More words and Bigger Pictures

D.A. Forsyth, UIUC,
with input from J. Hockenmaier, UIUC, D. Hoiem, UIUC,
T. Berg, SUNYSB, P. Duygulu, Bilkent,
K. Barnard, U. Arizona,
A. Farhadi, U. Washington
I. Endres, A. Sadeghi, B. Liao, Y. Wang, V. Hedau, K. Karsch,
al of UIUC
G. Wang, NTU,
Nicolas Loeff, RGM Advisors
10 years old

8 months old

Computer Vision
A Modern Approach
SECOND EDITION

FORSYTH | PONCE
Why is visual object recognition useful?

- If you want to act, you must draw distinctions
- For robotics
  - recognition can predict the future
    - is the ground soggy?
    - is that person doing something dangerous?
    - does it matter if I run that over?
    - which end is dangerous?
- For information systems
  - recognition can unlock value in pictures
    - for search, clustering, ordering, inference, ...
- General engineering
  - recognition can tell what people are doing
- If you have vision, you have some recognition system
Query on

“Rose”

Example from Berkeley Blobworld system

Annotation results in complementary words and pictures
Annotation results in complementary words and pictures

Query on

Example from Berkeley Blobworld system
Annotation results in complementary words and pictures

Query on

“Rose”

and

Example from Berkeley Blobworld system
Example: Predicting word tags

It was there and we didn’t

It was there and we predicted it  
It wasn’t and we did

Substantial literature; this figure from Loeff Farhadi 08; see also Quattoni Darrell 07
Words and pictures affect one another

Marc by Marc Jacobs

Zappos.com

soft and glassy patent calfskin trimmed with natural vachetta cowhide, open top satchel for daytime and weekends, interior double slide pockets and zip pocket, seersucker stripe cotton twill lining, kate spade leather license plate logo, imported 2.8” drop length 14”h x 14.2”w x 6.9”d

Katespade.com

It's the perfect party dress. With distinctly feminine details such as a wide sash bow around an empire waist and a deep scoopneck, this linen dress will keep you comfortable and feeling elegant all evening long. Measures 38” from center back, hits at the knee.
* Scoopneck, full skirt.
* Hidden side zip, fully lined.
* 100% Linen. Dry clean.

bananarepublic.com

Conclusion

- Recognition is subtle
  - strong basic methods based on classifiers
  - serious problems with intellectual underpinnings
- Important recognition technologies coming
  - the unfamiliar
  - phrases
  - geometry
  - selection
- Crucial open questions
  - dataset bias
  - links to utility
A belief space about recognition

- Object categories are fixed and known
  - Each instance belongs to one category of $k$

- Good training data for categories is available

- Object recognition = $k$-way classification

- Detection = lots of classification
Obtain dataset

Build features

Mess around with classifiers, probability, etc

Produce representation
Features

- Principles
  - illumination invariant (robust) -> gradient orientation features
  - windows always slightly misaligned -> local histograms
- HOG, SIFT features (Lowe, 04; Dalal+Triggs 05)
Classification works well
<table>
<thead>
<tr>
<th>AnswerPhone</th>
<th>GetOutCar</th>
<th>HandShake</th>
<th>HugPerson</th>
<th>Kiss</th>
<th>SitDown</th>
<th>SitUp</th>
<th>StandUp</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="AnswerPhone" /></td>
<td><img src="image2" alt="GetOutCar" /></td>
<td><img src="image3" alt="HandShake" /></td>
<td><img src="image4" alt="HugPerson" /></td>
<td><img src="image5" alt="Kiss" /></td>
<td><img src="image6" alt="SitDown" /></td>
<td><img src="image7" alt="SitUp" /></td>
<td><img src="image8" alt="StandUp" /></td>
</tr>
<tr>
<td><img src="image9" alt="AnswerPhone" /></td>
<td><img src="image10" alt="GetOutCar" /></td>
<td><img src="image11" alt="HandShake" /></td>
<td><img src="image12" alt="HugPerson" /></td>
<td><img src="image13" alt="Kiss" /></td>
<td><img src="image14" alt="SitDown" /></td>
<td><img src="image15" alt="SitUp" /></td>
<td><img src="image16" alt="StandUp" /></td>
</tr>
<tr>
<td><img src="image17" alt="AnswerPhone" /></td>
<td><img src="image18" alt="GetOutCar" /></td>
<td><img src="image19" alt="HandShake" /></td>
<td><img src="image20" alt="HugPerson" /></td>
<td><img src="image21" alt="Kiss" /></td>
<td><img src="image22" alt="SitDown" /></td>
<td><img src="image23" alt="SitUp" /></td>
<td><img src="image24" alt="StandUp" /></td>
</tr>
<tr>
<td><img src="image25" alt="AnswerPhone" /></td>
<td><img src="image26" alt="GetOutCar" /></td>
<td><img src="image27" alt="HandShake" /></td>
<td><img src="image28" alt="HugPerson" /></td>
<td><img src="image29" alt="Kiss" /></td>
<td><img src="image30" alt="SitDown" /></td>
<td><img src="image31" alt="SitUp" /></td>
<td><img src="image32" alt="StandUp" /></td>
</tr>
</tbody>
</table>

