

Dataset Issues in Object Recognition

J. Ponce, TL Berg, M. Everingham, DA Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, BC Russell, A. Torralba, CKI Williams, J. Zhang and A. Zisserman

Lecture Notes in Computer Science, 2006

Presented by: Elad Meuzuman



The Importance of Image Databases

- Learning visual object models
- Testing the performance of classification, detection and localization algorithms
- Play a key role in the recent resurgence of category-level recognition research
- Common ground for algorithm development and evaluation

Outline

- Existing datasets and lessons learnt from them
- Innovative ways to gather very large, annotated datasets from the WWW
- Recommendations for future datasets

Bonus:

- A comparison of affine region detectors and descriptors

Drawbacks of Current Datasets

- Limited range of variability:
 - Viewpoints and orientations tend to be similar
 - Sizes and image positions are normalized
 - One instance of an object per image
 - Little or no occlusion and background clutter
- Algorithms may exploit them (no need to scale or rotation invariance)
- Not sufficiently challenging
- Need for new datasets with more realistic and less restrictive image conditions: multiple object class instances within a single image, with partial occlusion and truncation

Caltech 101

- Pictures of objects belonging to 101 categories
- About 40 to 800 images per category. Most categories have about 50 images
- The size of each image is roughly 300 x 200 pixels
- Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato
- Became a de facto standard for evaluating algorithms for multi-class category-level recognition
- Inter-class variability but no intra-class variability
- No clutter, the objects are centered, stereotypical pose

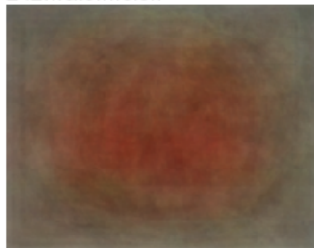
The Caltech 101 Average Image



PASCAL 2006 & Caltech 256 Average Images



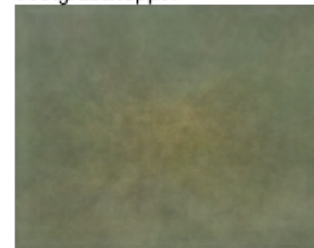
242.watermelon



171.refrigerator



093.grasshopper



162.picnic-table



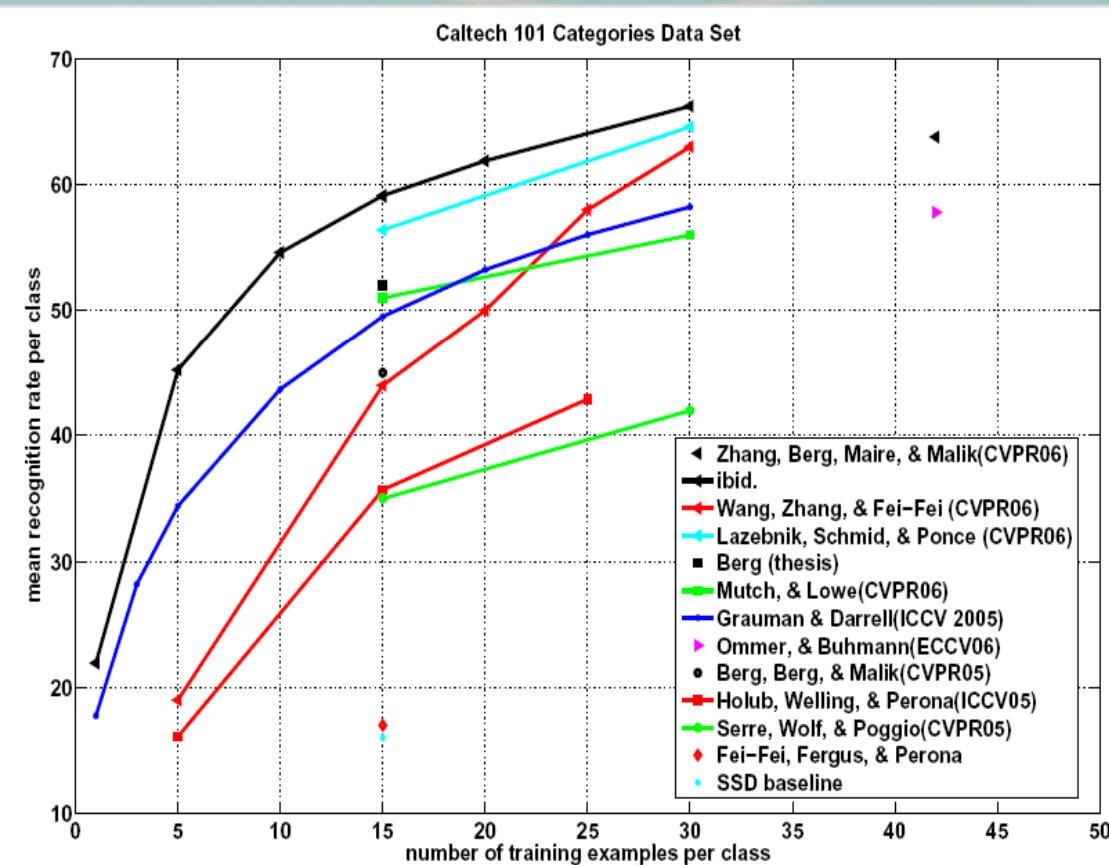
014.blimp



257.clutter



Performance on Caltech 101



- Performance improves with the number of training samples
- Algorithm using SVMs as classifiers tend to do well and include the two top performers
- The classification rate steadily improves with time
- Best 3: Totally different models for image categories: a bag of features, a spatial pyramid, and Bayesian model
- The improvement might be thanks to learning technique and not thanks to the model itself
- There is fine-tuning

