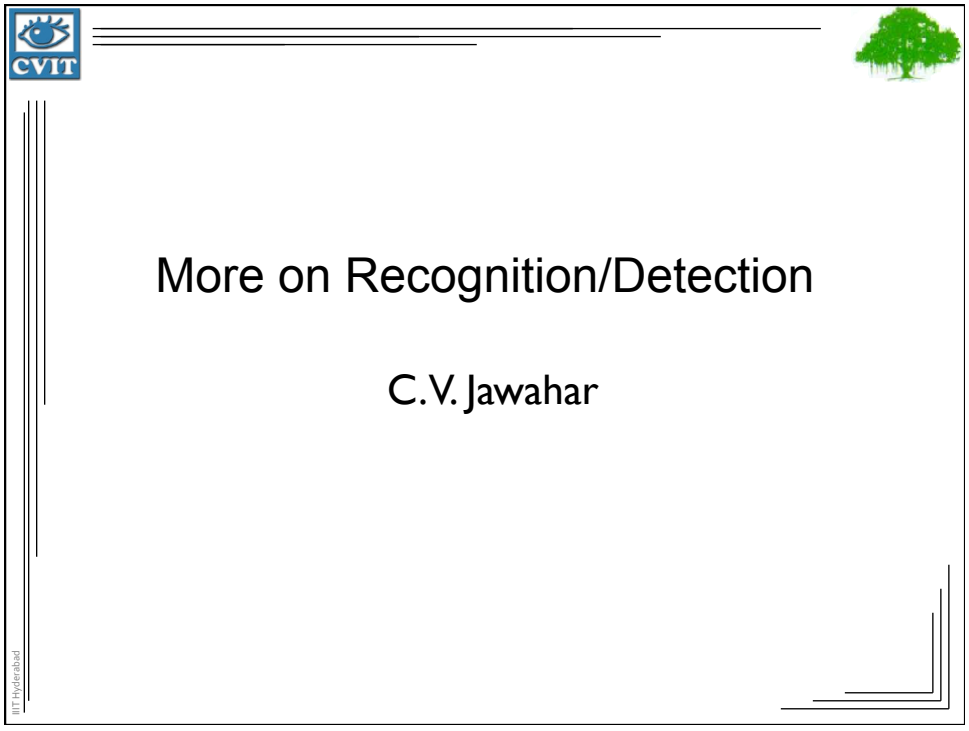
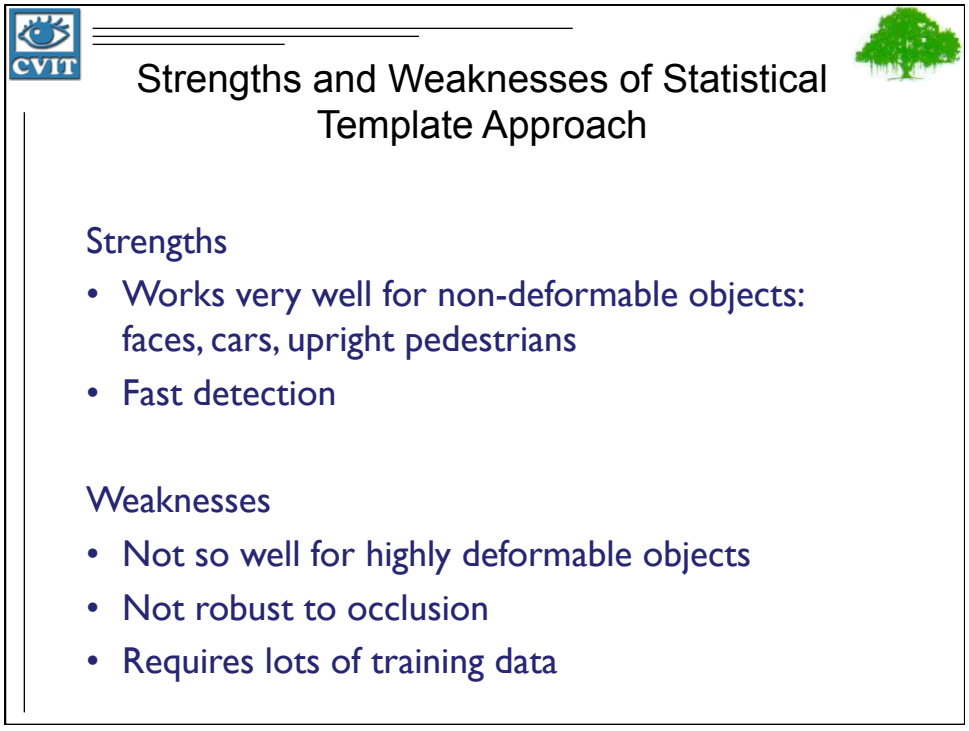

CVIT
More on Recognition/Detection
C.V. Jawahar
IIIT Hyderabad


CVIT
Strengths and Weaknesses of Statistical Template Approach
Strengths
• Works very well for non-deformable objects: faces, cars, upright pedestrians
• Fast detection
Weaknesses
• Not so well for highly deformable objects
• Not robust to occlusion
• Requires lots of training data