Movies and captions: Laptev et al 08
Detection with a classifier
Conclusion

- Recognition is subtle
  - strong basic methods based on classifiers
  - serious problems with intellectual underpinnings
- Important recognition technologies coming
  - the unfamiliar
  - phrases
  - geometry
  - selection
- Crucial open questions
  - dataset bias
  - links to utility
A belief space about recognition

- Object categories are fixed and known
  - Each instance belongs to one category of $k$
- Good training data for categories is available
- Object recognition = $k$-way classification
- Detection = lots of classification

Platonism? Obvious nonsense
What have we inherited from this view?

• Deep pool of information about feature constructions
• Tremendous skill and experience in building classifiers
• Much practice at empiricism
  • which is valuable, and hard to do right

• Subtleties
  • What about the unfamiliar?
  • What kinds of things should we recognize?
  • What environmental knowledge helps?
  • What should we say about pictures?
  • How does utility affect the output?
A belief space about recognition

- Object categories are fixed and known
  - Each instance belongs to one category of k

- Good training data for categories is available

- Object recognition = k-way classification

- Detection = lots of classification

Platonism?

Obvious nonsense
Are these monkeys?
Big questions

- What signal representation should we use?

- What should we say about visual data?

PLUMBING
Classifiers, probability
(Light entertainment)

MODELS
What aspects of the world should we represent and how?

Obtain dataset
Build features
Mess around with classifiers, probability, etc
Produce representation
Bias

Frequencies in the data may misrepresent the application

- Because the labels are often wrong
  - $P(\text{labelled}|X)$ is not uniform
  - eg obscure but important objects in complex clutter
  - eg pedestrians in crowds

- Because of what gets labelled
  - eg. pictures from the web are selected - not like a camera on head
  - eg. “Profession” labelling for faces in news pictures

- Because of what gets collected
  - Label error
  - Label bias
  - Curation bias

Should not be perjorative

$X=\text{data}$
Size doesn’t make bias go away

- And could make it worse...
  - eg your dataset collector really likes red cars

- cf next slide
Google “rooms”
Bias isn’t always bad

• If all the faces on the web are politicians
  • one needs only to be good at politicians to be good at the web

• If no users can tell an ape from a monkey
  • you might not have to either

• If people really only want to search videos for “kissing”
  • then you don’t need a general activity recognition strategy
Induction

• Fundamental principle of machine learning
  • if the world is like the dataset, then future performance will be like training
    • Chernoff bounds, VC dimension, etc., etc.

• But what if the world can’t be like the dataset?
Pedestrian Detection

- Pedestrian detection:
  - We may not run down people who behave strangely
    - want “will fail to detect with frequency ...”
    - can do “...” IF test set is like training set
  - There is a large weight of easy cases which may conceal hard cases

- Resolution (frankly implausible)
  - ensure that training set is like test set

- Resolution (perhaps)
  - try only to learn things that are “fairly represented” in datasets
  - i.e. build models
Object recognition

- The world can’t be like the dataset because
  - many things are rare in plausible datasets
    - but not in the world
  - this exaggerates bias

- Strategies
  - train by comparison to similar objects
  - represent in terms of pooled properties
Many things are rare

Wang et al, 10, LabelMe data cf word frequencies, which also tend to be like this
Defenses against Bias

- Appropriate feature representations
  - eg illumination invariance

- Appropriate intermediate representations
  - which could have less biased behavior
  - perhaps attributes? scenes? visual phrases?

