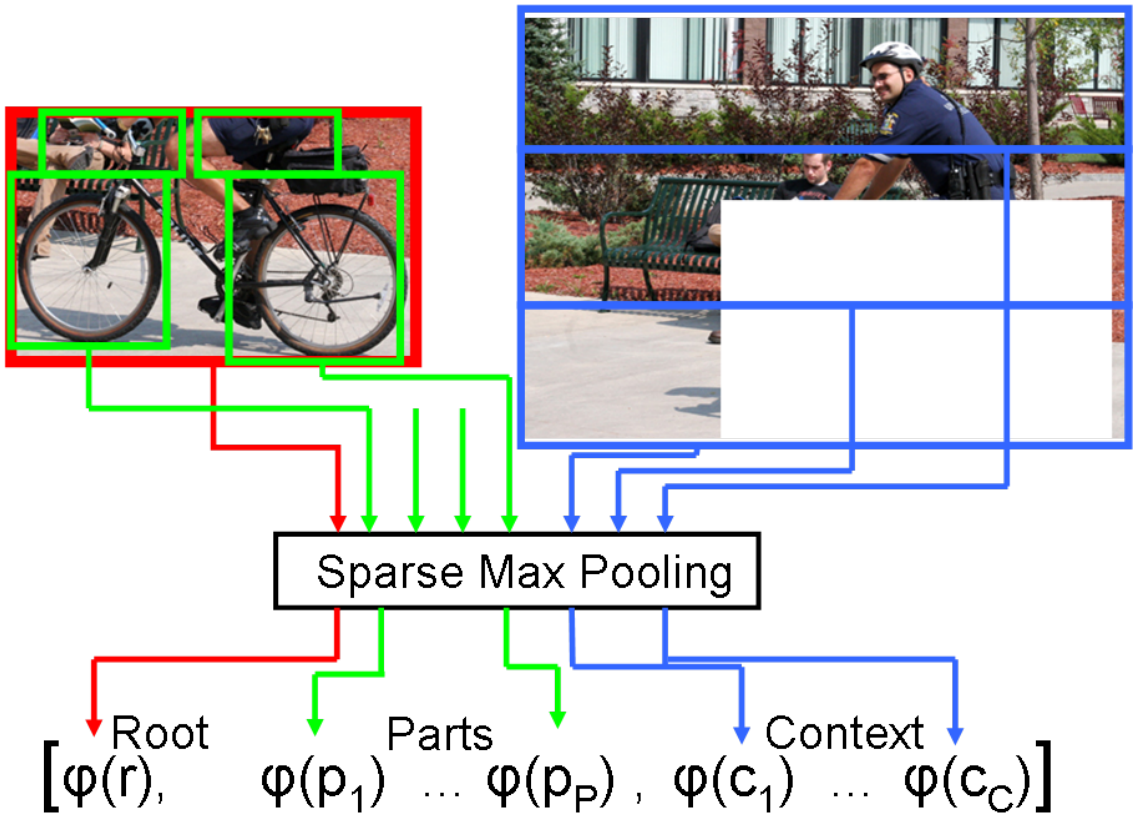
System Overview



Sparse Max Pooling Model



Sparse Max Pooling Model



$O = [r, p_1, ..., p_P, c_1, ..., c_C]$: A candidate object configuration

$$f(O) = \omega_r \phi(r) + \sum_{i=1}^{P} \omega_{p_i} \phi(p_i) + \sum_{i=1}^{C} \omega_{c_i} \phi(c_i)$$

UCDAVIS COLLEGE OF ENGINEERING

Sparse Max Pooling Model

Visual words: $V = [v_1, ..., v_{|V|}]$, where $v_i \in \mathfrak{R}^{128}$ is a cluster center in SIFT space.

Local features falling into a window: $F = f_i{}_{i=1}^{|F|}$ where $f_i \in \mathfrak{R}^{128}$ is a SIFT descriptor.

Each feature f_i is quantized into a $|V|$-dimensional sparse vector s_i that approximates f_i using some existing sparse coding algorithm and the dictionary V, that is, $f_i \approx s_i V$.