Tricks of the trade

- Details in feature computation really matter
  - E.g., normalization in Dalal-Triggs improves detection rate by 27% at fixed false positive rate
- Template size
  - Typical choice is size of smallest detectable object
- "Jittering" to create synthetic positive examples
  - Create slightly rotated, translated, scaled, mirrored versions as extra positive examples
- Bootstrapping to get hard negative examples
  1. Randomly sample negative examples
  2. Train detector
  3. Sample negative examples that score > -1
  4. Repeat until all high-scoring negative examples fit in memory

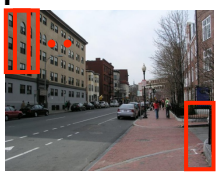
Consumer application: iPhoto 2009

Influential Works in Detection

- Sung-Poggio (1994, 1998) : ~1750 citations
  - Basic idea of statistical template detection (I think), bootstrapping to get "face-like" negative examples, multiple whole-face prototypes (in 1994)
- Rowley-Baluja-Kanade (1996-1998) : ~3400
  - "Parts" at fixed position, non-maxima suppression, simple cascade, rotation, pretty good accuracy, fast
- Schneiderman-Kanade (1998-2000,2004) : ~1700
  - Careful feature engineering, excellent results, cascade
- Viola-Jones (2001, 2004) : ~11,000
  - Haar-like features, Adaboost as feature selection, hyper-cascade, very fast, easy to implement
- Dalal-Triggs (2005) : ~3250
  - Careful feature engineering, excellent results, HOG feature, online code
- Felzenszwalb-Huttenlocher (2000): ~2500
  - Efficient way to solve part-based detectors
- Felzenszwalb-McAllester-Ramanan (2008)? ~1500
  - Excellent template/parts-based blend

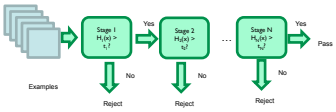
Things to remember

- Sliding window for search



- Features based on differences of intensity (gradient, wavelet, etc.)
  - Excellent results require careful feature design

- Boosting for feature selection
- Integral images, cascade for speed

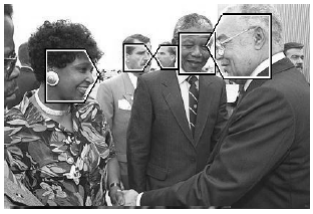


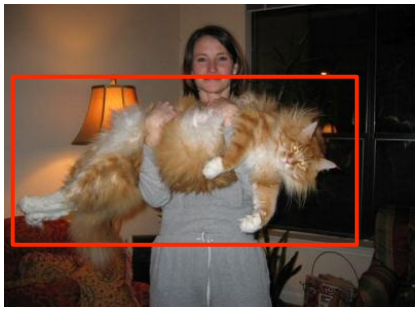
- Bootstrapping to deal with many, many negative examples

Goal: Detect all instances of objects

Cars

Faces

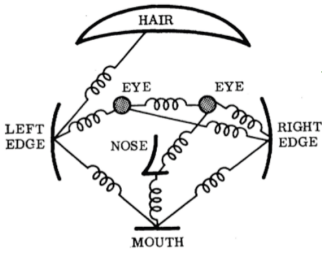


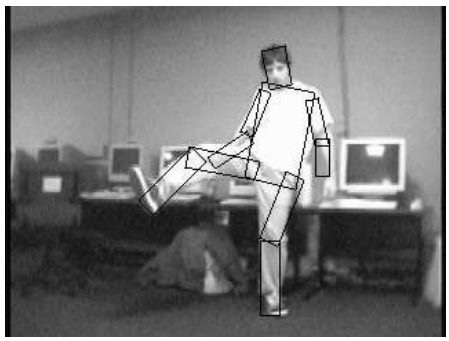


Cats

Object models: Challenging ones ..

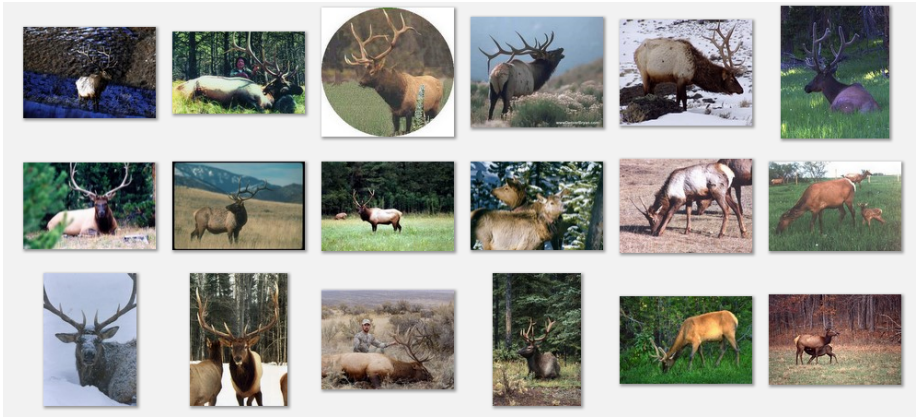
- Articulated parts model
  - Object is configuration of parts
  - Each part is detectable