- Appropriate representations of knowledge
  - eg geometry --- pedestrian example
Conclusion

- Recognition is subtle
  - strong basic methods based on classifiers
  - serious problems with intellectual underpinnings
- Important recognition technologies coming
  - the unfamiliar
  - phrases
  - geometry
  - sentences
- Crucial open questions
  - dataset bias
  - links to utility
Name in common use among sailors in 19th century is deeply shocking to modern ears; appears in Aubrey Maturin novels by Patrick O’Brien
britchka

brougham
The Unfamiliar
Vision for driving
Vision for driving

- Car
- 4-Legged Animal
- Head
- Walking Left
What is an object like?

Viz comic, issue 101
General architecture

Farhadi et al 09; cf Lampert et al 09
Direct Attribute Prediction

Known classes

Unknown classes

Attribute layer

Image features

Lampert ea 09; Farhadi ea 09

Stuff attributes
Attribute predictions for unknown objects

Farhadi et al. 09; cf Lampert et al. 09
Object categories in test set are not same categories as in training set
Known objects could be unfamiliar

- By being different from the typical

- Pragmatics suggests this is how adjectives are chosen
  - If we are sure it’s a cat, and we know that
    - an attribute is different from normal
    - the detector is usually reliable
  - we should report the missing/extra attribute
General architecture

Farhadi et al 09; cf Lampert et al 09
“Man with a dog on a leash.”
“Man in camouflage clothes restraining a vicious attack dog with a leash.”
Missing attributes

- **Aeroplane** No “wing”
- **Car** No “window”
- **Boat** No “sail”
- **Aeroplane** No “jet engine”
- **Motorbike** No “side mirror”
- **Car** No “door”
- **Bicycle** No “wheel”
- **Sheep** No “wool”
- **Train** No “window”
- **Sofa** No “wood”
- **Bird** No “tail”
- **Bird** No “leg”
- **Bus** No “door”
Extra attributes

Bird
“Leaf”

Bus
“face”

Motorbike
“cloth”

DiningTable
“skin”

People
“Furn. back”

Aeroplane
“beak”

People
“label”

Sofa
“wheel”

Bike
“Horn”

Monitor
window”
Indirect Direct Attribute Prediction

Known classes

Unknown classes

Attribute layer

Image features

Lampert ea 09

Stuff attributes
Indirect Attribute Prediction

- **Training**
  - learn predictors for known classes, usual procedure
  - y-a, a-z links from object semantics
    - all instances of a class have the same attribute vector
- **Test**
  - inference

- **Property:**
  - attributes from class predictions
    - so non-visual prediction should be OK
  - attribute predictions are “like” natural attribute vectors
Attribute Correlations

Lampert ea 09 after Osherson ea 91; Kemp ea 06
Datasets - I

• a-Pascal
  • mark up Pascal VOC 2008 with 64 attributes (using Amazon Turk)
  • all of it!

• a-Yahoo
  • 12 additional classes, from Yahoo, with attributes (Amazon Turk)
  • chosen to “mask” Pascal classes
    • Wolf (dog); Centaur (people, horses); goat (sheep); etc.

• Approx 1M annotations! ($600)

• Accuracy
  • Turk inter-annotator agreement 84.1%
  • UIUC inter-annotator agreement 84.3%
  • Turk UIUC agreement 81.4%

Farhadi ea 09
Datasets - II

- Animals with attributes
  - 30475 images
  - animals in 50 classes, min 92 per class
  - classes have attributes from Osherson, 91
  - 85 attributes in total
  - attribute markup inherited from class

Lampert ea 09
Datasets - III

Cross Category Object REcognition Dataset

2780 Images – from ImageNet
3192 Objects – 28 Categories
26695 Parts – 71 types
30046 Attributes – 34 types
1052 Material Images – 10 types

Endres et al 10; Farhadi ea 10

http://vision.cs.uiuc.edu/CORE
UIUC PASCAL Sentence Dataset

• 5 Sentences from AMT: “Please describe the image in one complete but simple sentence.”
• Quality control: qualification test + AMT grading task
• 8000 images for ~$1000

A large sheep standing between large trees in a rural area.

