Size isn’t important
or
What do big visual datasets tell us?

D.A. Forsyth, UIUC
(and I’ll omit the guilty)
Conclusion

- Not much, if the emphasis is on size
- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day
- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations
- Recognition problems are hard to frame
  - excess certainty may be dangerous
What could big datasets tell us?  
(by virtue of being big)

- Good magnitude estimates of small effects
- A more accurate estimate of what the world is like
  - frequencies, etc

- Collective search is more significant than it gets credit for
  - Problem:
    - publish a dataset
    - people try methods, keep ones that do well
    - hence, results suffer from intense selection bias
  - Bigger datasets -> weaker recognition statistics
    - Because the categories are genuinely harder?
    - Because collective search is much harder?
Conclusion

- Not much, if the emphasis is on size

- **Collecting datasets is highly creative**
  - rather than a nuisance activity
  - tools are getting better by the day

- **Bias, weird frequencies are a major issue**
  - There are no best practices for avoiding problems
  - May shape our representations

- **Recognition problems are hard to frame**
  - excess certainty may be dangerous
Non-parametric regression

With a broad view of “match”, “smooth”, all classifiers fit into this story
A=picture, B=category

- Far too many to select one!
- Fergus et al 05; Fergus et al 04; Fei-Fei 06; Berg 05; Everingham et al Pascal Challenge reports 06, 07, 08;
  - etc etc etc etc etc

Table 1. Overall classification performance of the system, in various configurations, to 4289 control images and 565 test images. Configuration F is the primary configuration of the grouper, fixed before the experiment was run, which reports a nude present if either a girdle, a limb-segment girdle or a spine group is present, but not if a limb group is present. Other configurations represent various permutations of these reporting conditions; for example, configuration A reports a person present only if girdles are present. There are fewer than 15 cases, because some cases give exactly the same response.

Forsyth et al 96, 01
Conclusion

• Not much, if the emphasis is on size

• Collecting datasets is highly creative
  • rather than a nuisance activity
  • tools are getting better by the day

• Bias, weird frequencies are a major issue
  • There are no best practices for avoiding problems
  • May shape our representations

• Recognition problems are hard to frame
  • excess certainty may be dangerous
Bias

- Frequencies in the data may misrepresent the application
  - Because the labels are often wrong
    - $P(\text{labelled}|X)$ is not uniform
  - Because of what gets labelled
    - eg obscure but important objects in complex clutter
    - eg pedestrians in crowds
  - Because of what gets collected
    - eg. pictures from the web are selected - not like a camera on head
    - eg. “Profession” labelling for faces in news pictures

Should not be perjorative
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 people really only want to search videos for “kissing”
  - then you don’t need a general activity recognition strategy
Bias is pervasive
Size doesn’t make bias go away

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

- cf next slide
Label error

• Fact of life
  • people label things wrong

• Can fix when there are many instances
  • consistency (Zhao et al 08)
  • smoothing (Berg, 06; Li, 06; Wang 08; Collins 08)

• Might be able to fix with hierarchy+generalization
  • we should never mix up “cat”’s and “truck”’s
Label bias: the choice of what is labelled

- \( P(\text{labelled}|X) \) is not uniform
  - or \( P(X|\text{labelled}) \) is not the same as \( P(X|\text{not labelled}) \)

- There are models
  - problem sometimes called dataset shift, see (Quinonero-Candela 09)
  - can be addressed with, say, large unlabelled datasets
    - build smoothed estimate of \( p(\text{labelled}|X) \), reweight

- Important effect
  - can make high capacity classifiers generalize better than low capacity
  - (maybe) be very cautious about linear SVM’s
Curation bias

- Collected data is not a fair sample of X
  - labelled AND unlabelled data

- Images on the web are “curated”

- Iconography seems to be a big effect here
  - visual “modes” of representation
    - see Berg+Berg 09
  - we might not see them all
  - cf Google image search with Flickr

X=data
Y=labels
X_i = unlabelled examples
(X_j, Y_j)=labelled examples
Iconographic phenomena

Berg+Berg 09; see Jing+Baluja 08
Google “rooms”