Images from Felzenszwalb

Deformable objects



Images from Caltech-256

Slide Credit: Duan Tran

Deformable objects

Images from D. Ramanan's dataset

Slide Credit: Duan Tran

Compositional objects

Parts-based Models

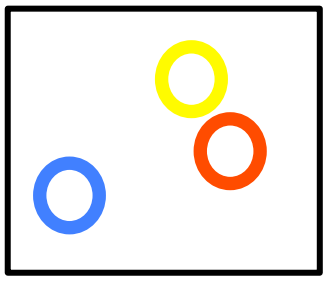
Define object by collection of parts modeled by

1. Appearance
2. Spatial configuration

Slide credit: Rob Fergus

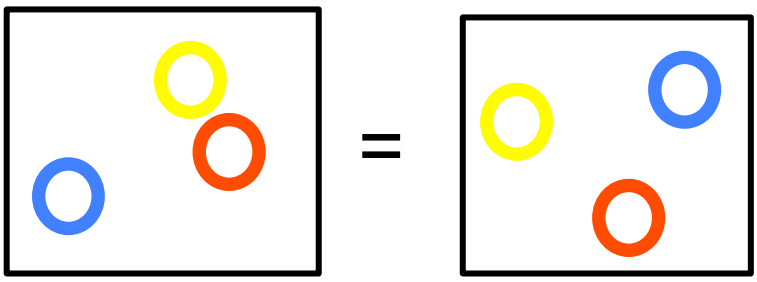
How to model spatial relations?

- One extreme: fixed template



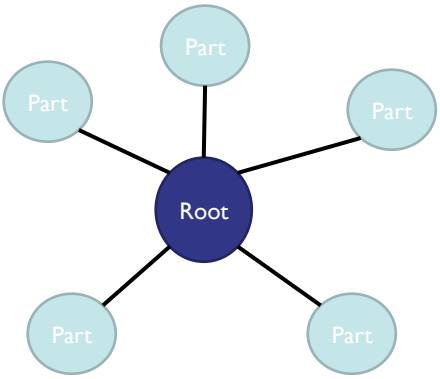
How to model spatial relations?

- Another extreme: bag of words



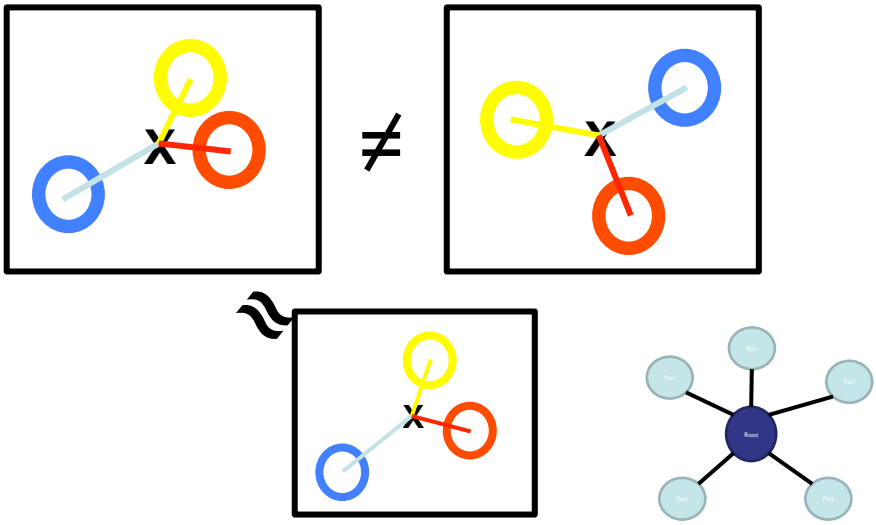
How to model spatial relations?

- Star-shaped model



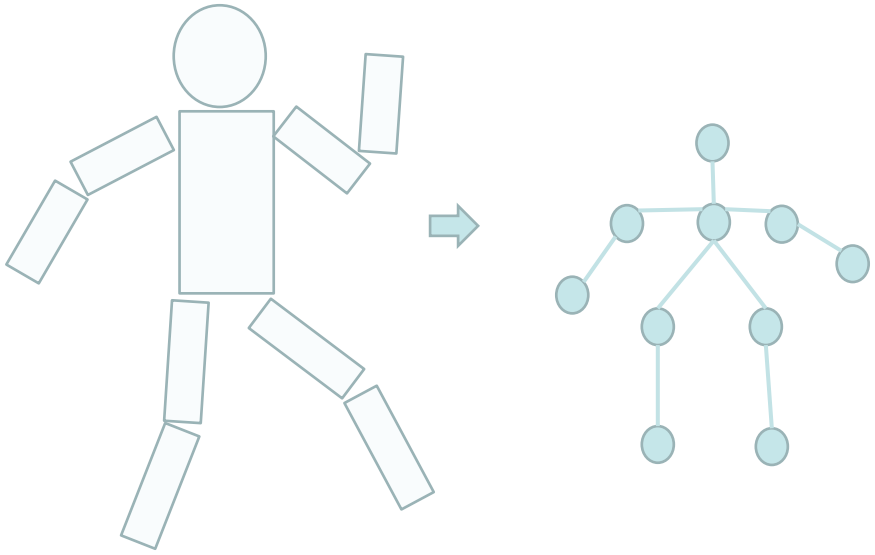
How to model spatial relations?