The PASCAL Visual Object Classes Challenge

- Yearly challenge (first challenge ran at 2005)

Objectives:

- To compile a standardized collection of object recognition databases
 - To provide standardized ground truth object annotations across all databases
 - To provide a common set of tools for accessing and managing the database annotations
 - To run a challenge evaluating performance on object class recognition
-
- 2 test sets:
 - images from standard sources (i.e Caltech101 sets)
 - Images from new sources (Google image search, local photographs, etc.) harder with greater variability of scale, pose, background clutter and degree of occlusion



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

PASCAL 2005 Challenge

bike



cars



motorbikes



people



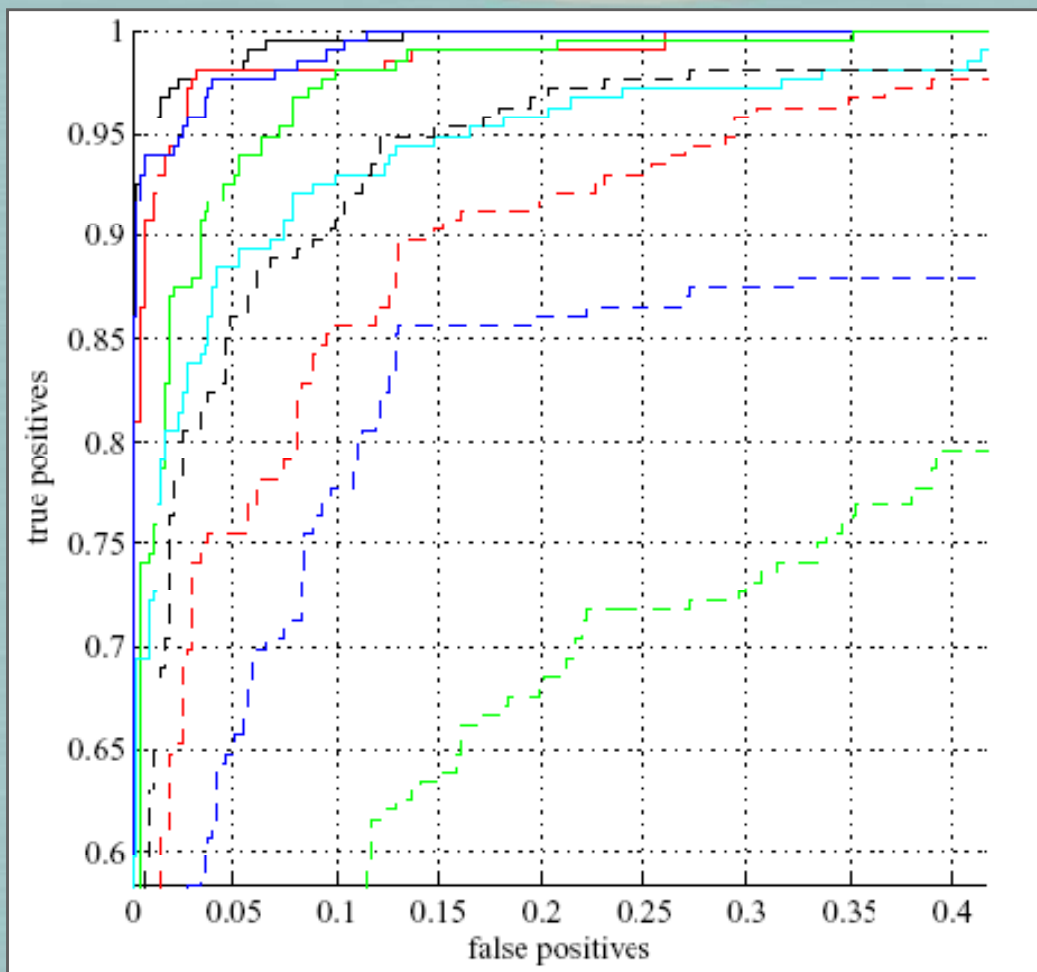
training

test set 1

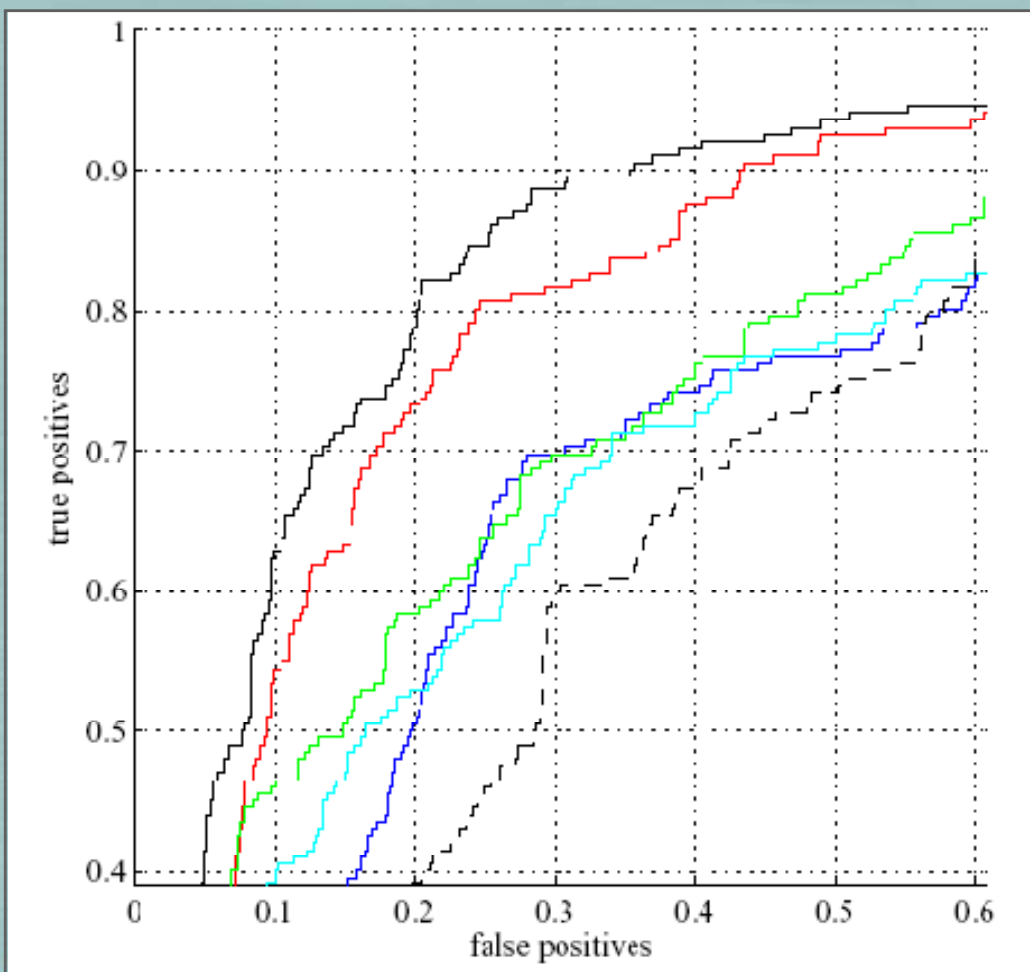
test set 2

PASCAL 2005 Results (motorbikes)

Test set 1



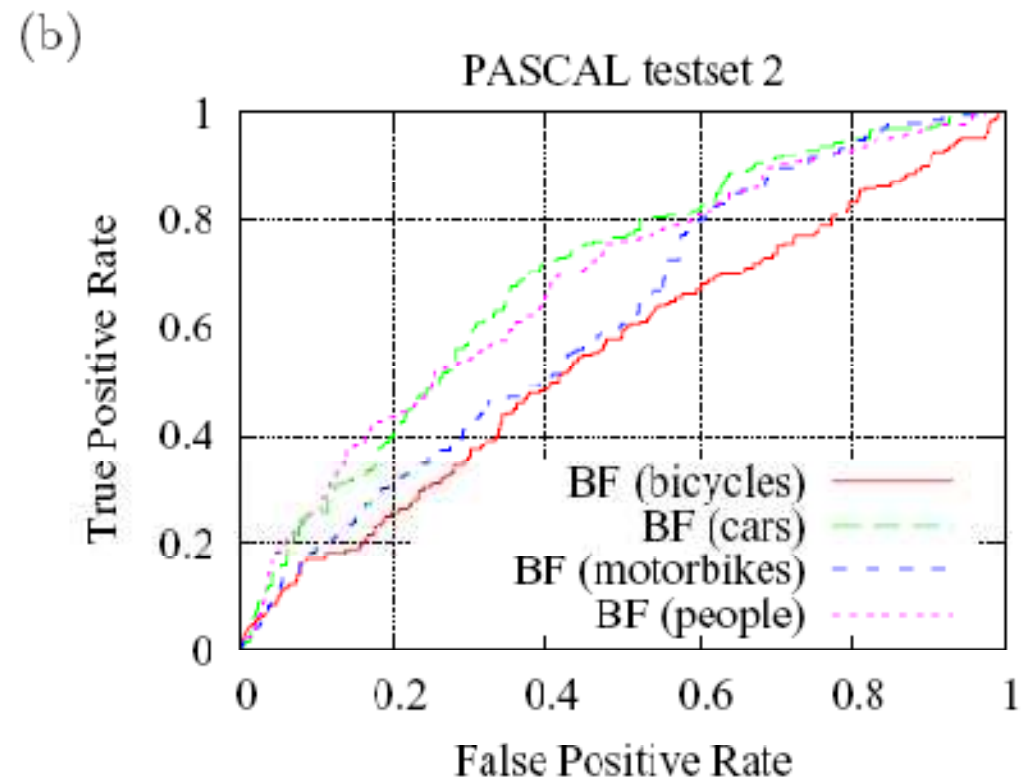
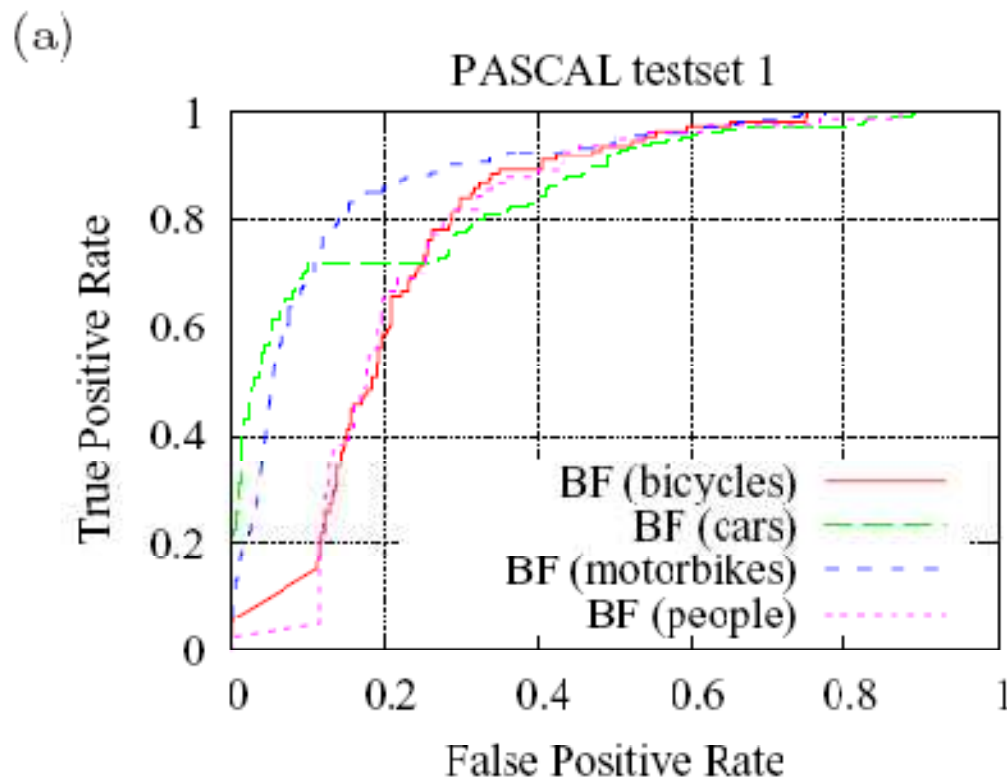
Test set 2



