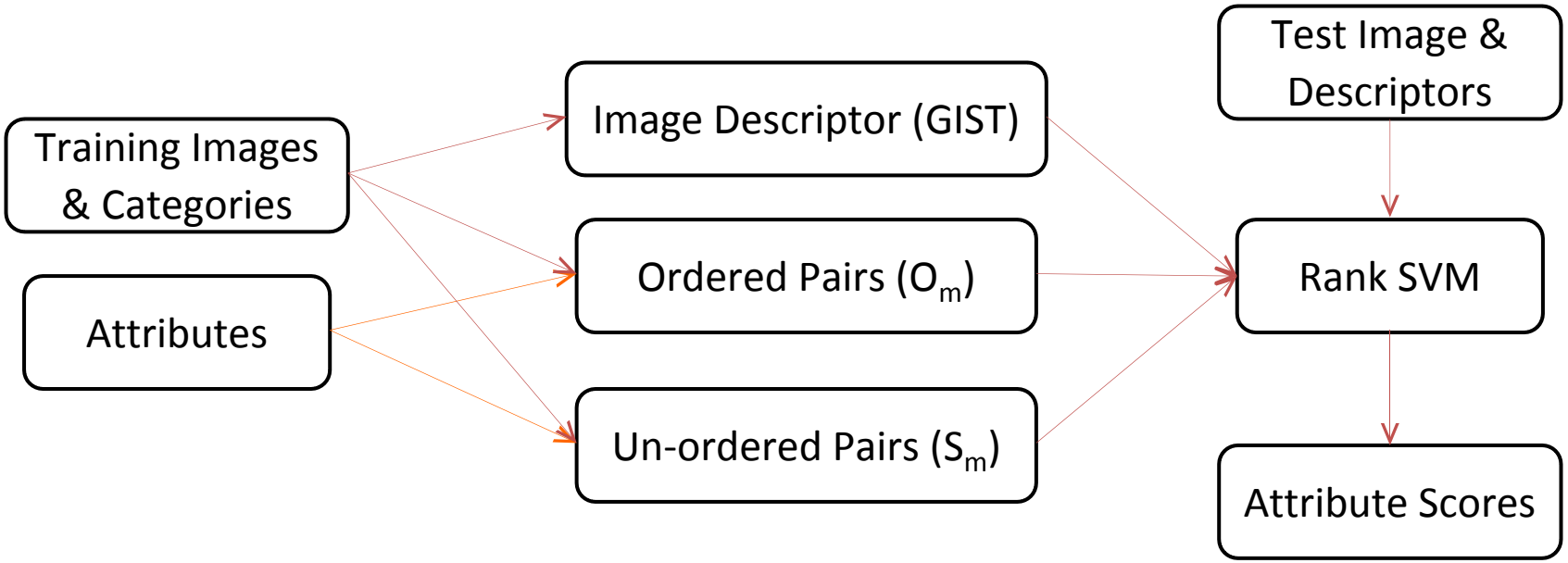
Relative Attributes

Experiments

Sanmit Narvekar
Department of Computer Science
The University of Texas at Austin

October 19, 2012

Overview



1. How does the type of "pairs" supervision given affect how well an attribute is learned?
2. Do we need a continuous relative ranking, or would discrete work better?
3. How do we know whether the attributes are learning the features they correspond to?

Analyzing Type of Supervision

The paper does this

- Category-level training pairs
  - Easy to obtain more pairs, which may not all be “correct”
- Instance-level training pairs
  - Harder to obtain, but more “correct”

Categories are different people or scene types

Analyzing Type of Supervision

- Compare which attributes perform better for which type of supervision

- Masculinity and smiling
- Naturalness and openness
- Evaluated on 10 random pairs of images