UCDAVIS COLLEGE OF ENGINEERING

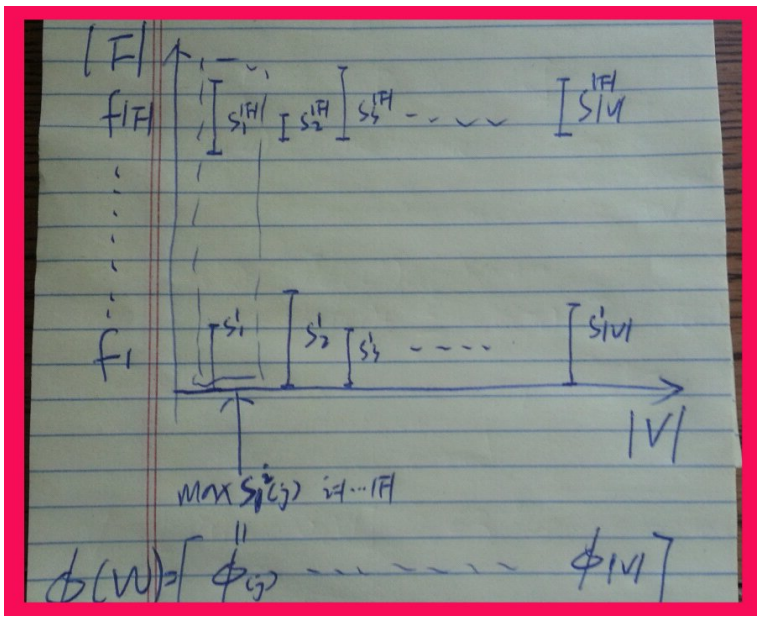
Sparse Max Pooling Model

Taking each encoding s_i as input, the SMP representation of W is given by:

$\phi(W) = [\phi^1, ..., \phi^{|V|}]$, *where*

$\phi^j = max(s_i(j)), i = 1, ..., |F|,$

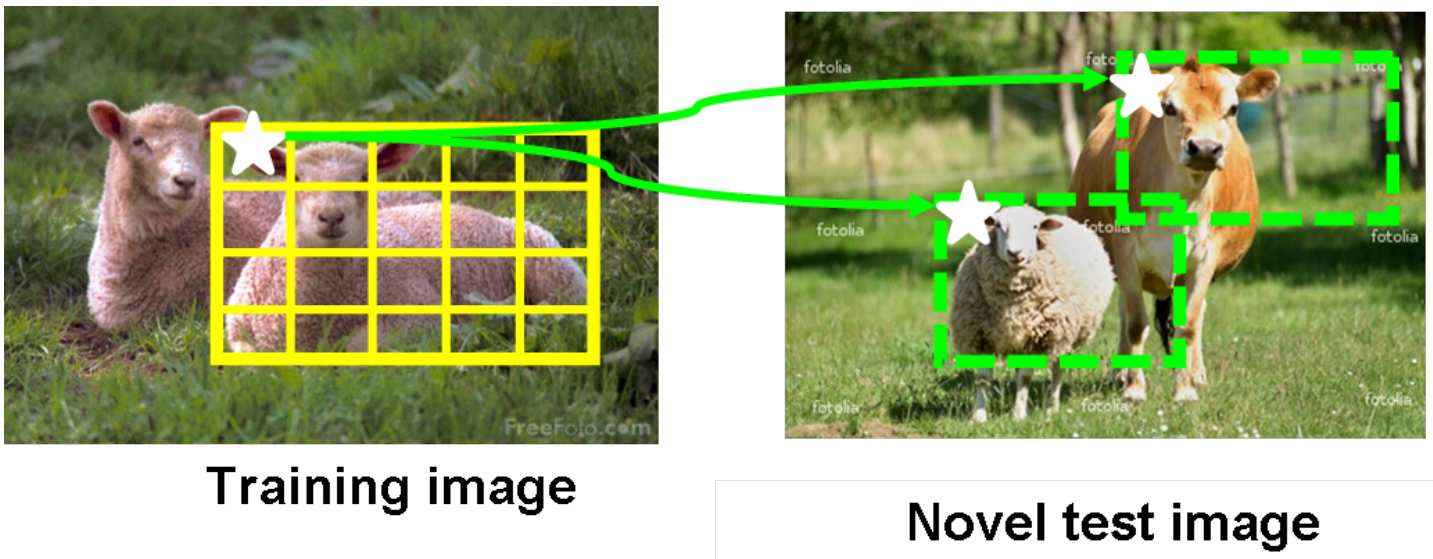
Normalize $\phi(W)$ by its L_2 norm.