- Star-shaped model



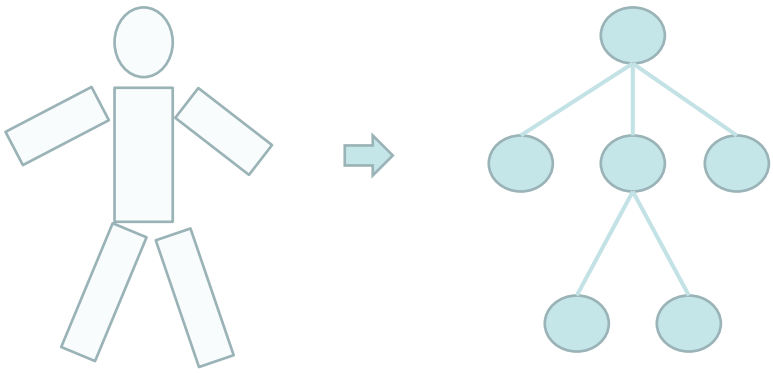
How to model spatial relations?

- Tree-shaped model

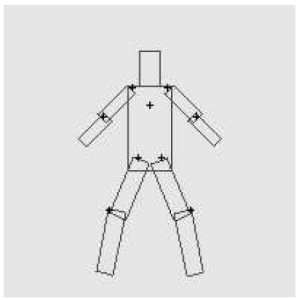


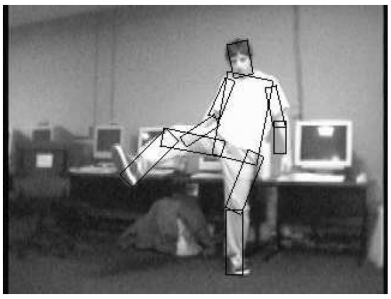
How to model spatial relations?

- Tree-shaped model



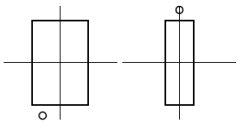
Pictorial Structures Model



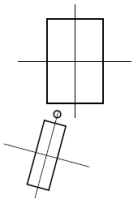


Part = oriented rectangle

Spatial model = relative size/orientation



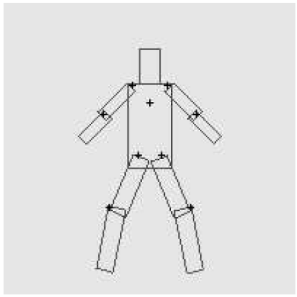
a



b

Felzenszwalb and Huttenlocher 2005

Pictorial Structures Model



$$P(L|I,\theta) \propto \left(\prod_{i=1}^{n} p(I|l_i, u_i) \prod_{(v_i,v_j)\in E} p(l_i, l_j|c_{ij}) \right)$$

Appearance likelihood

Geometry likelihood

Modeling the Appearance

- Any appearance model could be used
  - HOG Templates, etc.
  - Here: rectangles fit to background subtracted binary map
- Can train appearance models independently (easy, not as good) or jointly (more complicated but better)

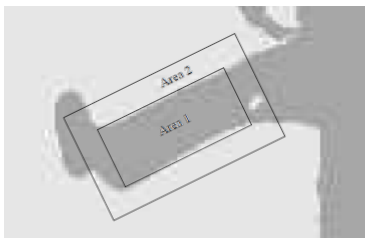
$$P(L|I,\theta) \propto \left(\prod_{i=1}^{n} p(I|l_i, u_i) \prod_{(v_i,v_j)\in E} p(l_i, l_j|c_{ij}) \right)$$

Appearance likelihood

Geometry likelihood

Part representation

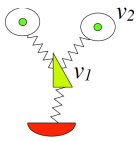
- Background subtraction



Pictorial structures model

Optimization is tricky but can be efficient

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

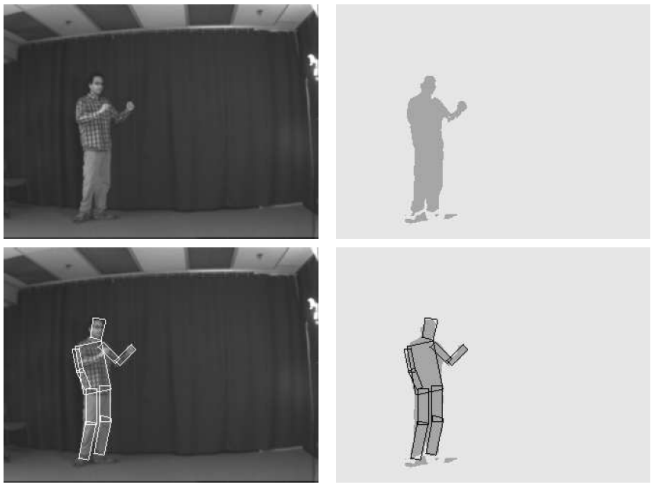


- For each l_1, find best l_2:

$$\text{Best}_2(l_1) = \min_{l_2} \left[m_2(l_2) + d_{12}(l_1, l_2) \right]$$

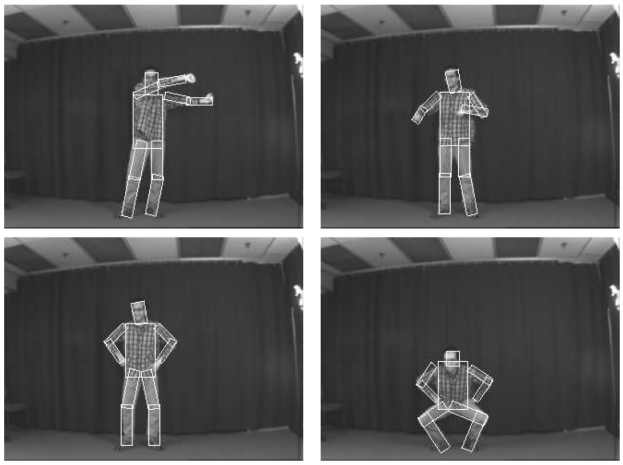
- Remove v_2, and repeat with smaller tree, until only a single part
- For k parts, n locations per part, this has complexity of $O(kn^2)$, but can be solved in ~$O(nk)$ using generalized distance transform

Results for person matching



24

Results for person matching



25

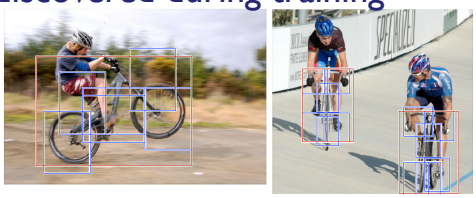
Enhanced pictorial structures

BMVC 2009

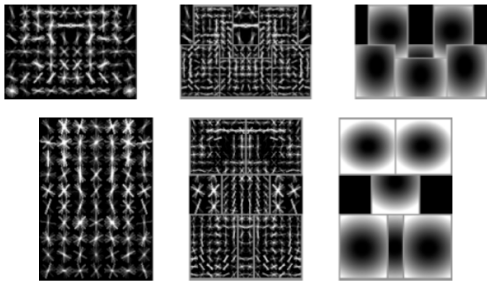
Deformable Latent Parts Model

Useful parts discovered during training

Detections



Template Visualization

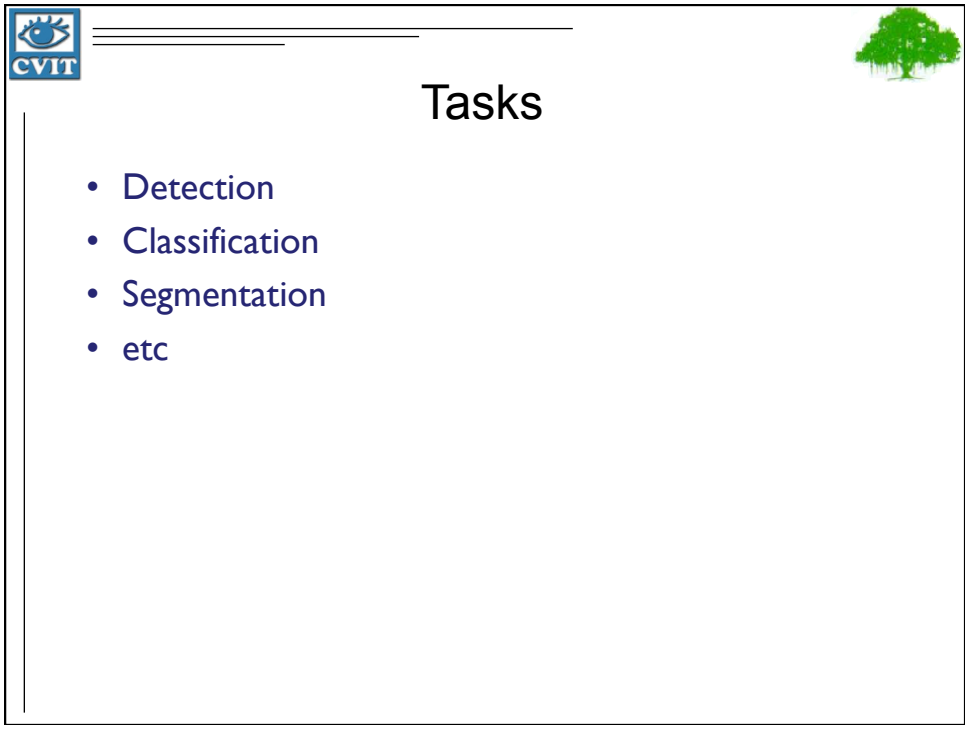


root filters coarse resolution | part filters finer resolution | deformation models

Felzenszwalb et al. 2008

Data Sets: The benchmarks that set goals

- Caltech
- PASCAL
- ImageNet
- Scene-15
- Sceene-67
- SUN



Scene category dataset

Fei-Fei & Perona (2005), Oliva & Torralba (2001)

http://www-cvr.ai.uiuc.edu/ponce_grp/data

Multi-class classification results (100 training images per class)

