Learning Everything about Anything

Santosh Divvala
University of Washington

Joint work with Carlos Guestrin & Ali Farhadi
How can we learn *everything* that is visual about *anything*?

All possible appearance variation that our models can learn
Q1. How to gather the training data (vocabulary, images, etc.)?
Ans. Benchmark datasets e.g., PASCAL VOC

Q2. How to model the visual variance?
Ans. Philosophy of Divide & Conquer
   Examples: Viewpoint, Aspect-Ratio, Taxonomy, Phrases, Phraselets, Attributes, etc.

Human Supervision
Problem with Human Supervision

• Biased, non-comprehensive
Unbiased Look at Dataset Bias

Antonio Torralba
Massachusetts Institute of Technology
torralba@csail.mit.edu

Alexei A. Efros
Carnegie Mellon University
efros@cs.cmu.edu

Abstract

Datasets are an integral part of contemporary object recognition research. They have been the chief reason for the considerable progress in the field, not just as source of large amounts of training data, but also as means of measuring and comparing performance of competing algorithms. At the same time, datasets have often been blamed for narrowing the focus of object recognition research, reducing it to a single benchmark performance number. Indeed, some datasets, that started out as data capture efforts aimed at representing the visual world, have become closed worlds unto themselves (e.g. the Corel world, the Caltech-101 world, the PASCAL VOC world). With the focus on beating the latest benchmark numbers on the latest dataset, have we perhaps lost sight of the original purpose?

The goal of this paper is to take stock of the current state of recognition datasets. We present a comparison study using a set of popular datasets, evaluated based on a number of criteria including: relative data bias, cross-dataset generalization, effects of closed-world assumption, and sample value. The experimental results, some rather surprising, suggest directions that can improve dataset collection as well as algorithm evaluation protocols. But more broadly, the hope is to stimulate discussion in the community regard-

Figure 1. Name That Dataset: Given three images from twelve popular object recognition datasets, can you match the images with the dataset? (answer key below)
Problem with Human Supervision

• Biased, non-comprehensive

• Concept-specific expertise
“Tall Rabbit is shorter than Short Horse”
(Figure from Devi Parikh)
Problem with Human Supervision

• Biased, non-comprehensive

• Concept-specific expertise

• Scalability
The Human Annotation Tool
Lubomir Bourdev and Jitendra Malik

Phrasal Recognition Dataset
Download Phrasal Recognition Dataset (250MB)

This dataset contains 8 object categories from Pascal VOC that are suitable for studying the interactions between objects. The dataset is formatted like Pascal VOC dataset and is easy to use. This dataset contains:
- 2769 images
- 5067 bounding-box annotations
- 8 objects
- 17 visual phrases
- 120 image per visual phrase
- 1796 bounding boxes for for visual phrases
- 3271 bounding boxes for objects
- Objects:
  - person, bike, car, dog, horse, bottle, sofa, chair
Problem with Human Supervision

• Biased, non-comprehensive

• Concept-specific expertise

• Scalability

• Frozen (in time) decisions
Q1. How to gather the training data (vocabulary, images, etc.)?

Q2. How to model the visual variance?
The PASCAL Visual Object Classes (VOC) Challenge

Mark Everingham · Luc Van Gool · Christopher K. I. Williams · John Winn · Andrew Zisserman

Received: 30 July 2008 / Accepted: 16 July 2009 / Published online: 9 September 2009
© Springer Science+Business Media, LLC 2009

Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annually from 2005 to present, the challenge and its associated dataset has become accepted as the benchmark for object detection.

This paper describes the dataset and evaluation procedure. We review the state-of-the-art in evaluated methods for both classification and detection, analyse whether the methods are statistically different, what they are learning from the images (e.g. the object or its context), and what the methods find easy or confuse. The paper concludes with lessons learnt in the three year history of the challenge, and proposes directions for future improvement and extension.