... virtual tour > room photos 644 x 446 - 39k - jpg
www.mandalaybay.com

Bed Room Sets 599 x 402 - 33k - jpg
www.chiphi-pl.org

16 Creative and Sexy Art Hotel Rooms 468 x 354 - 11k - jpg
weburbanist.com
[ More from weburbanist.com ]

Rooms > 450 x 300 - 25k - jpg
www.radisson.com
[ More from www.radisson.com ]

Bookcase Secret Room Door 468 x 391 - 98k - jpg
weburbanist.com

The large room known today as the 350 x 353 - 48k - jpg
www.royalacademy.org.uk

To reserve a room call 640 x 480 - 93k - jpg
212-596-1200 ...
www.columbiaclub.org

Now let’s see some amazing rooms.
450 x 300 - 19k - jpg
freshome.com

Room for physically-challenged 600 x 396 - 24k - jpg
www.hotelnikkohanoi.com.vn

basement family room 450 x 325 - 49k - jpg
www.thisoldhouse.com

Handicap Room 300 x 301 - 22k - jpg
www.intl-house.howard-hotels.com

Spacious Guest Room 450 x 300 - 29k - jpg
www.radisson.com

Rooms may also include twin beds and ... 370 x 486 - 40k - jpg
www.inisrael.com

This bright room on the 2nd floor of ... 1728 x 1152 - 28k - jpg
biosphere.ec.gc.ca

These twenty rooms ...
468 x 352 - 97k - jpg
weburbanist.com

Texas’ enormous locker room facility ...
530 x 343 - 34k - jpg

Two Queen Room 450 x 300 - 25k - jpg
www.countryinn.com

trent room The Trent Room was first ...
346 x 450 - 54k - jpg

Image of changing room 450 x 388 - 75k - jpg

Tour the USC Marshall Capture Room 637 x 481 - 160k - jpg

large drawing room in two room suite
737 x 551 - 70k - jpg
Flickr “rooms”
Conclusion

• Not much, if the emphasis is on size

• Collecting datasets is highly creative
  • rather than a nuisance activity
  • tools are getting better by the day

• Bias, weird frequencies are a major issue
  • There are no best practices for avoiding problems
  • May shape our representations

• Recognition problems are hard to frame
  • excess certainty may be dangerous
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
  - this exaggerates bias

Wang et al, 10
Distributional semantics

• Most words are unusual
• Don’t know a word?
  • nearby words can tell you what it means
  • or how similar it is to a word you do know

“No; this my hand will rather the multitudinous seas incarnadine, making the green one red.”

“In one routine, describing his “ludicrously alpha” surfing instructor for the Forgetting Sarah Marshall shoot, he exclaims, “The sea were incarnadine wiv his testosterone!””
Bias affects representation

- Attribute style representations
  - because each attribute may have large unbiased training set
  - even when each category does not

Farhadi et al 09; Lampert 09
Bias affects representation

- Semantic parts
  - as opposed to variance suppressing
  - because many animals have legs, vehicles have wheels, etc.
    - again, may have large unbiased training set

Farhadi et al 10
Endres et al 10
Bias affects representation

- Other kinds of semantics
  - Ramanan’s activity example
    - where you are often reveals what you are doing
    - but how do we encode where you are
      - x-y coords?
      - near the stove?
Conclusion

• Not much, if the emphasis is on size

• Collecting datasets is highly creative
  • rather than a nuisance activity
  • tools are getting better by the day

• Bias, weird frequencies are a major issue
  • There are no best practices for avoiding problems
  • May shape our representations

• Recognition problems are hard to frame
  • excess certainty may be dangerous
Are these monkeys?
One belief space about recognition

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

- Object recognition = $k$-way classification

- Current data sets ok in principle
  - improve coverage
  - collect unbiased datasets with fair coverage

- Research agenda:
  - more features, better classifiers:
  - perhaps category hierarchies for statistical leverage (tying)

Platonism?
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
Another belief space about recognition

- **Categories are highly fluid**
  - opportunistic devices to aid generalization
    - affected by current problem
  - 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
Research agenda

• What should we mean by “category”?
  • how are categories created?
  • how can multiple category systems co-exist?
  • how can we sew together categorization and utility?

• What should we report about pictures?
  • What kind of clumps of meaning should we detect?
  • What should we say about things?