	Weak features (vocabulary size: 16)		Strong features (vocabulary size: 200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3

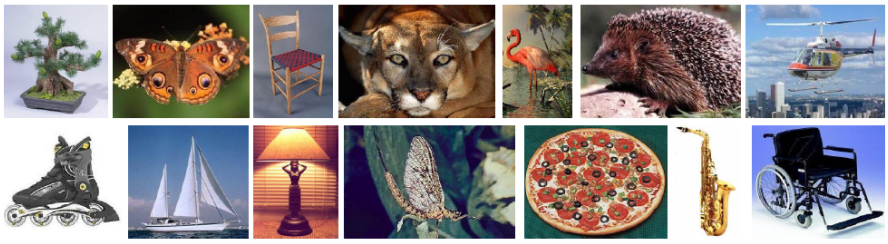
Fei-Fei & Perona: 65.2%

6

Caltech101 dataset

Fei-Fei et al. (2004)

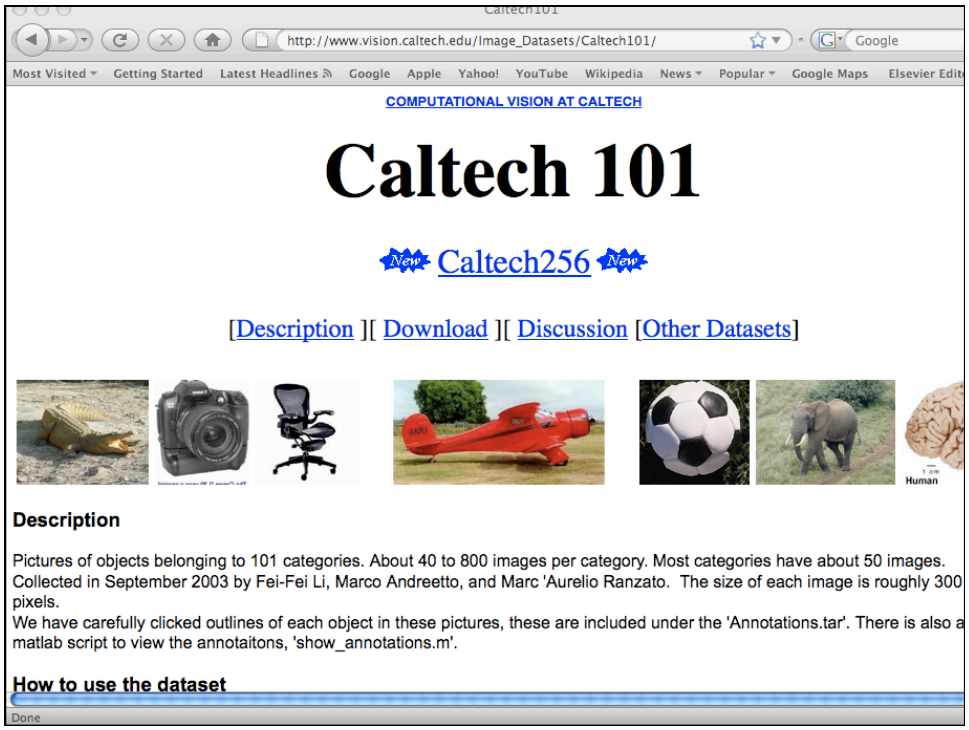
http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html



Multi-class classification results (30 training images per class)

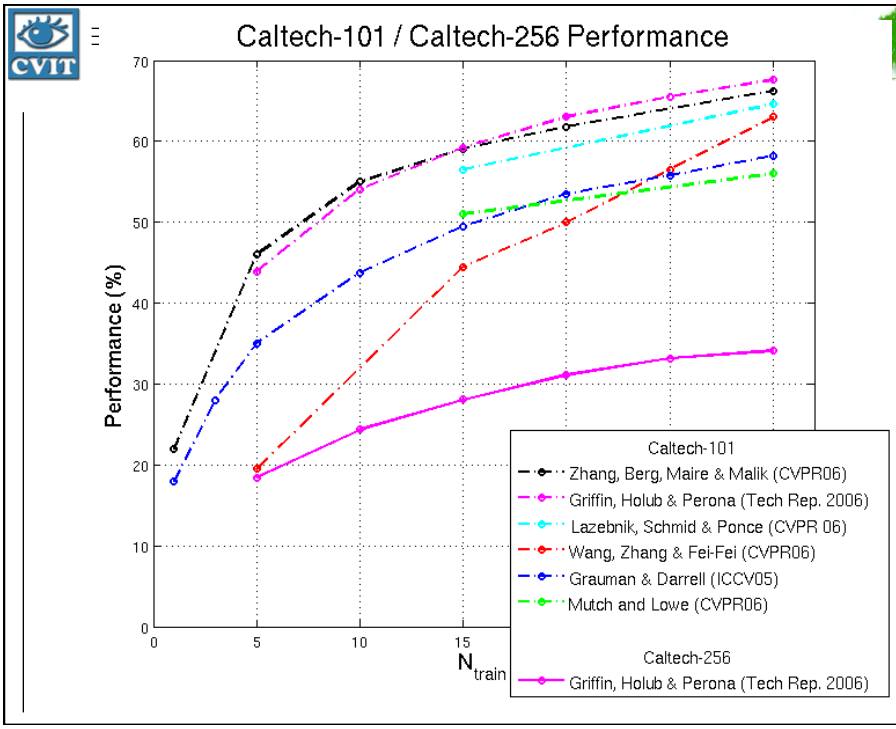
	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