A ram stands in the middle of a group of trees.

The sheep is standing under the trees.

A sheep standing in a forest.

A sheep under pine trees

http://vision.cs.uiuc.edu/pascal-sentences/

Rashtchian et al. HLT NAACL 2010
Attribute Discovery Data

• Gather pictures/captions of shoes, handbags, ties, earings, handbags
• Parse text into attributes
• Automatically learn which are visual
  – Visual attributes are more accurately classified
  – Human-Computer agreement on which attributes are visual: 70-90%
• Produces 37705 annotated examples
• Automatically characterize attribute localizability and type

http://tamaraberg.com/attributesDataset/index.html

Berg et al. ECCV 2010
SBU Captioned Photo Dataset

- Query images with captions from Flickr
- Filter: minimum length, at least two words from keyword list, at least one spatial preposition
- Dataset contains 1,000,000 captioned images

http://dsl1.cewit.stonybrook.edu/~vicente/sbucaptions/

Ordonez et al. NIPS 2011
Other Attribute Datasets

SUN Attributes Dataset

http://cs.brown.edu/~gen/sunattributes.html

Patterson Hays CVPR 2012
Other Attribute Datasets

PubFig

http://www.cs.columbia.edu/CAVE/databases/pubfig/  
Kumar et al. ICCV 2009
Latent Root

Visual attributes

Detector Responses

Sp: spatial part (gridded location)
Blc: basic level category
Sc: superordinate category

Farhadi ea 10

P: predicate
F: functional attribute
Asp: aspect
animal
blc: eagle

function: can bite
function: can fly
function: is predator
function: is carnivorous

part: eye
part: foot
part: head
part: leg
part: mouth
part: wing

Pose: extended_wings
Pose: objects_front

animal
function: can bite
function: can fly

part: eye
part: foot
part: head
part: leg
part: mouth
part: tail
part: wing

Pose: objects_front
Localizing unfamiliar categories

- Detect by:
  - Part detectors (e.g., leg - over several example categories)
  - BLC detectors (e.g., animal - ditto)
  - Vote on location

- Train on familiar animals/vehicles, test on unfamiliar
No horses or carriages in training set

Farhadi ea 10
Conclusion

• Recognition is subtle
  • strong basic methods based on classifiers
  • many meanings, useful in different contexts

• Important recognition technologies coming
  • attributes
  • phrases
  • geometry
  • sentences

• Crucial open questions
  • dataset bias
  • links to utility
Meaning comes in clumps

“Sledder”
Is this one thing?
Should we cut her off her sled?
Scenes

- Likely stages for
  - Particular types of object
  - Particular types of activity

Xiao et al 10
Scenes

Torralba et al ’93
• Composites
  • easier to recognize than their components
  • because appearance is simpler
Issue: what should one recognize?

- Single objects
  - potentially inaccurate
  - which ones?
  - crosstalk between detectors
    - eg bottles and humans

- Visual phrases
  - chosen opportunistically for accuracy
  - potentially far too many
  - which ones?
  - crosstalk between detectors
    - eg bottle, person, person drinking from bottle, etc.
Decoding

• Take pool of detector responses, decide what to believe

• Standard pastime, usually unremarked

  • eg test against threshold (pretty much everyone, all the time)
  • eg greedy algorithm (Desai et al 09; Kang et al 06)
  • vote on location (Bourdev+Malik 09)
  • single classifier looks at all best detector responses (Maji et al 11)
  • each detector response retested, using others as features (Farhadi Sadeghi 11)

• Probably much richer topic than currently allowed
Decoding

We use visual phrase and object models to make independent predictions. We then combine the predictions by a decoding algorithm that takes all detection responses and decides on the final outcome. Note that visual phrase recognition works better than recognizing the participating objects. For example, the horse detector does not produce reliable predictions about horses in this picture, while the "person riding horse" detector finds one instance. Our decoding then successfully adds two examples of horses and removes two wrong predictions of people by looking at other detections in the vicinity.