Accuracies

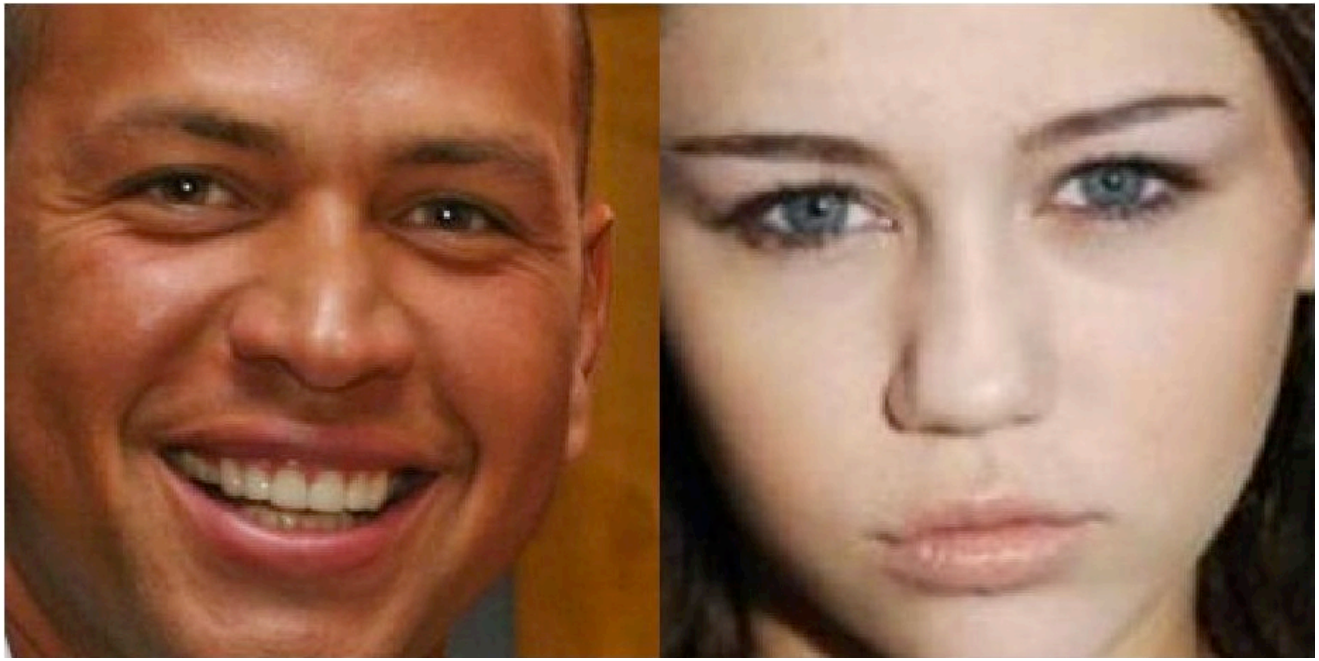
	Masculinity	Smiling
Categorical	0.90	0.70
Instance	0.80	0.70

