CS 2770: Computer Vision
Unsupervised Learning, Discovery
Active Learning

Prof. Adriana Kovashka
University of Pittsburgh
April 6, 11, 13, 2017
Motivation

- So far we’ve assumed access to plentiful labeled data
Crowdsourcing

Task: Dog?

Workers

Broker

Answer: Yes

Pay: $0.01

www.mturk.com

Is this a dog?

- Yes
- No

$0.01

Alex Sorokin
Crowdsourcing via games

• The ESP Game
  – Two-player online game
  – Partners don’t know each other and can’t communicate
  – Objective of the game: type the same word
  – The only thing in common is an image

THE ESP GAME

PLAYER 1

GUESSING: CAR
GUESSING: HAT
GUESSING: KID
SUCCESS!
YOU AGREE ON CAR

PLAYER 2

GUESSING: BOY
GUESSING: CAR
SUCCESS!
YOU AGREE ON CAR

Motivation

• So far we’ve assumed access to plentiful labeled data
• What if we have limited or no labeled data?
• One approach: learn from unlabeled data (unsupervised learning)
  – Mine for interesting patterns (discovery)
• Another approach: carefully choose which data to label (active learning, human-in-the-loop)
Plan for last few lectures

• Unsupervised learning
  – From surrounding regions (ICCV 2015)
  – From temporal order (ECCV 2016)
  – From ego-motion (ICCV 2015)
  – From surrounding objects (CVPR 2010)

• Visual discovery
  – Elements of architectural styles (SIGGRAPH 2012)
  – Gradual style changes in time/space (ICCV 2013)

• Active human-in-the-loop learning
  – Attributes for fine-grained recognition (ECCV 2010)

• My lab’s research (if time)
Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei Efros and Abhinav Gupta

ICCV 2015
ImageNet + Deep Learning

Do we even need semantic labels?

Do we even need this task?


Context as Supervision
[Collobert & Weston 2008; Mikolov et al. 2013]

house, where the professor lived without his wife and child; or so he said jokingly sometimes: “Here’s where I live. My house.” His daughter often added, without resentment, for the visitor’s information, “It started out to be for me, but it’s really his.” And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked “Kitty” and half full of eternal milk; but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter’s preference was for the store-bought gimmicks and appliances, the toasters and carpet sweepers of Lilliput, but she knew that most adult visitors would
Context Prediction for Images

Semantics from a non-semantic task

Relative Position Task

8 possible locations

Patch Embedding

Input

Nearest Neighbors

Note: connects *across* instances!

Architecture

Softmax loss
Fully connected

Fully connected

Fully connected
Max Pooling
Convolution
Convolution
Convolution
LRN
Max Pooling
Convolution
LRN
Max Pooling
Convolution
Tied Weights

Fully connected
Max Pooling
Convolution
Convolution
Convolution
LRN
Max Pooling
Convolution
LRN
Max Pooling
Convolution

Patch 1

Patch 2

What is learned?

Pre-Training for R-CNN

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Pre-train on relative-position task, w/o labels


[Girshick et al. 2014]
VOC 2007 Performance
(pretraining for R-CNN)

<table>
<thead>
<tr>
<th>Method</th>
<th>% Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet Labels</td>
<td>54.2</td>
</tr>
<tr>
<td>Ours</td>
<td>46.3</td>
</tr>
<tr>
<td>No Pretraining</td>
<td>40.7</td>
</tr>
</tbody>
</table>

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert
ECCV 2016
Fig. 1: (a) A video imposes a natural temporal structure for visual data. In many cases, one can easily verify whether frames are in the correct temporal order (shuffled or not). Such a simple sequential verification task captures important spatiotemporal signals in videos. We use this task for unsupervised pre-training of a Convolutional Neural Network (CNN). (b) Some examples of the automatically extracted positive and negative tuples used to formulate a classification task for a CNN.
Fig. 2: (a) We sample tuples of frames from high motion windows in a video. We form positive and negative tuples based on whether the three input frames are in the correct temporal order. (b) Our triplet Siamese network architecture has three parallel network stacks with shared weights up to the fc7 layer. Each stack takes a frame as input, and produces a representation at the fc7 layer. The concatenated fc7 representations are used to predict whether the input tuple is in the correct temporal order.
Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Initialization</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>Random</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>50.2</strong></td>
</tr>
<tr>
<td>HMDB51</td>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>UCF Supervised</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>
Learning image representations tied to ego-motion