Part-based Representation + Linear Classifier = Good trade-off between model complexity and accuracy.

Generating Candidate Root Windows

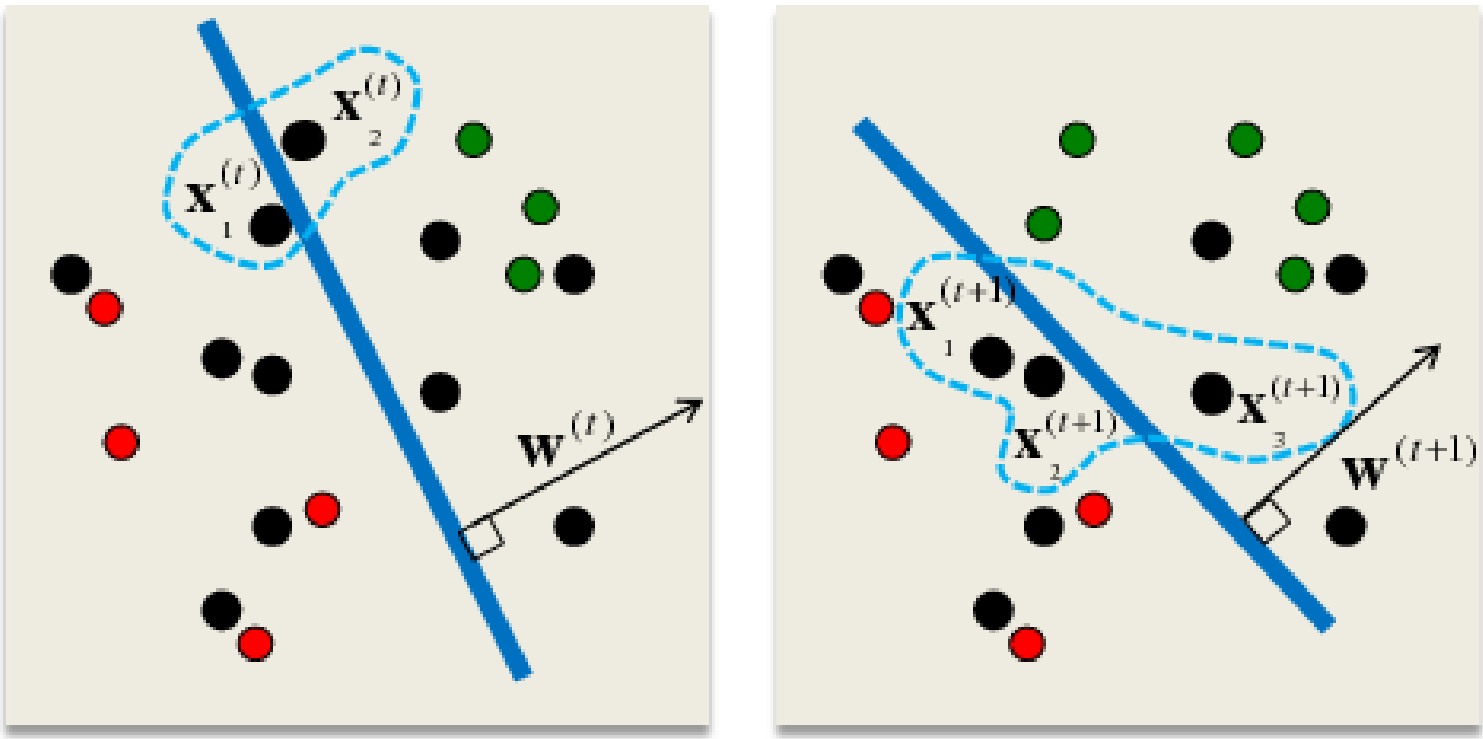
Know the representation and scoring function. How to generate the candidates?