Keywords Database · Benchmark · Object recognition · Object detection

1 Introduction

The PASCAL Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available dataset of images and annotation, together with standardised evaluation software; and (ii) an annual competition and workshop. The VOC2007 dataset consists of annotated consumer photographs collected from the flickr\(^1\) photo-sharing web-site. A new dataset with ground truth annotation has been released each year since 2006. There are two principal challenges: classification—"does the image contain any instances of a particular object class?" (where the object classes include cars, people, dogs, etc.), and detection—"where are the instances of a particular object class in the image (if any)?". In addition, there are two subsidiary challenges ("tasters") on pixel-level segmentation—assign each pixel a class label, and "person layout”—localise the head, hands and feet of people in the image. The challenges are issued with deadlines each year, and a workshop held to compare and discuss that year’s results and methods. The datasets and associated annotation and software are subsequently released and available for use at any time.

The objectives of the VOC challenge are twofold: first to provide challenging images and high quality annotation, together with a standard evaluation methodology—a "plug and play" training and testing harness so that performance of algorithms can be compared (the dataset component); and second to measure the state of the art each year (the competition component).

\(^1\)PASCAL stands for pattern analysis, statistical modelling and computational learning. It is an EU Network of Excellence funded under the IST Programme of the European Union.

\(^2\)http://www.flickr.com/
Table 1 Queries used to retrieve images from flickr. Words in bold show the “targeted” class. Note that the query terms are quite general—including the class name, synonyms and scenes or situations where the class is likely to occur

- **horse**, gallop, jump, buck, equine, foal, cavalry, saddle, canter, buggy, mare, neigh, dressage, trial, racehorse, steeplechase, thoroughbred, cart, equestrian, paddock, stable, farrier
- **motorbike**, motorcycle, minibike, moped, dirt, pillion, biker, trials, motorcycling, motorcyclist, engine, motocross, scramble, sidecar, scooter, trail
- **person**, people, family, father, mother, brother, sister, aunt, uncle, grandmother, grandma, grandfather, grandpa, grandson, granddaughter, niece, nephew, cousin
- **sheep**, ram, fold, fleece, shear, baa, bleat, lamb, ewe, wool, flock
- **sofa**, chesterfield, settee, divan, couch, bolster
- **table**, dining, cafe, restaurant, kitchen, banquet, party, meal
- **potted plant**, pot plant, plant, patio, windowsill, window sill, yard, greenhouse, glass house, basket, cutting, pot, cooking, grow
- **train**, express, locomotive, freight, commuter, platform, subway, underground, steam, railway, railroad, rail, tube, underground, track, carriage, coach, metro, sleeper, railcar, buffet, cabin, level crossing
- **tv/monitor**, television, plasma, flatscreen, flat screen, lcd, crt, watching, dvd, desktop, computer, computer monitor, PC, console, game
Gathering Vocabulary

- Comprehensive
- Concept-specific

Figures from newyorker.com, huffpost.com
Dependency N-grams
Dependency N-grams

_Start_ John has short black hair _END_
Dependency N-grams

_START_ John has short black hair _END_

-----------------------------------------------

Raw Ngrams

John short
John has ...
... short black hair
Dependency N-grams

Raw Ngrams

John  short
John has  ...
...  short black hair
Dependency N-grams

<table>
<thead>
<tr>
<th>Raw Ngrams</th>
<th>Annotated Ngrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>John short</td>
<td><em>START</em> John</td>
</tr>
<tr>
<td>John has</td>
<td>John NOUN</td>
</tr>
<tr>
<td>...</td>
<td>John has_VERB</td>
</tr>
<tr>
<td>... short black hair</td>
<td>hair <em>END</em></td>
</tr>
<tr>
<td>...</td>
<td>John <em>VERB</em> short</td>
</tr>
</tbody>
</table>
Dependency N-grams