Dinesh Jayaraman and Kristen Grauman

ICCV 2015
The kitten carousel experiment [Held & Hein, 1963]

Key to perceptual development: self-generated motion + visual feedback
Problem with today’s visual learning

**Status quo:** Learn from “disembodied” bag of labeled snapshots.

**Our goal:** Learn in the context of *acting* and *moving* in the world.

Our idea: **Ego-motion ↔ vision**

**Goal:** Teach computer vision system the connection: “how I move” ↔ “how my visual surroundings change”

Ego-motion motor signals + Unlabeled video

Ego-motion ↔ vision: view prediction

After moving:

Ego-motion ↔ vision for recognition

Learning this connection requires:

➢ Depth, 3D geometry
➢ Semantics
➢ Context

Also key to recognition!

Can be learned without manual labels!

Our approach: unsupervised feature learning using egocentric video + motor signals

Approach idea: Ego-motion equivariance

**Invariant features**: unresponsive to some classes of transformations

\[ z(gx) \approx z(x) \]

**Equivariant features**: predictably responsive to some classes of transformations, through simple mappings (e.g., linear)

\[ z(gx) \approx M_g z(x) \]

Invariance **discards** information; equivariance **organizes** it.

Approach idea: Ego-motion equivariance

**Training data**
Unlabeled video + motor signals

**Equivariant embedding**
organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Approach overview

Our approach: unsupervised feature learning using egocentric video + motor signals

1. Extract training frame pairs from video
2. Learn ego-motion-equivariant image features
3. Train on target recognition task in parallel

Training frame pair mining

Discovery of ego-motion clusters

Ego-motion equivariant feature learning

**Given:**

\[ x_i \]

\[ g \]

\[ g x_i \]

**Desired:** for all motions \( g \) and all images \( x \),

\[ z_\theta(g x) \approx M_g z_\theta(x) \]

**Unsupervised training**

\[ z_\theta(x_i) \rightarrow M_g \rightarrow || M_g z_\theta(x_i) - z_\theta(g x_i) ||_2 \]

**Supervised training**

\[ z_\theta(x_i) \rightarrow M_g \rightarrow \max \text{ loss } L_C(x_k, y_k) \]

\[ \theta, M_g \text{ and } W_{z_\theta(g x_i)} \text{ jointly trained} \]

Summary

**APPROACH**

- Ego-motion training pairs
- Neural network training
- Equivariant embedding

**RESULTS**

- Scene and object recognition
- Next-best view selection

- Football field?
- Pagoda?
- Airport?
- Cathedral?
- Army base?

- Cup
- Frying pan

Results: Recognition

Learn from **unlabeled car video** (KITTI)

Exploit features for **static scene classification**
(SUN, 397 classes)

Apse, Window seat, Art school, Library, Auditorium, Bus interior, Cathedral, Freeway, Guardhouse

Geiger et al, IJRR '13

Xiao et al, CVPR '10

Results: Recognition

Do ego-motion equivariant features improve recognition?

Up to 30% accuracy increase over state of the art!

Object-Graphs for Context-Aware Category Discovery

Yong Jae Lee and Kristen Grauman
CVPR 2010
Goal

- Discover *new* object categories, based on their relation to categories for which we have trained models

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Existing approaches

Previous work treats unsupervised visual discovery as an appearance-grouping problem.

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Our idea

How can seeing previously learned objects in novel images help to discover new categories?

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Our idea

Discover visual categories within unlabeled images by modeling interactions between the unfamiliar regions and familiar objects.