Faces Dataset

	Naturalness	Openness
Categorical	0.90	0.80
Instance	0.80	0.90

Scenes Dataset

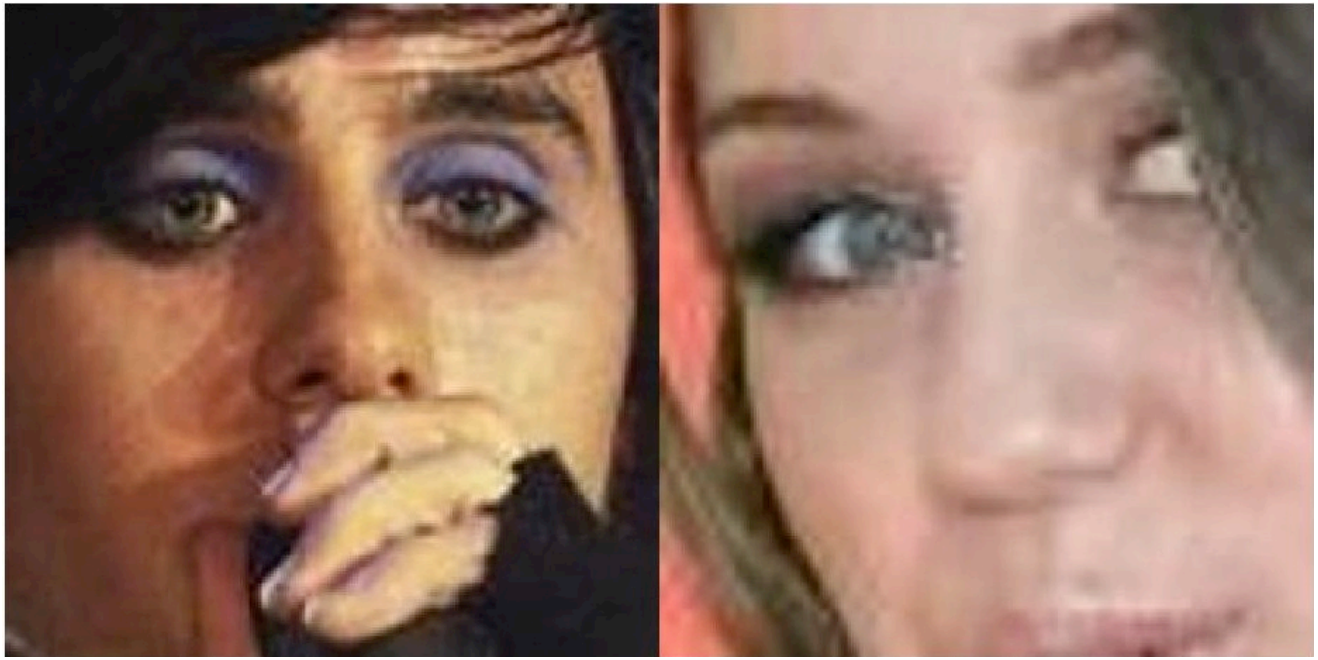
Masculinity and Smiling



	Smiling	
Categorical	-0.1354	0.1155
Instance	0.1091	0.0852

Miley usually smiles more than Alex, so the categorically trained classifier got confused
Attributes that vary within classes are trained better on instances

Masculinity and Smiling

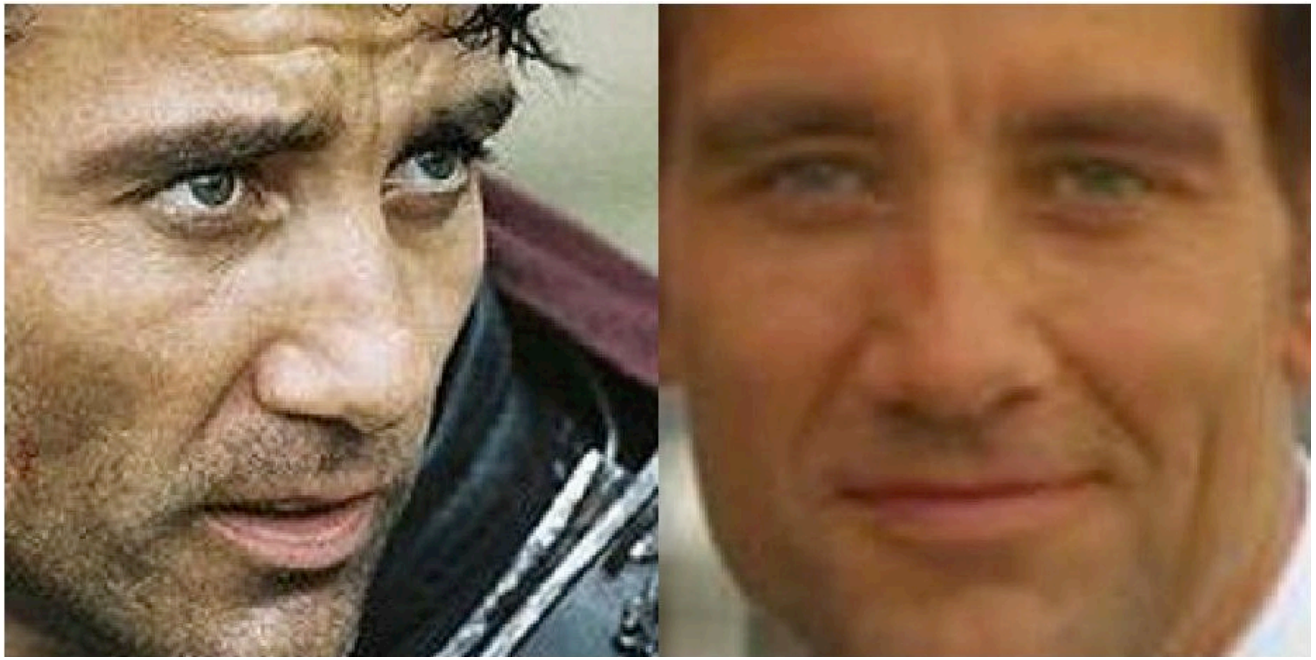


	Smiling	
Categorical	-0.2777	0.0697
Instance	0.0384	-0.1093

Occlusion interferes with the inference.