Raw Ngrams

<table>
<thead>
<tr>
<th>Raw Ngrams</th>
<th>Annotated Ngrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>John short</td>
<td><em>START</em> John NOUN hair=&gt;short</td>
</tr>
<tr>
<td>John has ...</td>
<td>... John has_VERB hair=&gt;black ...</td>
</tr>
<tr>
<td>... short black hair</td>
<td>hair_END_ John_VERB_short NOUN_&lt;=has</td>
</tr>
</tbody>
</table>

Lin et al., ACL 2012
Approach

- Approximately 5000 N-grams per concept
- Several visually non-salient N-grams e.g., “last horse”, “particular horse”, etc.
Good vs. Bad N-grams

“Last Horse”

“Reining Horse”

Top Google Image Search Results
Approach

• Pruning method
  – Download thumbnail images from Google
  – Split data and train/test a (HOG+SVM) classifier
  – If A.P. < *thresh*, discard N-gram

• Reduces #N-grams to approx. 1000 (from 5000)
Superfluous List of N-grams

“Sleigh horse” ⇔ “Sledge horse”

“Plow horse” ⇔ “Plough horse”

“Eating horse” ⇔ “Grazing horse”

“Cantering horse” ⇔ “Loping horse”
Space of Visual Variance
Space of Visual Variance
Find Subset of N-grams with good Quality & Coverage (Diversity)

\[
\max_S \sum_{i \in V} d_i \cdot O(i, S)
\]
Find Subset of N-grams with good Quality & Coverage (Diversity)

$$\max_S \sum_{i \in V} d_i \times \mathcal{O}(i, S)$$

$$\mathcal{O}(i, S) = \begin{cases} 
1 & i \in S \\
1 - \prod_{j \in S} (1 - e_{i,j}) & i \notin S
\end{cases}$$
Find Subset of N-grams with good Quality & Coverage (Diversity)

\[
\max_S \sum_{i \in V} d_i \ast \mathcal{O}(i, S)
\]

\[
\mathcal{O}(i, S) = \begin{cases} 
1 & \text{if } i \in S \\
1 - \prod_{j \in S} (1 - e_{i,j}) & \text{if } i \notin S
\end{cases}
\]

such that \( |S| \leq k \)
Sample Merging Results

“Iran Majlis” ⇔ “Iran Parliament”

“Angry Shouting” ⇔ “Angry Screaming”

“Cute Doctor” ⇔ “Women Doctor”

“Iran Leader” ⇔ “Iran Khomeini”

“Angry Mob” ⇔ “Angry Protestors” ⇔ “Angry Crowd”

“Doctor Explaining” ⇔ “Doctor Discussing” ⇔ “Consulting Doctor”
Approach

Discovering Vocabulary

User Input "horse"

Search Google books Ngrams

Ngram list
- eating horse
- last horse
- grazing horse
- reining horse

Pruning

Discard visually non-salient Ngrams e.g., "last horse"

Merge synonyms e.g., <eating horse, grazing horse>

Super Ngrams
Approach

- Download 200 images per super N-gram
- Discard near-duplicates and bad-aspect images
- Split data for training and validation
Approach

- Train separate DPM per super N-gram
- Initialize DPM with bounding boxes as full images
PASCAL VOC vs. Google Image Search

Sample PASCAL VOC Chair Images

“Needlepoint Chair”
“Willow Chair”
“Lincoln Chair”
Approach

- Train separate DPM per super N-gram
- Initialize DPM with bounding boxes as full images
- Components based on appearance clustering
Components act as noise sinks

"Reining Horse Components"
Components act as noise sinks

Noise sinks
Top Google Image Search Results

“Jumping Horse”

“Hunter Horse”
Approach

Discovering Vocabulary

User Input "horse"

Search

Ngram list
- eating horse
- last horse
- grazing horse
- reining horse

Pruning

Discard visually non-salient Ngrams e.g., "last horse"

Merge synonyms e.g., <eating horse, grazing horse>

Super Ngrams

Download Images

Learn Mixture of Root Filter Components

Prune Noisy ones, Merge Similar ones, Add Parts
Approach

Discovering Vocabulary

User Input "horse"