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Learn “Known” Categories

- Offline: Train region-based classifiers for $N$ “known” categories using labeled training data.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Identifying Unknown Objects

Input: unlabeled pool of novel images

Compute multiple-segmentations for each unlabeled image

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Identifying Unknown Objects

- For all segments, use classifiers to compute posteriors for the $N$ “known” categories.
- Deem each segment as “known” or “unknown” based on resulting entropy.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graphs

• Model the topology of category predictions relative to the unknown (unfamiliar) region.
• Incorporate uncertainty from classifiers.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graphs

An unknown region within an image

Closest nodes in its object-graph

- Consider spatially near regions above and below, record distributions for each known class.

\[ g(s) = [ H_0(s), H_1(s), \ldots, H_R(s) ] \]

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graph matching

- Object-graphs are very similar $\rightarrow$ produces a strong match

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graph matching

- Object-graphs are partially similar $\rightarrow$ produces a fair match

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Clusters from region-region affinities

\[ K(s_i, s_j) = K_{app}(s_i, s_j) + K_{obj-graph}(s_i, s_j) \]

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object Discovery Accuracy

- Four datasets

- Multiple splits for each dataset; varying categories and number of knowns/unknowns

- Train 40% (for known categories), Test 60% of data

- Textons, Color histograms, and pHOG Features

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object Discovery Accuracy

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Example Object-Graphs

- Color in superpixel nodes indicate the predicted known category.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Examples of Discovered Categories

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Plan for last few lectures

• Unsupervised learning
  – From surrounding regions (ICCV 2015)
  – From temporal order (ECCV 2016)
  – From ego-motion (ICCV 2015)
  – From surrounding objects (CVPR 2010)

• Visual discovery
  – Elements of architectural styles (SIGGRAPH 2012)
  – Gradual style changes in time/space (ICCV 2013)

• Active human-in-the-loop learning
  – Attributes for fine-grained recognition (ECCV 2010)

• My lab’s research (if time)
What Makes Paris Look like Paris?

Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, Alexei Efros

SIGGRAPH 2012
One of these is from Paris
Raise your hand if...

...this is Paris

Raise your hand if...

We showed 20 subjects:
- 100 Random Street View Images
- 50 from Paris
- They classified Paris vs non-Paris
- Accuracy: 79%

How do they know?

We showed 20 subjects:
- 100 Random Street View Images
- 50 from Paris
- They classified Paris non-Paris
- Accuracy: 79%

How do they know?

Our Goal:

*Given a large geo-tagged image dataset, we automatically discover visual elements that characterize a geographic location.*

Why might this be a useful task?

Our Hypothesis

• The visual elements that capture Paris:
  – Frequent: Occur often in Paris
  – Discriminative: Are not found outside Paris

Step 1: Nearest Neighbors for Every Patch
Using normalized correlation of HOG features as a distance metric

patch

nearest neighbors

Step 2: Find the Parisian Clusters by Sorting

Good Patches may have Bad Neighbors!

- The naïve distance metric gives equal weight to the vertical bar and the sign.

Step 3: Updating the Similarity Function

- Learn a similarity function that separates Paris from not-Paris
  - I.e. reweight the dimensions of the feature space
  - Recast problem as classification & use SVMs

Resulting Matches

Approach Summary

1. A Cluster for Every Patch

Our Approach

1. A cluster for every patch
2. Find clusters that are mostly Parisian

1. A cluster for every patch
2. Find clusters that are mostly Parisian
3. Refine clusters by making them more Parisian

Paris: A Few Top Elements

Elements from Prague

Elements from London

Elements from Barcelona

In the U.S.