But, we know Miley usually smiles more than Alex. Does this count?

Masculinity and Smiling



	Masculinity	
Categorical	-0.1211	0.0829
Instance	0.7310	0.4664

SAME PERSON?!

Masculinity is technically a categorical attribute
However, even categorical attributes can vary intra-class in unexpected ways

Naturalness and Openness

Naturalness

Categorical	0.5463	-0.0561
Instance	0.2931	-0.0162

And some things inevitably come down to taste

Need for Relative Attributes

- Do we really need continuous relative attributes?

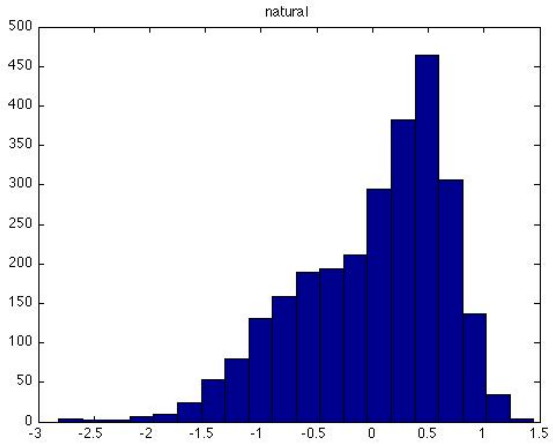
  OR

- Do some attributes form distinct groups?
  - male vs. female
  - natural vs. artificial
  - Could be more than 2 groups…
  - Then use a discrete ranking system?

  Analyze the histogram of rankings across attributes and their mean shift cluster centers

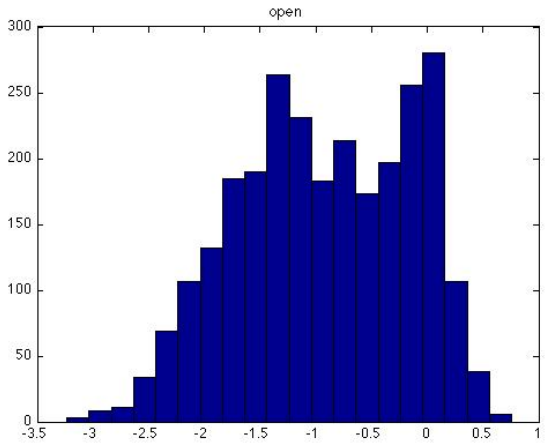
Relative Attributes (OSR)

Natural



Mean shift clusters

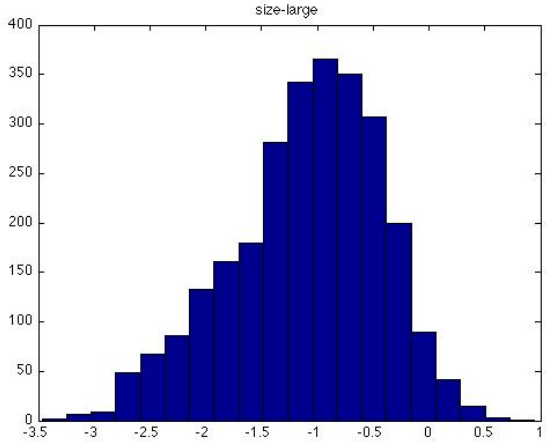
(0.4013, -2.5863)



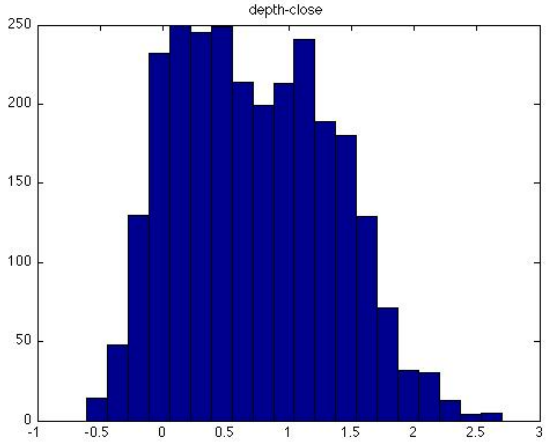
(-0.1120, -1.3116)