Our decoding then successfully adds two examples of horses and removes two wrong predictions of people by looking at other detections in the vicinity.

Decoding is an inevitable part of multiple object detection. The decoder may need to boost some detections and suppress others based on local context.

There is an analogy to machine translation problems where the alignment has to be established between phrases and areas of images. One might think of our system as having a phrase table with entities like "person", "horse", and "person riding horse". The ultimate goal is to look at all phrases and find the longest phrase that matches. This procedure is often called decoding in machine translation. Our decoder has to take into account that some of the detectors should overlap and when they overlap it has to decide which of the overlapping detectors are worth reporting.

In this paper we show the benefits of opportunistically selecting basic atoms of recognition and the significant gain in directly detecting visual phrases. Our contributions are:

1. Introducing visual phrases as categories for recognition.
2. Introducing a novel dataset for phrasal recognition.
3. Showing that considering visual phrases provides a significant gain over state of the art object detectors coupled with the state of the art methods of modeling interactions.
4. Introducing a decoding algorithm that takes into account specific properties of interacting objects in multiple levels of abstraction.
5. Producing state of the art performance results in multi-class object recognition.

Related Works

Object Recognition: Due to limited space we only mention the most relevant works in object recognition. Deformable templates and part based models are of the most successful methods in object recognition. In this paper we use the state of the art detectors in using deformable part models. This work considers multiple roots to model the appearance changes due to viewpoint or inherent intra-class variations.

Object Interactions: All methods that model interactions between objects neglect the change in the appearance of objects due to interactions with other objects. We differ from all by taking this effect into account. Gupta et al. model these interactions by modeling the prepositions and adjectives that relate nouns. Yao and Li model the...
Decoding helps

Table 3: Phrasal recognition helps object detection. This table compares the performance of our decoding with that of state-of-the-art object detectors and state-of-the-art multiclass recognition methods.

Figure 7: Rows 2 and 3 depict our results before and after decoding, respectively. The same applies to rows 4 and 5. For example, in image "a" our decoding boosts the confidence of the bicycle classifier and suppresses the confidences of wrong person detections using reliable "person riding bicycle" detection. In image "c" a confident "dog lying on sofa" detector improves the confidence of the sofa detection and decreases the confidences of wrong person detections. In image "d" the "person sitting on chair" detector increases the confidence of the chair detection.

Our decoding shows that visual phrases help object detection and vice versa. In image "b" the confident sofa detection boosts the confidence of "dog lying on sofa" detection.

Figures showing the effectiveness of phrasal recognition and decoding in object detection.
Conclusion

- **Recognition is subtle**
  - strong basic methods based on classifiers

- **Important recognition technologies coming**
  - the unfamiliar
  - phrases
  - geometry
  - selection

- **Crucial open questions**
  - dataset bias
  - links to utility
Environmental knowledge

Hoiem et al 06
Environmental knowledge is powerful

Hoiem et al. 06
Vanishing points

- Cluster long straight edges into three clusters (after Rother, 02)
Estimating layout

- Choice of layout = 4DOF in image
- Search cost function
  - learned from examples
Clutter maps
Detecting beds

• **Natural strategy**
  • mark up data, apply Felzenswalb et al, ’08 (FMRG)

• **Problem**
  • changes in viewpoint lead to changes in appearance
    • FMRG doesn’t know this - must be less efficient

• **Using a room box**
  • rectify the image to each face of the room box
  • look for FACES of beds in each rectified image using FMRG
  • find three that share a corner
Detecting beds - I
Detecting beds - II

True positives

False positives
Detecting beds - III

- Beds constrain rooms
  - are axis-aligned
  - can’t pierce walls

- Variants
  - Box only (OK)
  - Box + 2D (better)
  - Jointly estimate room box, bed box(es) (best)
Joint estimation helps
Joint estimation helps
Conclusion

- **Recognition is subtle**
  - strong basic methods based on classifiers
  - many meanings, useful in different contexts
- **Important recognition technologies coming**
  - attributes
  - phrases
  - geometry
  - selection
- **Crucial open questions**
  - dataset bias
  - links to utility
Selection: What is worth saying?