The Importance of Context in Object Recognition Databases

- Zhang, *et al*: Local features and kernels for classification of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhone-Alpes (2005)
- A bag-of-features
- Harris and Laplacian regions, along with their SIFT descriptors
- Support vector machines (SVMs) using the Earth Mover's Distance as a kernel
- uses both foreground and background features
- Foreground features (FF) located within the object region
- Original Background features (BF) are replaced by 2 specially constructed alternative sets: random and constant natural scene (fixed camera observing a natural scene over an extended period of time)

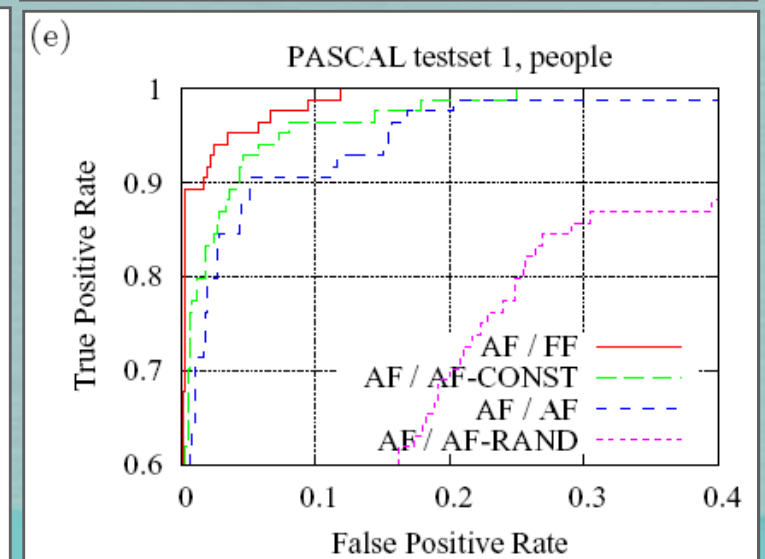
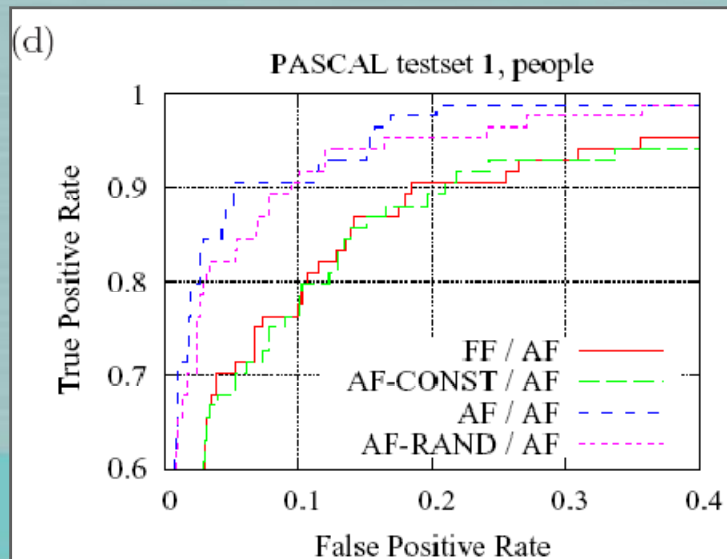
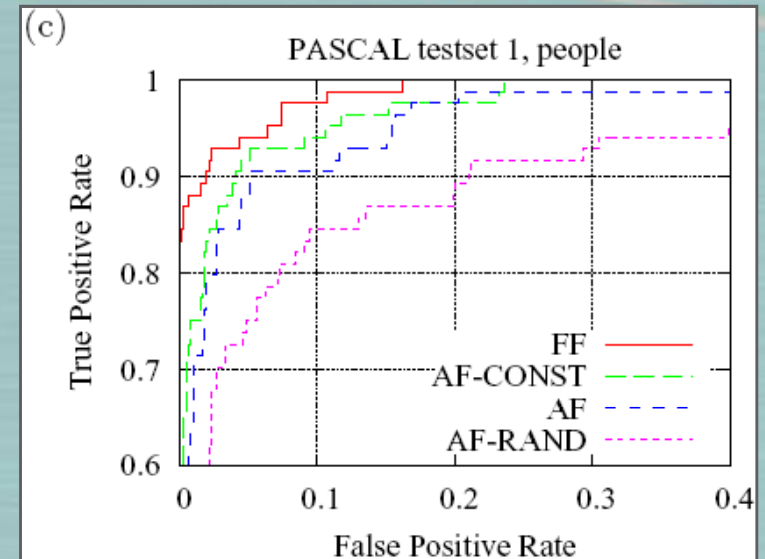
Is it the object or its background, which is recognized?