• What information is important?
  • Texture, yes; but: support? shape? geometry? context?
  • Goals and intentions?
Co-existing category systems

Monkey or Plastic toy or both or irrelevant

Some of this depends on what you’re trying to do, in ways we don’t understand

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

“Sledder”
Is this one thing?
Should we cut her off her sled?
Clumps of meaning
What should we report?

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
A man stands next to a train on a cloudy day
A backpacker stands beside a green train
This is a picture of a man standing next to a green train
There are two men standing on a rocky beach, smiling at the camera.
This is a person laying down in the grass next to their bike in front of a strange white building.
Selection

- (No-one was hurt; I checked)
How many adults were on the platform and what were they doing?
What’s going to happen to the baby?

What outcome do we expect?

How are other people feeling?

What will they do?
What should we do about datasets?

- **Recognize and beware of fallacies**
  - Good datasets are big implies big datasets are good
  - If you know your problem well, you can collect an unbiased dataset

- **Always train on dataset A and test on B**
  - this isn’t the same as a train/test split of A

- **Throw away more data than we’re doing**
  - it tends to go off, and when it has gone off, it’s poisonous

- **Come up with new methods to identify and manage bias**
  - How?

- **Come up with richer notions of categorical annotation**
Conclusion

- Not much, if the emphasis is on size
  - strong classification methodologies are no substitute for thought

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous
Obtain dataset

Build features

Mess around with classifiers, probability, etc

Produce representation
Obtain dataset

Build features

Light entertainment
(the way we do it)

Mess around with classifiers, probability, etc

Produce representation
Big questions

• What signal representation should we use?

• What should we say about visual data?

Taxonomy
The Unfamiliar

• What do you say about it?
  • Attributes?

• Are many categories rare?
  •
Distributional semantics

- Most words are unusual
- Don’t know a word?
  - nearby words can tell you what it means
  - or how similar it is to a word you do know

“No; this my hand will rather the multitudinous seas incarnadine, making the green one red.”

“In one routine, describing his “ludicrously alpha” surfing instructor for the Forgetting Sarah Marshall shoot, he exclaims, “The sea were incarnadine wiv his testosterone!””
Are most things unfamiliar?

Wang ea 10; labelme data
Collective search

• Problem:
  • publish a dataset
  • people try methods, keep ones that do well
  • hence, results suffer from intense selection bias

• Bigger datasets -> weaker recognition statistics
  • Because the categories are genuinely harder?
  • Because collective search is much harder?
Fallacy

Good datasets are big
implies
Big datasets are good
Fallacy

If you know your problem well you can collect an unbiased dataset
Conclusion

• Collecting datasets is highly creative
  • rather than a nuisance activity
  • tools are getting better by the day

• Bias, weird frequencies are a major issue
  • There are no best practices for avoiding problems
  • May shape our representations

• Recognition problems are hard to frame
  • excess certainty may be dangerous
How do we assess different datasets?

- By what they are for
  - activity vs category
- By what they cover
  - many cases vs few
- By how well they represent the problem
  - in some special cases, it is easy to tell
  - what is the problem?
- By how big they are
  - easy!
Conclusion

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous
Non-parametric regression

With a broad view of “match”, “smooth”, all classifiers fit into this story
A=Image, B=Body pose

- Rosales+Sclaroff, 00; Shakhnarovich+Darrell, 03
A=Image with hole, B=fill-in

Efros+Leung, 99; Hays+Efros 07
A=picture, B=location

Hays+Efros, 08
A=motion window, B=words

Laptev Perez 2007; see also Laptev et al 08
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters.

Berg et al 04, 05; Guillaumin et al 08; Everingham et al 06; Ozkan et al 06; Zhao et al 08; Yagnik et al 07; lots of others
Input Image

Find visually similar internet images

dog, pet, animal, Dogs! Dogs! Dogs!

dog, boxer, Dogs! Dogs! Dogs!, cool dogs..

Text summarizing

dog
Dogs! Dogs! Dogs!
A = picture, B = Sentence

A man stands next to a train on a cloudy day
A backpacker stands beside a green train
This is a picture of a man standing next to a green train
There are two men standing on a rocky beach, smiling at the camera.
This is a person laying down in the grass next to their bike in front of a strange white building.