Two girls take a break to sit and talk.

Two women are sitting, and one of them is holding something.

Two women chatting while sitting outside.

Two women sitting on a bench talking.

Two women wearing jeans, one with a blue scarf around her head, sit and talk.

Sentences from Julia Hockenmaier’s work

Rashtchian ea 10

For language people: Pragmatics - what is worth saying?
Some factors conducive to being mentioned (Berg et al. 12)

<table>
<thead>
<tr>
<th>Top10</th>
<th>Prob</th>
<th>Last10</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>firework</td>
<td>1.00</td>
<td>hand</td>
<td>0.15</td>
</tr>
<tr>
<td>turtle</td>
<td>0.97</td>
<td>cloth</td>
<td>0.15</td>
</tr>
<tr>
<td>horse</td>
<td>0.97</td>
<td>paper</td>
<td>0.13</td>
</tr>
<tr>
<td>pool</td>
<td>0.94</td>
<td>umbrella</td>
<td>0.13</td>
</tr>
<tr>
<td>airplane</td>
<td>0.94</td>
<td>grass</td>
<td>0.13</td>
</tr>
<tr>
<td>bed</td>
<td>0.92</td>
<td>sidewalk</td>
<td>0.11</td>
</tr>
<tr>
<td>person</td>
<td>0.92</td>
<td>tire</td>
<td>0.11</td>
</tr>
<tr>
<td>whale</td>
<td>0.91</td>
<td>smoke</td>
<td>0.09</td>
</tr>
<tr>
<td>fountain</td>
<td>0.89</td>
<td>instrument</td>
<td>0.07</td>
</tr>
<tr>
<td>flag</td>
<td>0.88</td>
<td>fabric</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 1. Probability of being mentioned when present for various object categories (ImageCLEF).
Berg et al 12: Objects more likely to be mentioned in uncommon context, figure shows probability of being mentioned conditioned on appearing
Predicting stylized narrations

Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder catches the ball after Fielder runs towards the ball. Fielder catches the ball before Fielder throws to the base. Fielder throws to the base and then Fielder at Base catches the ball at base.

Pitcher pitches the ball and then Batter hits. Fielder catches the ball after Batter hits.

Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder runs towards the ball and then Fielder catches the ball. Fielder throws to the base after Fielder catches the ball. Fielder throws to the base and then Fielder at Base catches the ball at base.

Pitcher pitches the ball and then Batter does not swing.