BF - background features

Performance on Different Combinations of Features

FF - Foreground features
BF - background features
BF-RAND - random scene backgrounds
BF-CONST - constant natural scene backgrounds
AF-CONST - FF + BF-CONST
AF-RAND - FF + BF-RAND
AF - all the features extracted from the original
Training/Testing



The Role of Background Features

- Even if the background has no negligible correlation with the foreground, using both foreground and background features for learning and recognition does not result in better performance
- high performance on datasets with high correlation between foreground and background does not necessarily mean high performance on real images with varying backgrounds
- When the training set has different image statistics than the test set, it is usually beneficial to train on the most difficult dataset available, since the presence of varied backgrounds during training improves the generalization ability of the classifier

Innovative Methods for Acquiring New Datasets

- Web-Based Annotation - building large annotated databases by relying on the collaborative effort of a large population of users:
 - ESP and Peekaboom internet games - “bored human intelligence”
 - LabelMe
- Data Collection as Recognition:
 - Starting from image search – image name and surrounding text
 - Starting from text search

ESP Game

gwap **ESP Game** x Tag a Tune Verbosity Squigl Matchin logged in x


Most Points Today

| | | |
|----|-------------|-------|
| 1 | lubeckm | 101 K |
| 2 | guest100752 | 74 K |
| 3 | 92trebor | 44 K |
| 4 | Tink | 39 K |
| 5 | Anubis | 39 K |
| 6 | Linyclort | 34 K |
| 7 | Nissy | 32 K |
| 8 | regina | 30 K |
| 9 | Christi | 25 K |
| 10 | Chop Suey | 25 K |


score **0**  **ESP Game** time **2:14**
Concentrate...

What do you see?

taboo words
tree
building



guesses
house



BONUS!
5,000 PTS 

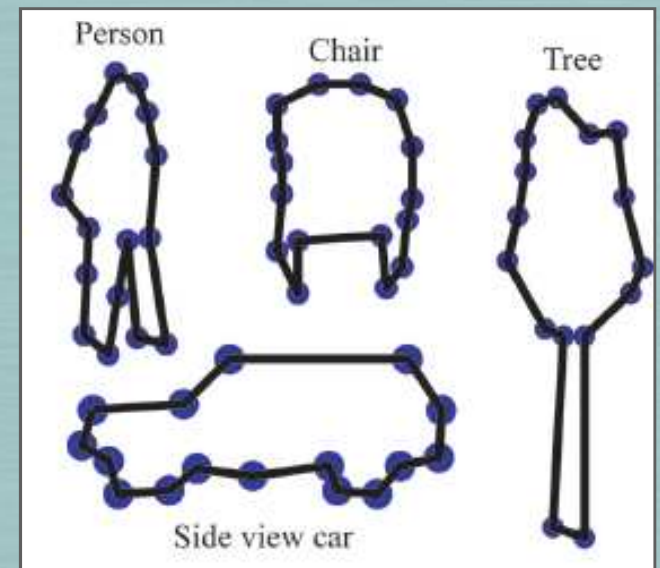
LabelMe



<http://labelme.csail.mit.edu/>

Images: 177,430

Annotated: 52,341



“Our goal is not to provide a new benchmark for computer vision. The goal of the LabelMe project is to provide a dynamic dataset that will lead to new research in the areas of computer vision and computer graphics. ”