Take 3000 top-ranked boxes from each unlabeled image (vs up to 10^5 boxes if using sliding window)

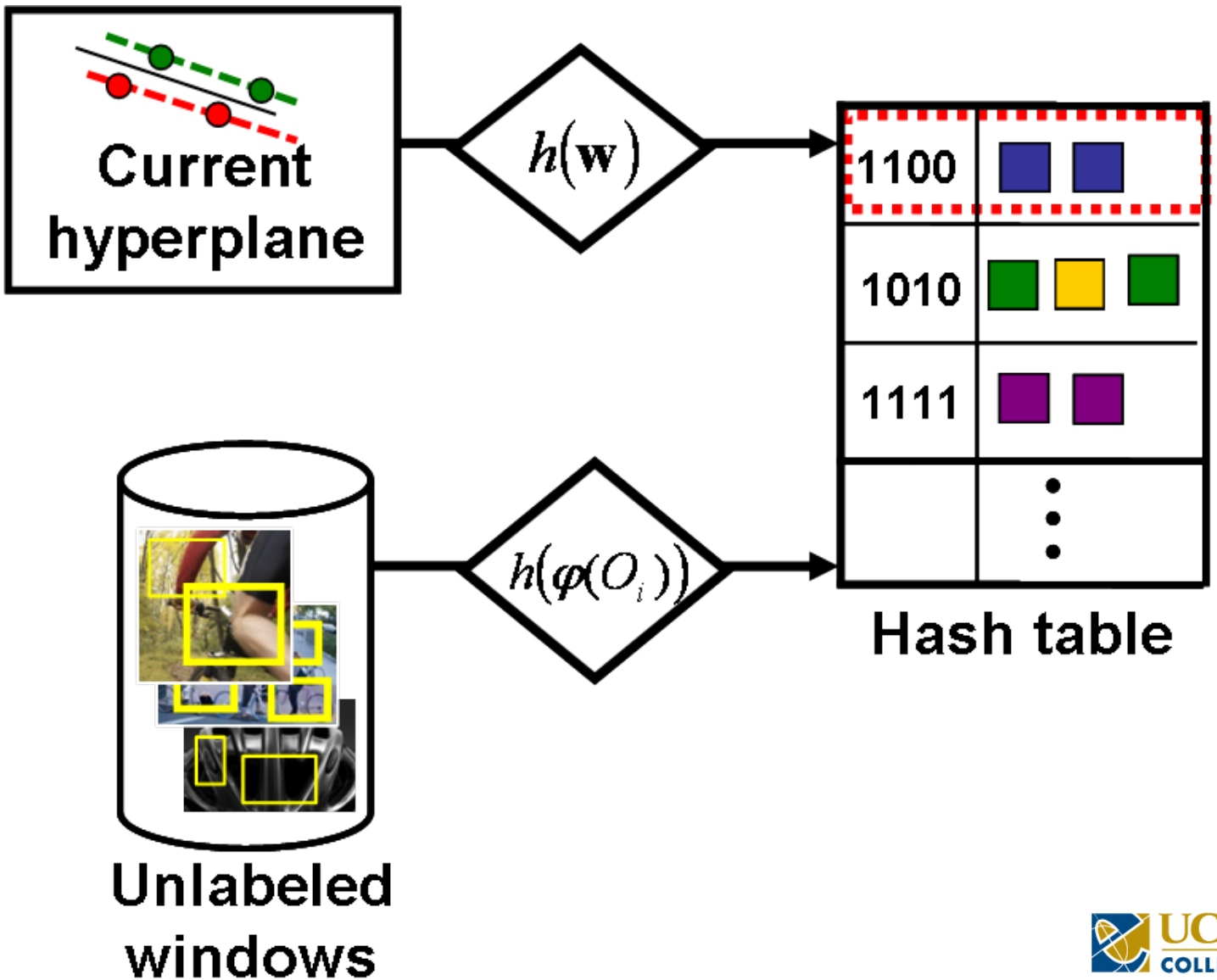
Large-scale Active Selection

Repeatedly query annotators to label the **most uncertain examples** in a massive pool of unlabeled data.