9

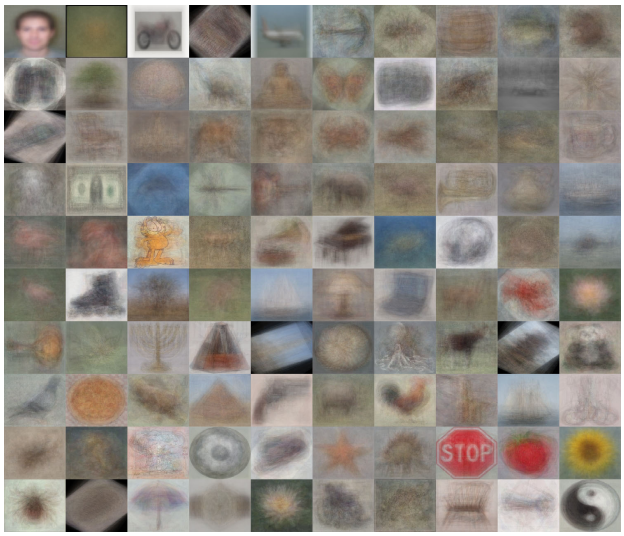


Caltech-101: Drawbacks

- Smallest category size is 31 images: $N_{train} \leq 30$
- Too easy?
  - left-right aligned
  - Rotation artifacts
  - Soon will saturate performance



When do statistical templates make sense?



Caltech 101 Average Object Images

The PASCAL Visual Object Classes Challenge 2007

The twenty object classes that have been selected are:

Person: person

Animal: bird, cat, cow, dog, horse, sheep

Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train

Indoor: bottle, chair, dining table, potted plant, sofa, tv/ monitor

TRECVID: Video Retrieval

- **Problem**: From a large (> 200 Hrs) of video (more than 1L shots, 2.5L key-frames), retrieve shots containing specific objects, scenes or actions. Increasing every year.

- Eg. Person singing, mountains, telephones etc.
- A separate training data is also provided for the development and validation.

Person Riding Bicycle: Results

SVM Classifier, Intersection Kernel, PHOW feature

Cityscape: Results

SVM Classifier, Intersection Kernel, PHOW feature

Female Face Close-up: Results

SVM Classifier, Intersection Kernel, PHOW feature

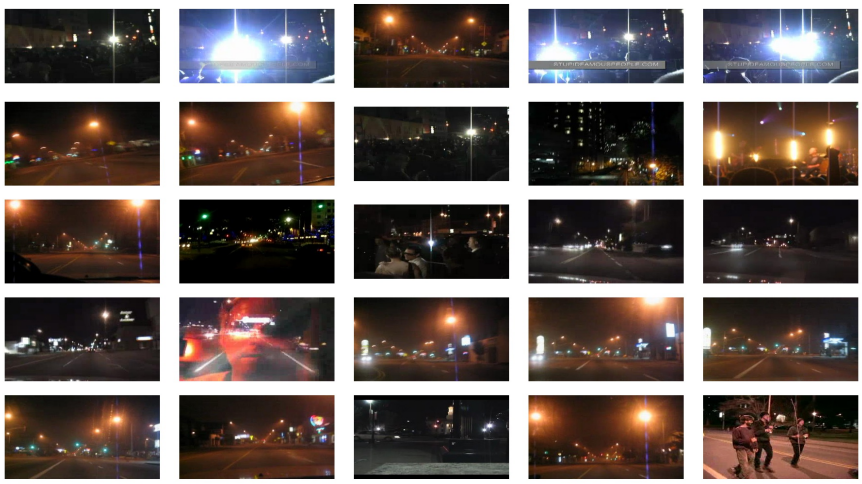
Airplanes: Results

SVM Classifier, Chi-2 Kernel, Kernel Maps, PHOW feature

Demonstration or Protest: Results

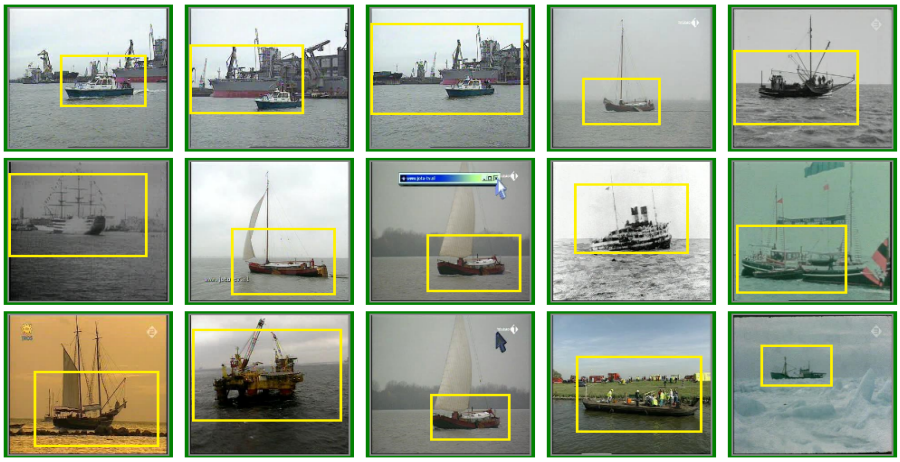
SVM Classifier, Chi-2 Kernel, Kernel Maps, PHOW feature