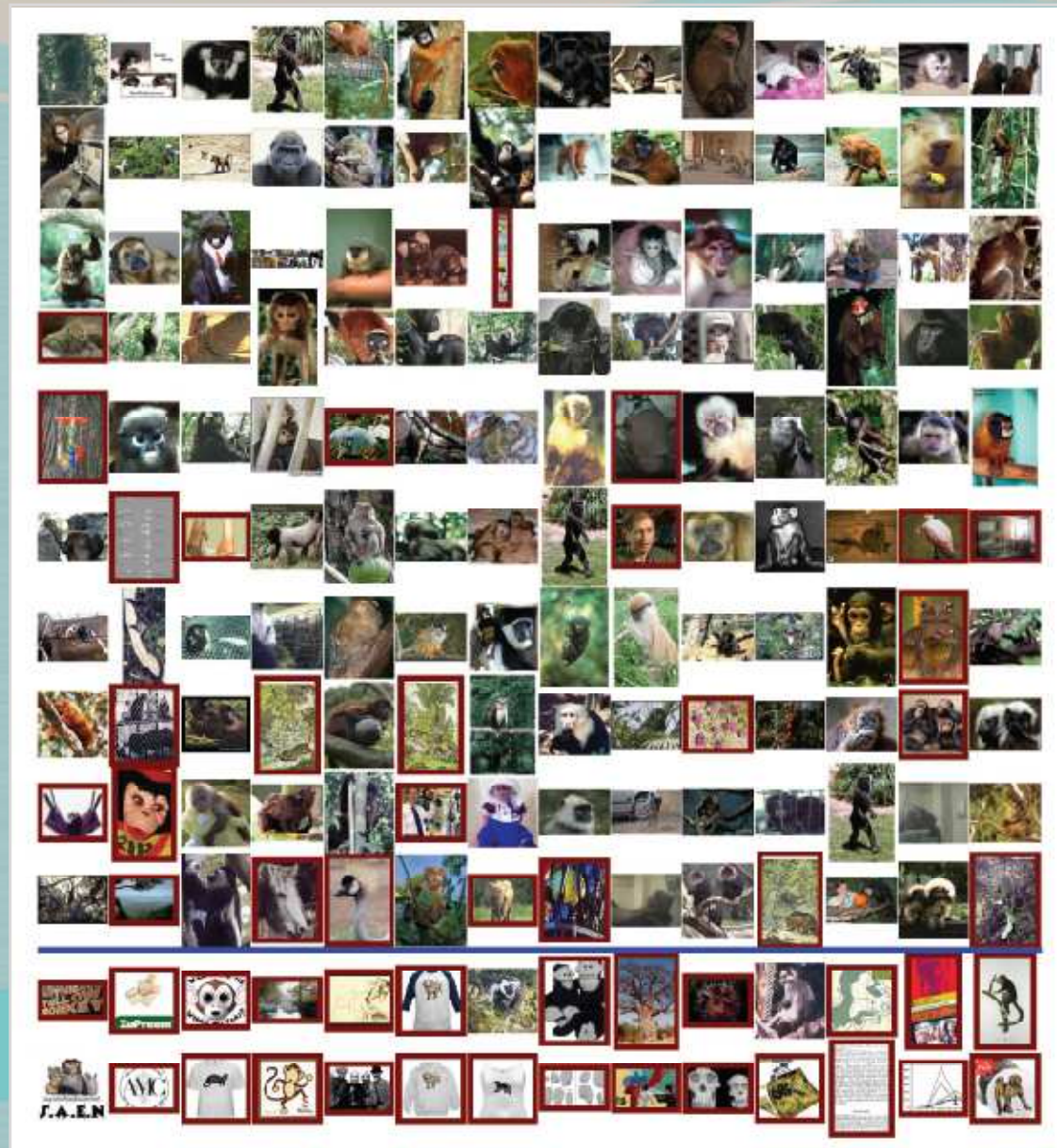
Data Collection as Recognition - Starting from Image Search

- Fergus *et al* – unsupervised clustering for “good” images from Google:
 1. Google’s image search
 2. Each picture is described by a bag of visual words
 3. Clustering into visual components using pLSA
 4. Relevant clusters selection using the first images from the search results (search is done for several languages using Google’s translation tool)

Data Collection as Recognition - Starting from text search

- Berg and Forsyth method for creating large, high-quality dataset using both textual and visual features:
 1. Text search -> images of sufficient size are extracted
 2. Selecting a set of visual exemplars:
 1. LDA for the Web pages to discover a set of latent topics for each category
 2. Select highly likely words for each topic
 3. For each topic images are ranked by their nearby words likelihood -> 30 exemplars are selected for each topic
 4. User labels each topic as relevant or background (using the exemplars and words) all the relevant topics (and background) are merged together
 3. Voting on all the other image using image and nearby words information
- 81% of the top 500 images are correct

“Monkey” Semi-automatically Generated Dataset



Recommendations

- Large datasets with ground truth labels are needed
- Labels should provide information about the classes, shape, locations, and more
- Future databases should have intra-class variability and different levels of difficulty
- There is a need for rigorous evaluation protocols for the algorithms over datasets
- Tools for testing specific aspects of algorithms on datasets would be extremely useful
- Gathering the statistics of the results should be done with “caution”
- Meta analysis of category level object recognition algorithms could prove to be fruitful



Caltech-256

- Smallest category size now 80 images
- About 30K images
- Harder
 - Not left-right aligned
 - No artifacts
 - Performance is halved
 - More categories
- Performance are halved (even less)
- New and larger clutter category



Slide credit: *Griffin, Holub, Perona*



A comparison of affine region detectors and descriptors

References:

- “A performance evaluation of local descriptors”, Mikolajczyk and Schmid, International Journal of Computer Vision 2005
- “A Comparison of Affine Region Detectors”, Mikolajczyk, Tuytelaars, Schmid, Zisserman, Matas, Schaffalitzky, Kadir, Van Gool, PAMI 2005
- Slides credit: Cordelia Schmid