Large-scale Active Selection

Using Hyperplane Hashing Function:



Large-scale Active Selection

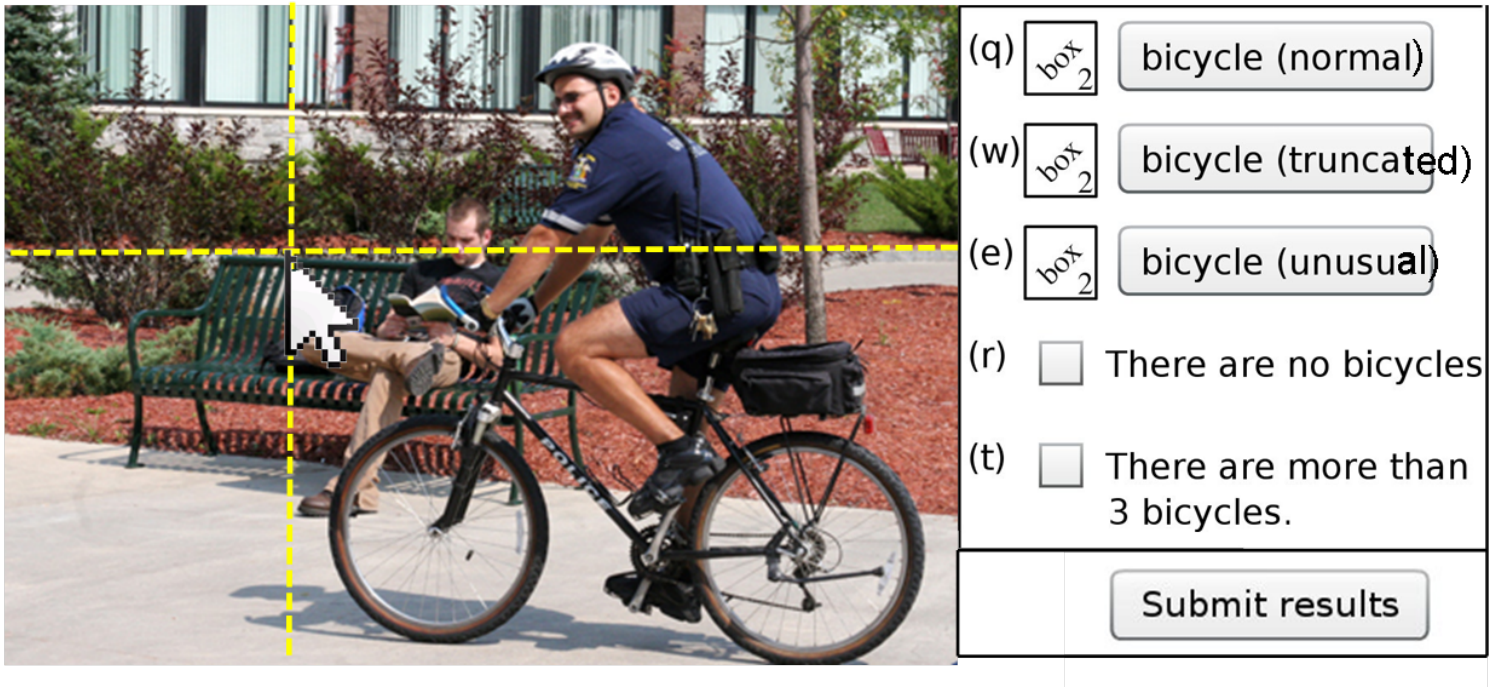
$$h_H(z) = \begin{cases} h_{u,v}(\phi(O_i), \phi(O_i)), \text{if } z \text{ is a database vector} \\ h_{u,v}(\omega, -\omega), \text{if } z \text{ is a query hyperplane} \end{cases}$$

where the component function is defined as:

$$h_{\mathbf{u},\mathbf{v}}(\mathbf{a}, \mathbf{b}) = [sign(\mathbf{u}^T\mathbf{a}), sign(\mathbf{v}^T\mathbf{b})] \quad (1)$$

Evaluate scores for the returns of the hash functions. Choose top T images.
Only need to evaluate the examples that fall into a particular hash bucket-typically less than 0.1% of the total number of unlabeled examples.