Farhadi et al 10
Recognition datasets

• **Collection strategies**
  • Web pix + fix
    • Flickr
    • Google image search
    • Microsoft image search
  • Existing collections
    • Corel
  • Photograph yourself
  • Photograph isolated, then enrich
Gotchas!

- Web pix+fix
  - Bias (more later!)
  - Might be few of the right kind (Sapp et al 08)

This difficulty probably exaggerated
Gotchas!

- **Existing collections**
  - mainly stock photo’s like Corel
  - Massive bias issues with corel
    - one can predict CD number from color histogram rather accurately (Chappelle et al, 99)

- **Photograph yourself**
  - hard work
Gotchas!

- **Enriching**
  - Use a probabilistic “model” to
    - enrich background
    - vary foreground
  - **DANGER**
    - strong unnatural high frequencies at blend
    - unnatural illumination relations
    - no surface texture distortion

- **Random**
  - Example: aspect and symmetry

Sapp Saxena Ng, 08 AAAI
Recognition datasets

• **Taxonomy strategies**
  - Choose some categories (Fei-Fei 04; Griffin 07; Everingham 06)
  - Wordnet (Deng 09)
  - Other?

• **Labelling strategies**
  - query image search, check responses (Fei-Fei 04; Griffin 07; Everingham 06)
  - tagging by volunteers
    - benevolent people (Antonio’s mom) (Russell 08)
    - game players (von Ahn 04)
  - tagging by paid annotators (Yao 07; Sorokin 08)

  ● **Go to Alex and Fei-Fei’s tutorial on Friday**

  ● active learning (Berg, 06; Li, 06; Wang 08; Collins 08)
Turk experience outside vision

- HLT-NAACL workshop 2010
  - proceedings out two weeks ago
  - competition: make a nice NLP dataset for less than $100

<table>
<thead>
<tr>
<th>Why do you complete tasks in MTurk?</th>
<th>US</th>
<th>India</th>
</tr>
</thead>
<tbody>
<tr>
<td>To spend free time fruitfully and get cash (e.g., instead of watching TV)</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>For “primary” income purposes (e.g., gas, bills, groceries, credit cards)</td>
<td>15%</td>
<td>27%</td>
</tr>
<tr>
<td>For “secondary” income purposes, pocket change (for hobbies, gadgets)</td>
<td>60%</td>
<td>37%</td>
</tr>
<tr>
<td>To kill time</td>
<td>33%</td>
<td>5%</td>
</tr>
<tr>
<td>The tasks are fun</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td>Currently unemployed or part time work</td>
<td>30%</td>
<td>27%</td>
</tr>
</tbody>
</table>
Turk experience outside vision


Figure 1: Time spent, HITs completed, and amount earned from a survey of 1,000 Turkers by Ipeirotis (2010).
Design remains hard

- When we get poor results, is it because
  - the interface is poor (e.g. confusing buttons)
  - the task is hard (e.g. mark all pixels such that ...)
  - the task is unnatural (e.g. are red cats heavier than blue dogs)
Conclusion

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- **Bias, weird frequencies are a major issue**
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous
Conclusion

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous

You can’t get away from bias by saying you must know your problem well before you collect
Big questions

• What signal representation should we use?

• What should we say about visual data?
Taxonomy/Category problems

- A choice of taxonomy is a profound commitment
  - which may enhance/distort future research

- Examples:
  - Recognition
Object recognition = k class classification

- current data sets ok,
  - improve coverage
  - collect unbiased datasets with fair coverage

- research agenda:
  - more features, better classifiers:
  - perhaps category hierarchies for statistical leverage (tying)
Are these monkeys?
Object recognition = describing what objects are like

- most current datasets
  - are largely of the wrong form
    - and no declarative data about objects
  - bias is intrinsic
    - and intertwined with representation agendas

- research agenda
  - learning by reading
  - similarity
  - coping with induction issues
  - sensible responses to objects of unknown category
  - within class variance has semantics
  - architectures, representations, semantics

Representational agenda may be driven by bias in datasets
Conclusion

• Collecting datasets is highly creative
  • rather than a nuisance activity
  • tools are getting better by the day

• Bias, weird frequencies are a major issue
  • There are no best practices for avoiding problems
  • May shape our representations

• Recognition problems are hard to frame
  • excess certainty may be dangerous

You can’t get away from bias by saying you must know your problem well before you collect