Gupta ea 09
Rich(ish) sentences from simple intermediates

Object, action, scene

Farhadi ea 10
### Examples

<table>
<thead>
<tr>
<th>(pet, sleep, ground)</th>
<th>(dog, sleep, ground)</th>
<th>(animal, sleep, ground)</th>
<th>(animal, stand, ground)</th>
<th>(goat, stand, ground)</th>
<th>see something unexpected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, room)</td>
<td>(furniture, place, home)</td>
<td>(bottle, place, table)</td>
<td>(display, place, table)</td>
<td>Refrigerator almost empty.</td>
</tr>
<tr>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, room)</td>
<td>(furniture, place, home)</td>
<td>(bottle, place, table)</td>
<td>(display, place, table)</td>
<td>Foods and utensils.</td>
</tr>
<tr>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, room)</td>
<td>(furniture, place, home)</td>
<td>(bottle, place, table)</td>
<td>(display, place, table)</td>
<td>Eatables in the refrigerator.</td>
</tr>
<tr>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, room)</td>
<td>(furniture, place, home)</td>
<td>(bottle, place, table)</td>
<td>(display, place, table)</td>
<td>The inside of a refrigerator apples, cottage cheese, tupperwares and lunch bags.</td>
</tr>
<tr>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, room)</td>
<td>(furniture, place, home)</td>
<td>(bottle, place, table)</td>
<td>(display, place, table)</td>
<td>Squash apenny white store with a hand statue, picnic tables in front of the building.</td>
</tr>
<tr>
<td>(transportation, move, track)</td>
<td>(bike, ride, track)</td>
<td>(transportation, move, road)</td>
<td>(pet, sleep, ground)</td>
<td>(bike, ride, road)</td>
<td>A man stands next to a train on a cloudy day</td>
</tr>
<tr>
<td>(transportation, move, track)</td>
<td>(bike, ride, track)</td>
<td>(transportation, move, road)</td>
<td>(pet, sleep, ground)</td>
<td>(bike, ride, road)</td>
<td>A backpacker stands beside a green train</td>
</tr>
<tr>
<td>(transportation, move, track)</td>
<td>(bike, ride, track)</td>
<td>(transportation, move, road)</td>
<td>(pet, sleep, ground)</td>
<td>(bike, ride, road)</td>
<td>This is a picture of a man standing next to a green train</td>
</tr>
<tr>
<td>(transportation, move, track)</td>
<td>(bike, ride, track)</td>
<td>(transportation, move, road)</td>
<td>(pet, sleep, ground)</td>
<td>(bike, ride, road)</td>
<td>There are two men standing on a rocky beach, smiling at the camera.</td>
</tr>
<tr>
<td>(transportation, move, track)</td>
<td>(bike, ride, track)</td>
<td>(transportation, move, road)</td>
<td>(pet, sleep, ground)</td>
<td>(bike, ride, road)</td>
<td>This is a person laying down in the grass next to their bike in front of a strange white building.</td>
</tr>
<tr>
<td>(display, place, table)</td>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, furniture)</td>
<td>(bottle, place, table)</td>
<td>(furniture, place, table)</td>
<td>(furniture, place, home)</td>
</tr>
<tr>
<td>(display, place, table)</td>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, furniture)</td>
<td>(bottle, place, table)</td>
<td>(furniture, place, table)</td>
<td>(furniture, place, home)</td>
</tr>
<tr>
<td>(display, place, table)</td>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, furniture)</td>
<td>(bottle, place, table)</td>
<td>(furniture, place, table)</td>
<td>(furniture, place, home)</td>
</tr>
<tr>
<td>(display, place, table)</td>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, furniture)</td>
<td>(bottle, place, table)</td>
<td>(furniture, place, table)</td>
<td>(furniture, place, home)</td>
</tr>
<tr>
<td>(display, place, table)</td>
<td>(furniture, place, furniture)</td>
<td>(furniture, place, furniture)</td>
<td>(bottle, place, table)</td>
<td>(furniture, place, table)</td>
<td>(furniture, place, home)</td>
</tr>
</tbody>
</table>

Farhadi ea 10
Adding Attributes and Prepositions

1) Object(s)/Stuff

Input Image

- a) dog
- b) person
- c) sofa

2) Attributes

- dog: brown 0.01, striped 0.16, furry 0.26, wooden 0.2, feathered 0.06...
- person: brown 0.32, striped 0.09, furry 0.04, wooden 0.2, feathered 0.04...
- sofa: brown 0.04, striped 0.10, furry 0.06, wooden 0.08, feathered 0.08...

3) Prepositions

- dog: near(b,a) 1, against(b,a) 11, against(b,a) 0.04, beside(b,a) 0.24, beside(b,a) 0.17...
- person: near(c,a) 1, against(a,c) 3, against(a,c) 0.05, beside(a,c) 0.5, beside(c,a) 0.45...
- sofa: near(b,c) 1, near(c,b) 1, against(b,c) 0.67, against(c,b) 0.33, beside(b,c) 0.0, beside(c,b) 0.19...

4) Constructed CRF

5) Predicted Labeling

- null, person - b, against, brown, sofa - c
- null, dog - a, near, null, person - b
- null, dog - a, beside, brown, sofa - c

6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Kulkarni et al 11
Adding Attributes and Prepositions

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.

Here we see one person and one train. The black person is by the train.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

This is a picture of two dogs. The first dog is near the second furry dog.

This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.
Use an integer program to enforce discourse, etc constraints (objects should not be mentioned repeatedly)

ILP: Method (Berg ea 12, ACL paper)

HMM: Yang et al 11 (cf Kulkarni ea 11)

Human: Human annotator
Detecting visual text

“Car” is Visual

Discriminative, using largely lexical features

Not Visual

Dodge et al 12

Another dream car to add to the list, this one spotted in Hanbury St.

Shot out my car window while stuck in traffic because people in Cincinnati can’t drive in the rain.