Open

Large size



(-0.8582)



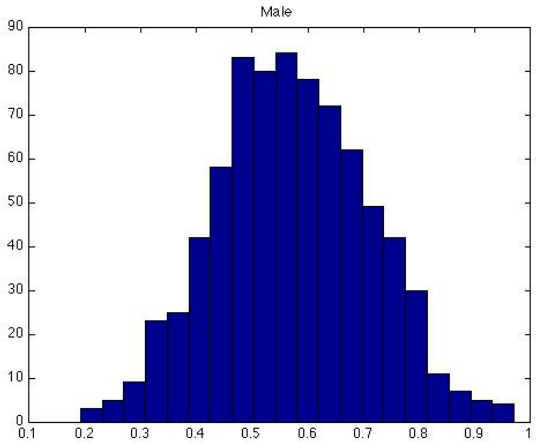
(0.9771, 0.2566)

Close depth

Most rankings have a Gaussian-like distribution, suggesting attributes are more amenable to representation by relative rankings rather than binary or discrete rankings

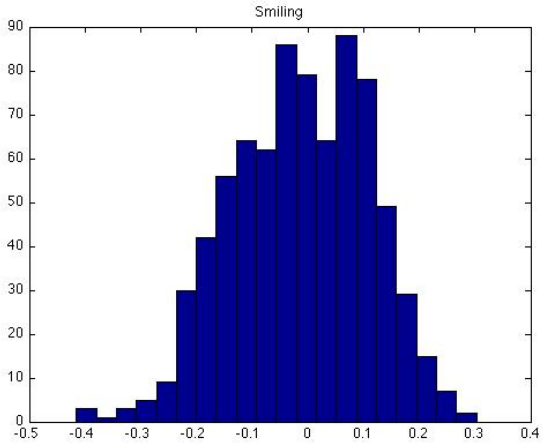
Relative Attributes (PubFig)

Male



(0.5728)

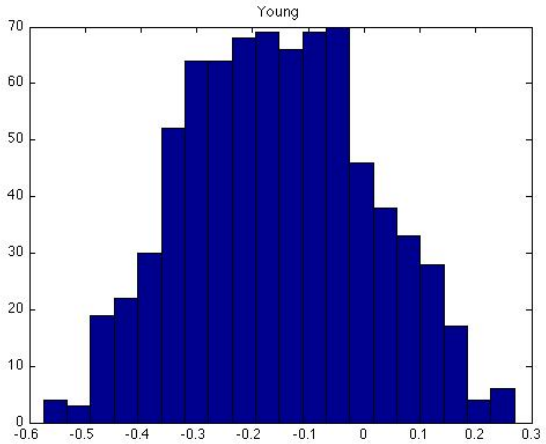
Smiling



(-0.0110)

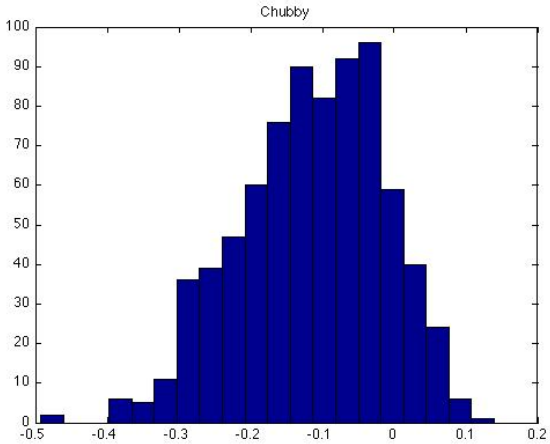
Gaussian even for "intrinsically" categorical attributes

Young



(-0.1543)

Chubby



(-0.1151)

In distributions where a lot of the mass is in the middle, binary attribute labels (representing the extrema) could be inappropriate

Attribute Localization

- How do you know whether the attributes learned correspond to their semantic meanings?
  - Especially when no labels, bounding boxes, etc. given

Object recognition

Learning airplane or sky?