Elements from San Francisco

Elements from Boston

Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time

Yong Jae Lee, Alexei A. Efros, and Martial Hebert

ICCV 2013
Our Goal

• Mine mid-level visual elements in temporally- and spatially-varying data and model their “visual style”

when?
Historical dating of cars

where?
Geolocalization of StreetView images

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Key Idea

1) Establish connections

2) Model style-specific differences

“closed-world”

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

- Sample patches and compute nearest neighbors

[Dalal & Triggs 2005, HOG]

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

Patch

Nearest neighbors

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

<table>
<thead>
<tr>
<th>Patch</th>
<th>Nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1929</td>
<td>1927 1929 1923 1930</td>
</tr>
<tr>
<td></td>
<td>1946 1948 1940 1939 1949</td>
</tr>
</tbody>
</table>

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

(a) Peaky (low-entropy) clusters

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

(b) Uniform (high-entropy) clusters

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

- Take top-ranked clusters to build correspondences

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

- Train a detector (HoG + linear SVM) [Singh et al. 2012]

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

Top detection per decade

[Singh et al. 2012]

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

• We expect style to change gradually...

1920s

1930s

1940s

Natural world “background” dataset

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

Top detection per decade

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Making visual connections

Top detection per decade

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Results: Example connections

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Training style-aware regression models

Regression model 1

Regression model 2

- Support vector regressors with Gaussian kernels
- Input: HOG, output: date/geo-location

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Training style-aware regression models

- Train image-level regression model using outputs of visual element detectors and regressors as features

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Results: Date/Geo-location prediction

Crawled from www.cardatabase.net

- 13,473 images
- Tagged with year
- 1920 – 1999

Crawled from Google Street View

- 4,455 images
- Tagged with GPS coordinate
- N. Carolina to Georgia

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Results: Date/Geo-location prediction

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>Doersch et al. ECCV, SIGGRAPH 2012</th>
<th>Spatial pyramid matching</th>
<th>Dense SIFT bag-of-words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cars</strong></td>
<td><strong>8.56</strong> (years)</td>
<td>9.72</td>
<td>11.81</td>
<td>15.39</td>
</tr>
<tr>
<td><strong>Street View</strong></td>
<td><strong>77.66</strong> (miles)</td>
<td>87.47</td>
<td>83.92</td>
<td>97.78</td>
</tr>
</tbody>
</table>

Mean Absolute Prediction Error

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Plan for last few lectures

• Unsupervised learning
  – From surrounding regions (ICCV 2015)
  – From temporal order (ECCV 2016)
  – From ego-motion (ICCV 2015)
  – From surrounding objects (CVPR 2010)

• Visual discovery
  – Elements of architectural styles (SIGGRAPH 2012)
  – Gradual style changes in time/space (ICCV 2013)

• Active human-in-the-loop learning
  – Attributes for fine-grained recognition (ECCV 2010)

• My lab’s research (if time)
Crowdsourcing

Training Labels → Find images near decision boundary

Training Images:
- Show images, Collect labels
- Find images near decision boundary

Image Features → Classifier Training → Trained Classifier

Unlabeled Images:
- Find images near decision boundary

James Hays
Active Learning

Traditional active learning reduces supervision by obtaining labels for the most informative or uncertain examples first.

Multi-Level Active Visual Learning

Choose not only *which images* to label, but *at what level* to label them.

**Weak labels:** informing about presence of an object

- Phone
- Phone
- Not Phone
- Not Phone

**Strong labels:** outlines demarking the object

**Stronger labels:** informing about parts of objects

Sudheendra Vijayanarasimhan
Best use of manual resources may call for combination of annotations at different levels.
Choice must balance cost of varying annotations with their information gain.
Visual Recognition With Humans in the Loop

Steve Branson, Catherine Wah, Boris Babenko, Florian Schroff, Peter Welinder, Pietro Perona, Serge Belongie
ECCV 2010
What kind of bird is this?

Parakeet Auklet

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Categories of Recognition

Basic-Level
- Airplane? Chair? Bottle? ...

Easy for Humans
Hard for computers

Subordinate
- American Goldfinch? Indigo Bunting?...

Hard for Humans
Hard for computers

Parts & Attributes
- Yellow Belly? Blue Belly?...

Easy for Humans
Hard for computers

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Recognition With Humans in the Loop

- Computers: reduce number of required questions
- Humans: drive up accuracy of vision algorithms

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Example Questions

You will be asked to answer a series of questions based on identifying visual features from the bird image on the left. Closely follow the specific instructions for each question. Holding the mouse over each selectable option for 1 second will provide additional instructions or examples.