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>POSITION</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td></td>
<td>74.7</td>
</tr>
<tr>
<td>+ Image</td>
<td>-</td>
<td>74.4</td>
</tr>
<tr>
<td>+ Bootstrap</td>
<td>Phrase</td>
<td>74.3</td>
</tr>
<tr>
<td>+ Spell</td>
<td>Phrase</td>
<td>75.3</td>
</tr>
<tr>
<td>+ Length</td>
<td>-</td>
<td>74.7</td>
</tr>
<tr>
<td>+ Words</td>
<td>Before</td>
<td>76.2</td>
</tr>
<tr>
<td>+ Wordnet</td>
<td>Phrase</td>
<td>76.1</td>
</tr>
<tr>
<td>+ Spell</td>
<td>After</td>
<td>76.0</td>
</tr>
<tr>
<td>+ Spell</td>
<td>Before</td>
<td>76.8</td>
</tr>
<tr>
<td>+ Wordnet</td>
<td>Before</td>
<td>77.0</td>
</tr>
<tr>
<td>+ Wordnet</td>
<td>After</td>
<td>75.6</td>
</tr>
</tbody>
</table>

Table 6: Results of feature ablation on LARGE data set.
Conclusion

• Recognition is subtle
  • strong basic methods based on classifiers
  • many meanings, useful in different contexts

• Important recognition technologies coming
  • attributes
  • phrases
  • geometry
  • sentences

• Crucial open questions
  • dataset bias
  • links to utility
Defenses against Bias

- Appropriate feature representations
  - eg illumination invariance

- Appropriate intermediate representations
  - which could have less biased behavior
  - perhaps attributes? scenes? visual phrases?

- Appropriate representations of knowledge
  - eg geometry --- pedestrian example
Conclusion

- Recognition is subtle
  - strong basic methods based on classifiers
  - many meanings, useful in different contexts
- Important recognition technologies coming
  - attributes
  - phrases
  - geometry
  - sentences
- Crucial open questions
  - dataset bias
  - links to utility
What should we say about visual data?

- Most important question in vision
  - What does the output of a recognition system consist of?

- A list of all objects present in scene, and locations
  - obvious nonsense - too big, too sensitive to defn of “object”

- A useful representation of reasonable size
  - dubious answer
    - Useful in what way?
    - How do we make the size reasonable?
Object categories depend on utility

Monkey or Plastic toy or both or irrelevant

Person or child or beer drinker or beer-drinking child or tourist or holidaymaker or obstacle or potential arrest or irrelevant or...

Some of this depends on what you’re trying to do, in ways we don’t understand
Plausible belief space about recognition

- Categories are highly fluid
  - opportunistic devices to aid generalization
    - affected by current problem, utility
  - instances can belong to many categories
    - simultaneously
  - at different times, the same instance may belong to different categories
  - categories are shaded
    - much “within class variation” is principled
  - Most categories are rare
  - Many might be personal, many are negotiated

- Understanding (recognition)
  - constant coping with the (somewhat) unfamiliar
  - bias is pervasive, affects representation

Notice that some of these issues have resonant ideas when one thinks about the “meaning” of language.
The big question

- How to insert object semantics into object recognition?
  - without being silly
  - what is useful knowledge?
  - where does it come from?
  - what is worth saying about objects?
  - what objects are worth saying things about?
  - how should categories be created and destroyed to meet pragmatic needs?
Conclusion

• **Recognition is subtle**
  • strong basic methods based on classifiers
  • serious problems with intellectual underpinnings

• **Important recognition technologies coming**
  • the unfamiliar
  • phrases
  • geometry
  • selection

• **Crucial open questions**
  • dataset bias
  • links to utility
More information
The end

- Thanks to
  - ONR, NSF, Google
What is to be done?

• Cross border raiding by vision, NLP communities is fertile
  • long may it continue
  • even if the details of the analogy are sometimes shaky

• Build a body of knowledge about everyday objects
  • “mundane” knowledge, hard to harvest from the web

• Build a theory of what it means to be “like” something
  • in what respect are things similar? how can we use this idea?

• Build a theory of knowing and reasoning about objects
  • as applied to the concrete world
  • linked to visual observations