Search Google books Ngrams

Ngram list:
- eating horse
- last horse
- grazing horse
- reining horse

Pruning

Discard visually non-salient Ngrams e.g., "last horse"

Merge synonyms e.g., <eating horse, grazing horse>

Super Ngrams

Model Learning

Download Images Google

Learn Mixture of Root Filter Components

Prune Noisy ones, Merge Similar ones, Add Parts

Horse Detector
Results

• Amazon EC2-friendly framework
• 100+ concepts, 15000+ variations, 3Million images
• List includes Objects, Actions, Scenes, Events, Places, Emotions, Celebrities, Professions, Attributes, etc.
Results

• Amazon EC2-friendly framework
• 100+ concepts, 15000+ variations, 3 Million images
• List includes Objects, Actions, Scenes, Events, Places, Emotions, Celebrities, Professions, Attributes, etc.
• Online system available: http://goo.gl/O99uZ2
Learn Everything about Anything

Enter a concept and get its extensive model!

* Required

Enter a concept *
concept can be an object (e.g., apple), action (e.g., jumping), place (e.g., london), emotion (e.g., happy), etc

Select its parts of speech *
For example, noun for objects, verb for actions, adjective for emotions, or other

Please Note:

1. If you submit a query, please check back after approximately 24 hours. Your model will be available here: http://goo.gl/Yvg9Cc

2. To browse all results, please see: http://goo.gl/AIAT3T

3. This system is doubly-anonymous. We are not recording/using any analytics data about incoming IPs or traffic.

Submit

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google.
Results

- EC2-friendly framework
- 100+ concepts, 15000+ variations, 3 Million images
- List includes Objects, Actions, Scenes, Events, Places, Emotions, Celebrities, Professions, Attributes, etc.
- Online system available: [http://goo.gl/O99uZ2](http://goo.gl/O99uZ2)
- Quantitative Results: Object & Action Detection

Uses PASCAL VOC 2007 training images

Uses objectness [Ferrari et al., 2010] for initialization (High A.P. for bike, car, motorbike and train)

Does not work for objects that are small or in cluttered scenes e.g., bottle, chair, TV, etc.

- Manually chosen Youtube videos
- Uses Objectness for initialization
- Does not work on static (10/20) classes e.g., bottle, TV, etc
Baseline methods use weak supervision (images, videos, objectness)

Our method is “Unsupervised” => Webly-Supervised

Beats SOA on 13/20 classes; impressive results for bottle, chair, sheep, tv

Almost on part with supervised DPM on 5/20 classes!
PASCAL VOC 2010 Action Detection

**VOC Challenge**

Given the person bounding box in an image, identify the action

**Our Goal**

Given a image, identify *and localize* the action in an “unsupervised” approach
PASCAL VOC 2010 Action Detection

- Use baseline as reported in [Phraselets, ECCV 2012]
PASCAL VOC 2010 Action Detection

Average Precision

- Use baseline as reported in [Phraselets, ECCV 2012]
Potential Applications: Co-segmentation
Potential Applications: Co-reference Resolution

Indira Gandhi was the third Indian prime minister. Mohandas Gandhi was the father of Indian Nationalism. Mrs. Gandhi was inspired by Mahatma Gandhi’s writings.
Potential Applications: Co-reference Resolution

..Indira Gandhi was the third Indian prime minister. Mohandas Gandhi was the father of Indian Nationalism. Mrs. Gandhi was inspired by Mahatma Gandhi’s writings..

“Indira Gandhi” <=> “Mrs. Gandhi”

“Mahatma Gandhi” <=> “Mohandas Gandhi”
Potential Applications: Temporal Evolution of Concepts
Potential Applications: Temporal Evolution of Concepts

1900 car

1925 car

1975 car

2000 car
Potential Applications: Deeper Image Interpretation
Potential Applications:
Deeper Image Interpretation

Which bounding box to pick?
Thank You