Online Annotation Requests

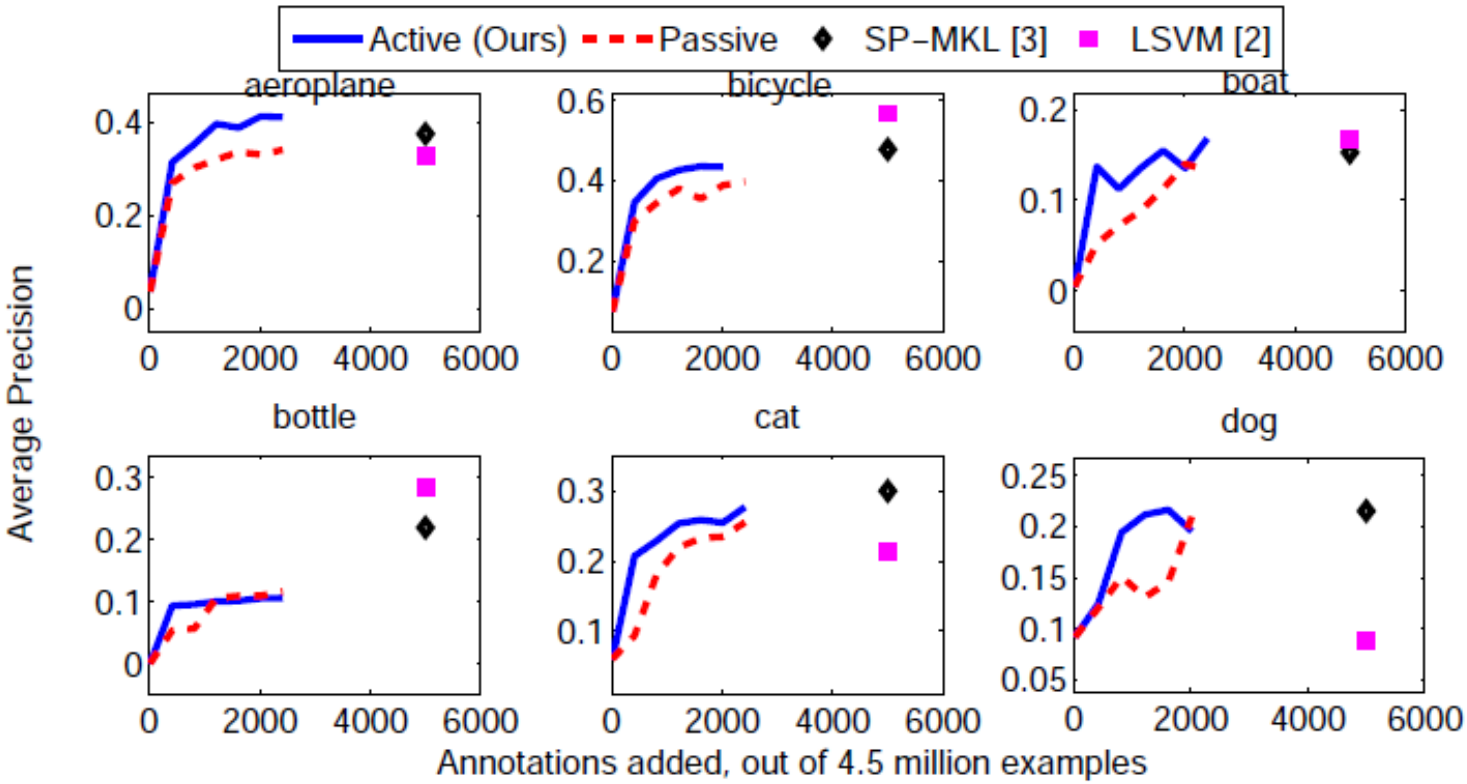


- Post selected images on Mechanical Turk;
- Provide multiple options to avoid incorrect boxes;
- Post same image to multiple 10 annotators and use mean shift for consensus.