Attribute-based recognition

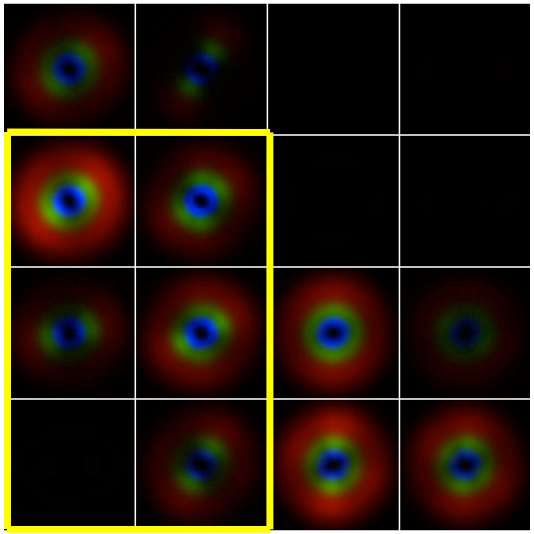
Learning high heels or no laces?

Seems more problematic in attribute-based recognition, since each attribute has semantic meaning, and is a part of a whole that can be hard to identify

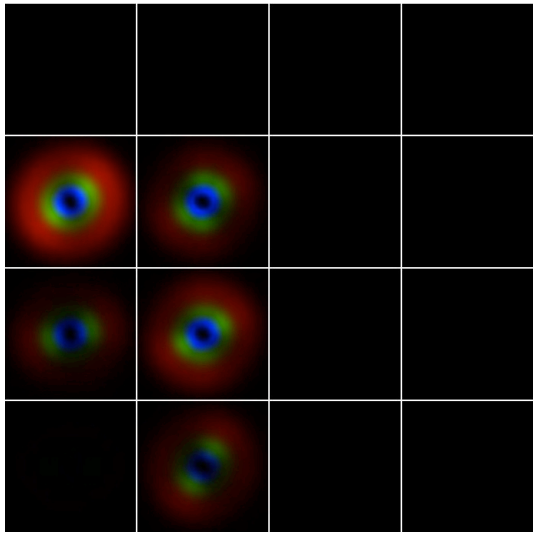
Attribute Localization

- **Task**: Determine whether the ranker is learning the attribute "high heels" in a dataset of shoes
- **Approach**:

Descriptor of whole image



Descriptor of heel area



Compare results of rankers trained on these different types

Attribute Localization

- Evaluate on 10 random pairs of images
- Images are automatically flipped if facing the wrong way
- Compare how well each method ranks high heels given
  - Image descriptor of the whole image
  - Image descriptor of only the heel area
  - Image descriptor of everything except the heel area

Accuracies

Whole Image	Relevant Area	Irrelevant Area
1.00	0.80	0.50

Suggests some contextual information was used for classification

Find the Highest Heel

Whole	0.6742	-0.6160
Relevant	-0.0342	-0.1440
Irrelevant	-0.1146	-0.0074

The “whole” and “relevant” descriptors both saw the missing heel in the right-side shoe
The straps might have mislead the “irrelevant area” classifier?

Find the Highest Heel

Whole	1.3252	1.8974
Relevant	-0.0154	-0.0181
Irrelevant	0.0612	-0.0910

The ranker fed the whole image descriptor could probably reason about heel height from the sole, since the heel itself was occluded.

Attribute captured, not captured, or assisted?

Summary

We looked at:

- Types of supervision, and its effects on attributes intrinsic to a class (masculinity) and where they can vary (smiling)
  - Category-level supervision
  - Instance-level supervision
- Need for continuous relative attributes, or whether attributes form “discrete” groups
  - How that affects different classes
- Attribute localization
  - Are we learning what we think we are?

References

- D. Parikh and K. Grauman. Relative Attributes. ICCV 2011.
- A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 2001.

- Links to existing code and data used:
  - GIST: http://people.csail.mit.edu/torralba/code/spatialenvelope/
  - Rank SVM: http://ttic.uchicago.edu/~dparikh/relative.html#code
  - Categorical and Instance Pair labels, extracted feature representations: http://www.cs.utexas.edu/~grauman/research/datasets.html

- Links to primary datasets used:
  - OSR: http://people.csail.mit.edu/torralba/code/spatialenvelope/
  - PubFig: http://www.cs.columbia.edu/CAVE/databases/pubfig/

Questions?