Affine covariant detectors

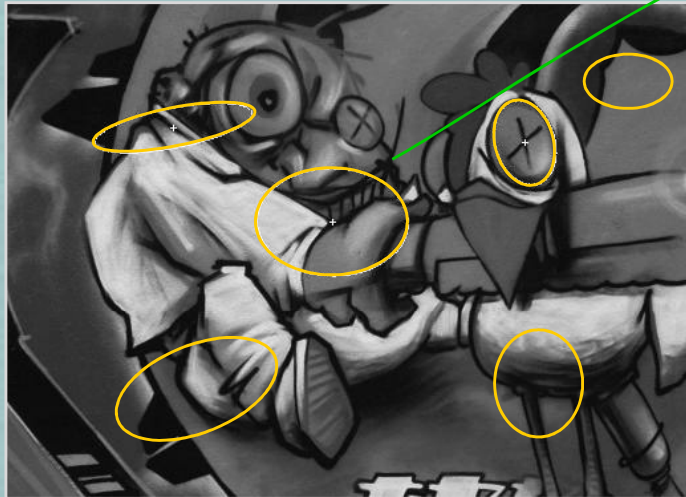
- *Harris-Affine (Mikolajczyk and Schmid'02, Schaffalitzky and Zisserman'02)*
- *Hessian-Affine (Mikolajczyk and Schmid'02)*
- *Maximally stable extremal regions (MSER) (Matas et al.'02)*
- *Intensity based regions (IBR) (Tuytelaars and Van Gool'00)*
- *Edge based regions (EBR) (Tuytelaars and Van Gool'00)*
- *Entropy-based regions (salient regions) (Kadir et al.'04)*

Dataset

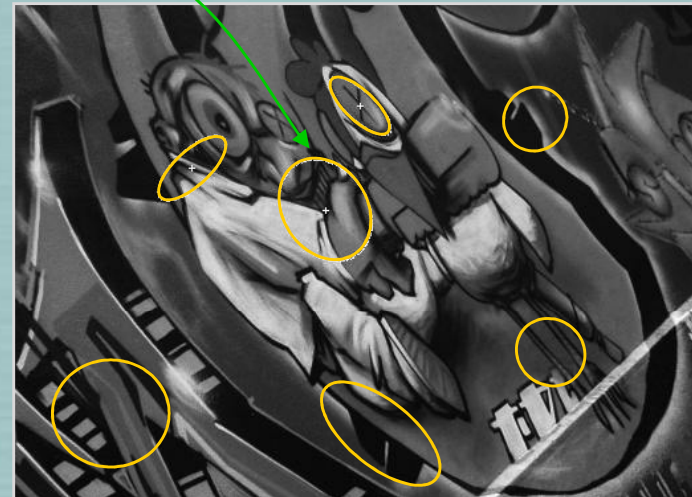
- Different types of transformation
 - Viewpoint change
 - Scale change
 - Image blur
 - JPEG compression
 - Light change
- Two scene types
 - Structured
 - Textured



Evaluation criterion

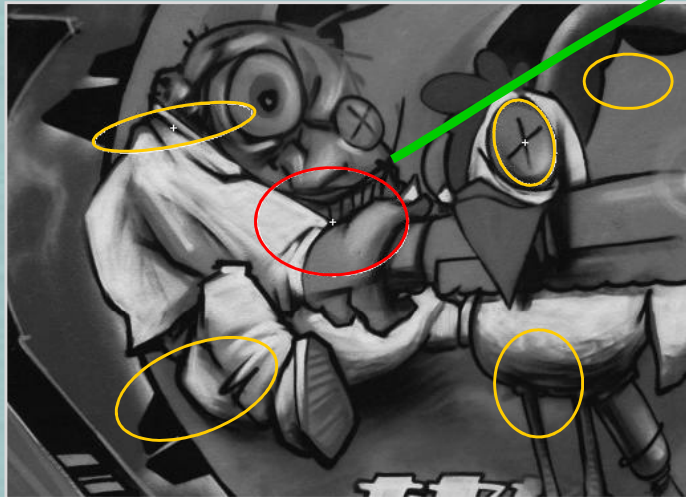


\xrightarrow{H}

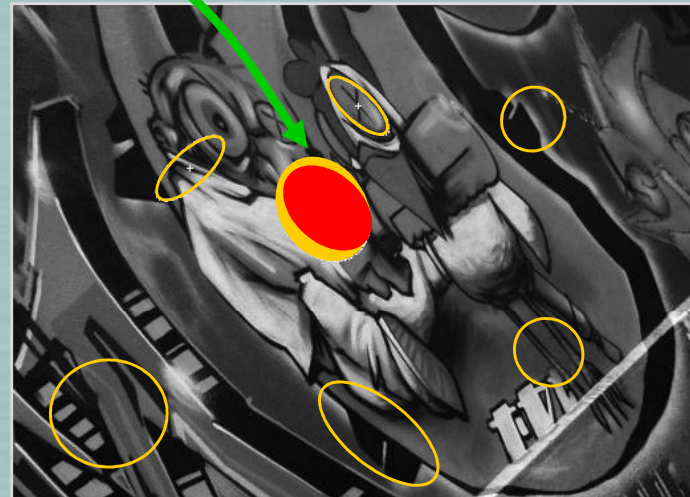


$$\text{repeatability} = \frac{\# \text{corresponding regions}}{\# \text{detected regions}} \cdot 100\%$$