UCDAVIS COLLEGE OF ENGINEERING

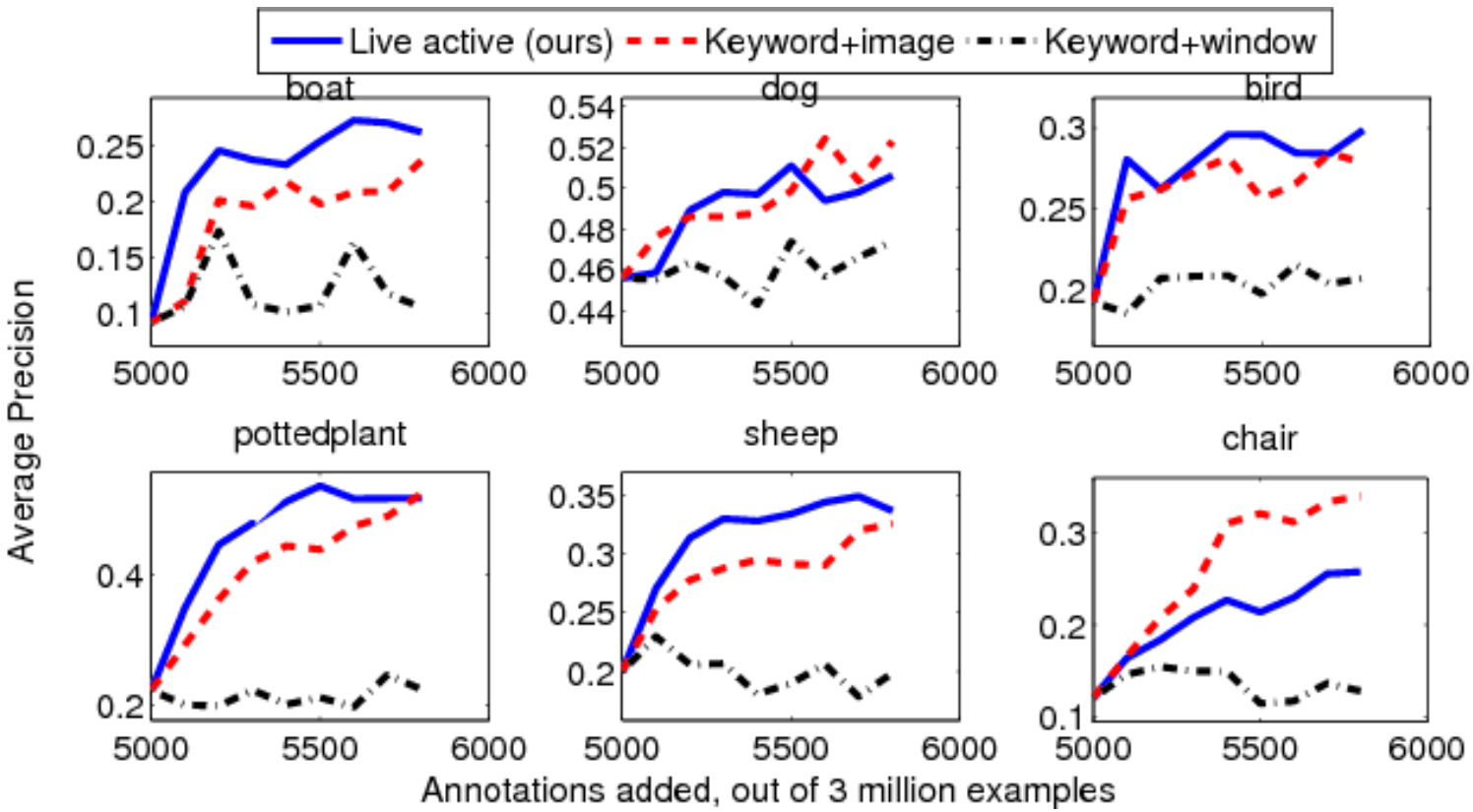
Results

Active selection outperforms passive baseline and we obtain close to state-of-art results using only one third of the data.



Results

We obtain dramatic improvements for most categories and active selection is better for 4/6 categories.



Results

	aeroplane	bird	boat	cat	dog	sheep	sofa	train
Ours	**48.4**	**15.8***	**18.9***	**30.7**	**25.3***	**28.8**	**33.0**	**47.7**
Previous best	37.6	15.3	16.8	30.0	21.5	23.9	28.5	45.3

Table 2. Categories for which our method yields the best AP on PASCAL VOC 2007, compared to any result we found in the literature. (*means extra Flickr data automatically obtained by our system was used to train.)