What is the color of the underparts of the bird?

Select at least one. If the underparts aren't visible, make your best guess, then select 'Guessing'. If the color is a mixture of two colors, select both (e.g., for blue-green select blue and green). If the underparts have multiple regions or patterns with multiple colors, select all relevant colors (e.g., for yellow with black stripes, select yellow and black).

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Example Questions

You will be asked to answer a series of questions based on identifying visual features from the bird image on the left. Closely follow the specific instructions for each question. Holding the mouse over each selectable option for 1 second will provide additional instructions or examples.

What is the pattern of the breast of the bird?

Select one. If the breast isn't visible, make your best guess, then select "Guessing".
Example Questions

You will be asked to answer a series of questions based on identifying visual features from the bird image on the left. Closely follow the specific instructions for each question. Holding the mouse over each selectable option for 1 second will provide additional instructions or examples.

What is the shape of the bill/beak?

Select one. If the beak isn't visible, make your best guess, then select "Guessing".

- All-purpose
- Cone
- Curved (up or down)
- Dagger
- Hooked
- Hooked Seabird
- Needle
- Spatulate
- Specialized

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Basic Algorithm

Input Image \((x)\)

Computer Vision

Max Expected Information Gain

Question 1:
Is the belly black?

A: NO

\(p(c \mid x, u_1)\)

Max Expected Information Gain

A: YES

\(p(c \mid x, u_1, u_2)\)

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Some Notation

- $c \in \{1,...C\}$: Bird type
- $Q = \{q_1,...q_n\}$: Set of possible questions
- $a_i \in A_i$: Possible answers to question $i$
- $r_i \in V$: Possible confidence in answer $i$ (Guessing, Probably, Definitely)
- $u_i = (a_i, r_i)$: User response
- $U^t$: History of user responses at time $t$
Question Selection

- **Select** the question (e.g. “What color is the belly?”) that gives the maximum info gain (entropy reduction)

\[
I(c; u_i | x, U^{t-1}) = \sum_{u_i \in A_t \times V} p(u_i | x, U^{t-1}) \left( H(c | x, u_i \cup U^{t-1}) - H(c | x, U^{t-1}) \right)
\]

- Probability of obtaining response \( u_i \) to evaluated question given image and response history
- Entropy when response is added to history
- Entropy over class at this iteration (before response to question is added to history)

where

\[
H(c | x, U^{t-1}) = -\sum_{c=1}^{C} p(c | x, U^{t-1}) \log p(c | x, U^{t-1})
\]
Probability of class labels

\[ p(c \mid x, u_1, u_2 \ldots u_t) \]

- Object Class
- Image
- Sequence of user responses

\[ \approx p(u_1, u_2 \ldots u_t \mid c) \cdot p(c \mid x)/Z \]

- Model of user responses
- Computer vision estimate (e.g. from an SVM)

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Modeling User Responses

Assume: 

$$p(u_1, u_2 \ldots u_t \mid c) \approx \prod_{i=1 \ldots t} p(u_i \mid c)$$

Then: 

$$p(u_i \mid x, U^{t-1}) = \sum_{c=1}^{C} p(u_i \mid c)p(c \mid x, U^{t-1})$$

Estimate $p(u_i \mid c)$ using Mechanical Turk statistics

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Birds 200 Dataset

• 200 classes, 6000+ images, 288 binary attributes

Black-footed Albatross
Groove-Billed Ani
Parakeet Auklet
Field Sparrow
Vesper Sparrow

Arctic Tern
Forster’s Tern
Common Tern
Baird’s Sparrow
Henslow’s Sparrow

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Results

Users drive performance: 19% $\rightarrow$ 68%

Fewer questions asked if CV used

Just Computer Vision
19%

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010
Examples

User Input Helps Correct Computer Vision

Magnolia Warbler

Is the breast pattern solid?
no (definitely)

Magnolia Warbler

Branson et al., “Visual Recognition With Humans in the Loop”, ECCV 2010