Evaluation criterion

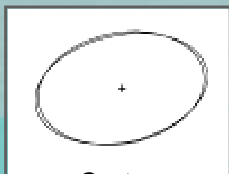


\xrightarrow{H}

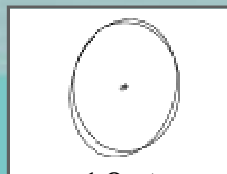


$$\text{repeatability} = \frac{\# \text{corresponding regions}}{\# \text{detected regions}} \cdot 100\%$$

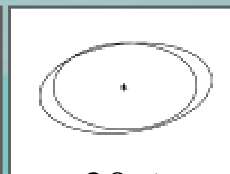
$$\text{overlap error} = \left(1 - \frac{\text{intersection}}{\text{union}}\right) \cdot 100\%$$



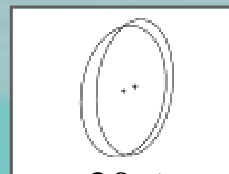
2%



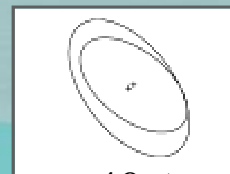
10%



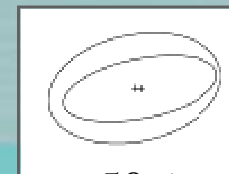
20%



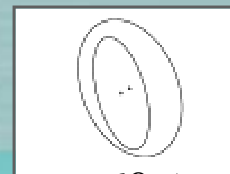
30%



40%



50%



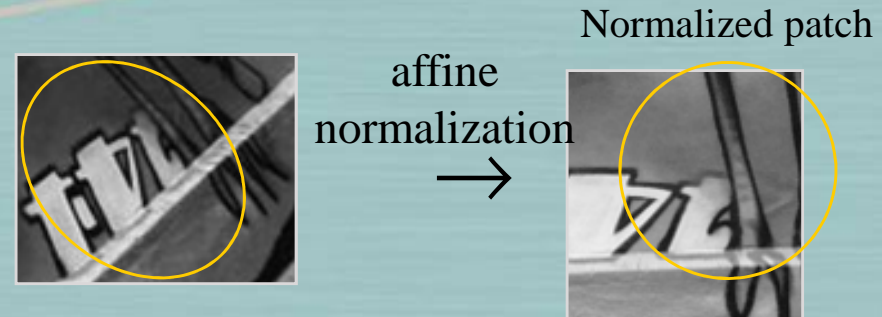
60%

Conclusion

- Results depend on transformation and scene type, no one best detector
- Performance of all declines slowly, with similar rates
- In many cases MSER was the best performer (Hessian-Affine second)
- Hessian-Affine and Harris-Affine – provide more regions than other detectors
- Edge based regions fail for texture scenes
- Detectors are complementary
 - MSER and EBR adapted to structured scenes
 - Harris-Affine and Hessian-Affine adapted to textured scenes

Descriptors

- Invariant to geometric and photometric transformations

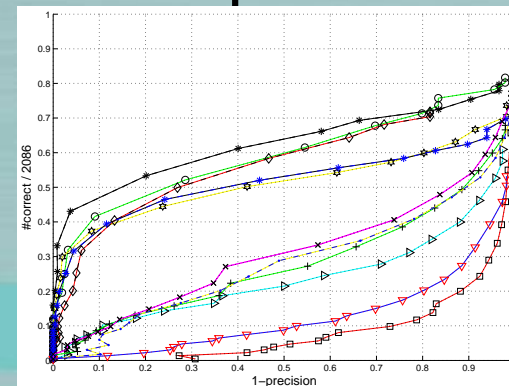


- Descriptors

- Sampled image patch
- Gradient orientation histogram - SIFT (Lowe'99)
- Shape context (Belongie et al.'02)
- PCA-SIFT (Ke and Sukthankar'04)
- Extended SIFT – SIFT with PCA dimensionality reduction
- Moment invariants (Van Gool'96)
- Gaussian derivative-based descriptors
 - *Differential invariants (Koenderink and van Doorn'87)*
 - *Steerable filters (Freeman and Adelson'91)*
- Complex filters (Baumberg'00, Schaffalitzky and Zisserman'02)

Comparison criterion

- Descriptors should be
 - Distinctive
 - Robust to changes on viewing conditions as well as to errors of the detector
- Different image transformation (the same as with the detectors)
- Different types of interest regions
- Different types of matching criterion
- Evaluation was done by looking on the recall with respect to precision
 - Recall: $\frac{\text{\#correct matches}}{\text{\#correspondences}}$
 - Precision: $\frac{\text{\#correct matches}}{\text{\# all matches}}$



Conclusion

- Performance of the descriptor is relatively independent of the detector
- Results similar for different matching strategies
- Dimension can be chosen optimally
- SIFT based descriptors perform best (high dimensional)

- A large set of good region detectors and descriptors exist
 - small extensions are possible, for example to deal with shape
- Good performance for recognizing an object/scene observed under different viewpoints and in a different context
 - invariance, occlusion, clutter
 - evaluation criteria tuned to this context
- <http://www.robots.ox.ac.uk/~vgg/research/affine/>



Thank You!