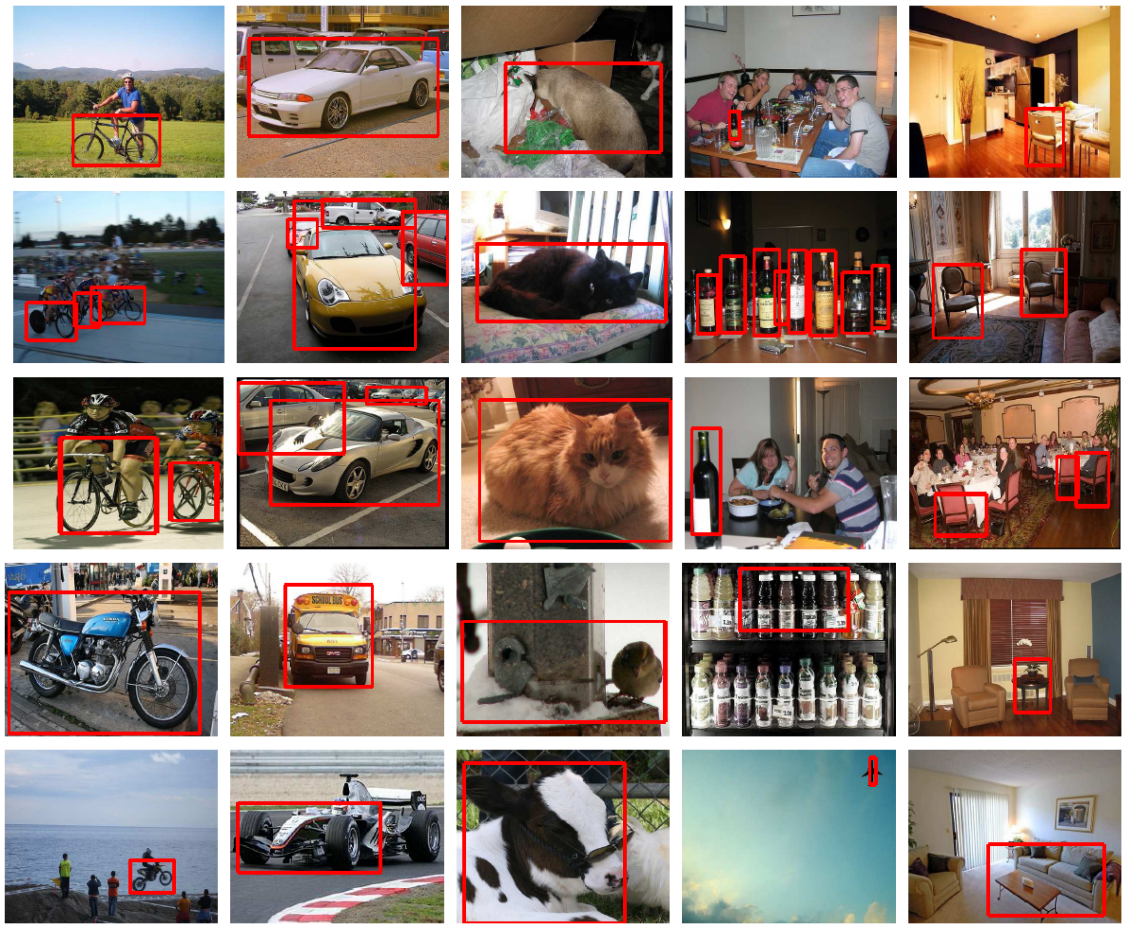
UCDAVIS COLLEGE OF ENGINEERING

Conclusion

- A novel efficient part-based linear detector that provides excellent performance;
- A jumping window and hashing scheme suitable for the proposed detector that retrieves relevant instances among millions of candidates;
- The first active learning results for which both image data and annotations are automatically obtained, with minimal involvement from vision experts.

Questions?

most contents referred to the project website:
http://vision.cs.utexas.edu/projects/livelearning/



Backup Slides: Why A Linear Model?

Backup Slides: Why A Linear Model?

- SVM training requires time linear in # training examples;

Backup Slides: Why A Linear Model?

- SVM training requires time linear in # training examples;
- Classification of novel instances requires constant time;

Backup Slides: Why A Linear Model?

- SVM training requires time linear in # training examples;
- Classification of novel instances requires constant time;
- Exact incremental classifier updates are possible;

Backup Slides: Why A Linear Model?

- SVM training requires time linear in # training examples;
- Classification of novel instances requires constant time;
- Exact incremental classifier updates are possible;
- Hash functions enable sub-linear time search to map a query-hyperplane to its nearest points.

Backup Slides: Where did the inial labeled images come from?

Backup Slides: Where did the inial labeled images come from?

- Trained using deformable part model;

Backup Slides: Where did the inial labeled images come from?

- Trained using deformable part model;
- Can be requested once directly from annotators;