Nighttime: Results



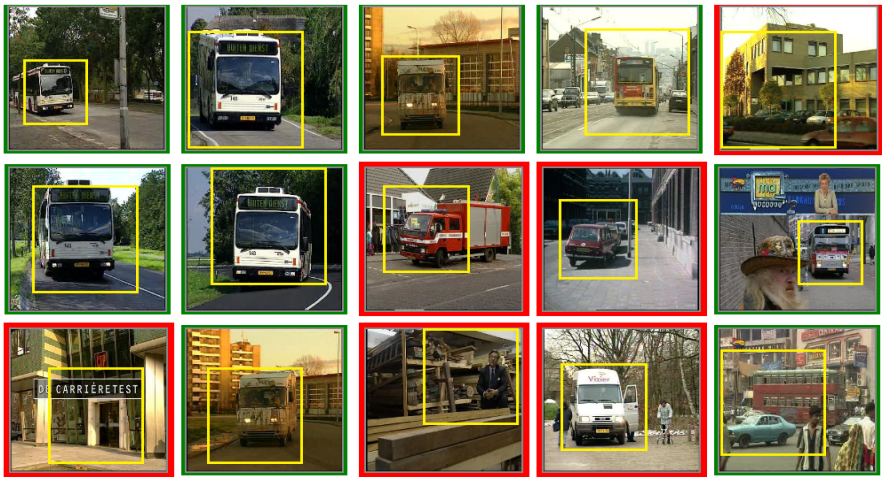
SVM Classifier, Chi-2 Kernel, Kernel Maps, PHOW feature

Boat-Ship: Results



Random Forest Classifier, PHOW feature, Sliding Window Search, Non-maxima suppression

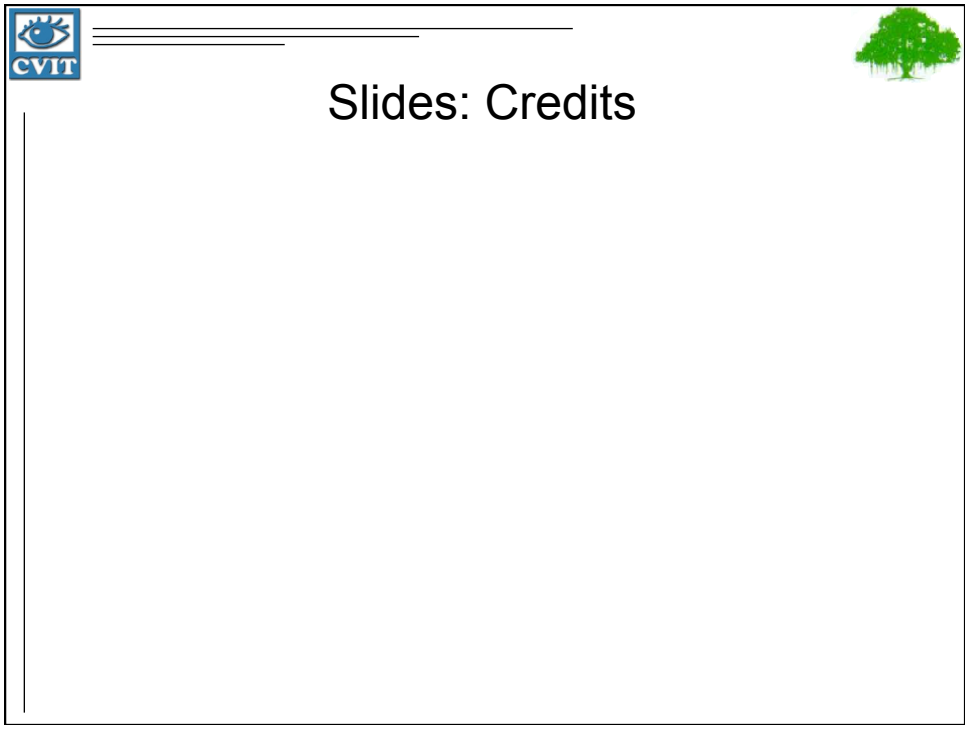
Bus: Results



Random Forest Classifier, PHOW feature, Sliding Window Search, Non-maxima suppression

What enabled this success?

- Modern Features
  - Invariant to popular transformations
  - Capable of capturing local and global (shape, colour, texture) characteristics reliably
- Machine Learning
  - Learn from examples rather than handcoding
  - New algorithms: effective, efficient
- Realistic Data
  - Huge amount; partly annotated
  - Regular competitions


CVIT
Slides: Credits