CS 1699: Intro to Computer Vision
Interactive Image Search with Attributes

Prof. Adriana Kovashka
University of Pittsburgh
November 5, 2015
We Need Search to Access Visual Data

- 144,000 hours of video uploaded to YouTube daily
- 4.5 million photos uploaded to Flickr daily
- 150 million top-level domains
- This data pool is too big to simply browse
Some Example Image Search Problems

“Who did the witness see at the crime scene?”

“Find all images depicting shiny objects.”

“Which plant is this?”
How Is Image Search Done Today?

- Keywords work well for categories with known name.
How Is Image Search Done Today?

- Unclear how to search without name or image
- Keywords are not enough!
Solution? **Interactively Describing** Visual Content

- It’s a yellow fruit.
- It has spikes.
- It is green on the inside.

Is it a lemon?

Is it a durian?

Is it a horned melon?
Solution? **Interactively Describing Visual Content**

- It’s a yellow fruit.
- It has spikes.
- It is green on the inside.

- Is it a lemon?
- Is it a durian?
- Is it a horned melon?
My Goal: Precise Communication for Search

- Potential to communicate more precisely the desired visual content
- Iteratively refine the set of retrieved images
Key Questions

1) How to open up communication?
   - Target is more smiling than

2) How to account for ambiguity of search terms?
How is *Interactive* Search Done Today?

**Keywords**

+ binary relevance feedback

- Traditional binary feedback imprecise; allows only coarse communication between user and system

Our Idea: Search via Comparisons

• Allow user to “whittle away” irrelevant images via comparative feedback on properties of results

“Like this... but with curlier hair”

Kovashka, Parikh, and Grauman, CVPR 2012
Visual Attributes

• High-level semantic properties shared by objects

yellow, spiky, smooth, natural, perspective, open, smiling, large-lips, long-hair, ornaments, metallic, high heel, red
Why Infer Properties

We want to be able to infer something about unfamiliar objects

If we can infer category names…

Familiar Objects

Cat  Horse  Dog

New Object

???

Farhadi et al., CVPR 2009
Why Infer Properties

We want to be able to infer something about unfamiliar objects

If we can infer properties…

Familiar Objects

- Has Stripes
- Has Ears
- Has Eyes
- Has Four Legs
- Has Mane
- Has Tail
- Brown
- Muscular
- Has Snout

New Object

- Has Stripes (like cat)
- Has Mane and Tail (like horse)
- Has Snout (like horse and dog)

*Farhadi et al., CVPR 2009*
Binary Attributes

- bright / not bright
- smiling / not smiling
- natural / not natural
Relative Attributes

We need ability to **compare** images by attribute “strength”
Learning Relative Attributes

• At test time, predict attribute strength of each database image
  • Input: Image features $x$
  • Output: Real-valued attribute strength $a_m(x)$

• At training time, learn a mapping between image features and attribute strength
  • Input: Pairs of ordered images with features
  • Output: Ranking functions $a_1, \ldots, a_M$

Parikh and Grauman, ICCV 2011
Learning Relative Attributes

• We want to learn a spectrum (ranking model) for an attribute, e.g. “brightness”.
• Supervision from human annotators consists of:

\[ O_m : \{ (\text{shoe}, \sim, \text{shoe}) , \ldots \} \]

\[ E_m : \{ (\text{shoe}, \sim, \text{shoe}) , \ldots \} \]

\[ \text{Ordered pairs} \]
\[ \text{Similar pairs} \]

Parikh and Grauman, ICCV 2011
Learning Relative Attributes

Learn a ranking function

\[ a_m(x_i) = \mathbf{w}_m^T x_i \]

that best satisfies the constraints:

\[ \forall (i, j) \in O_m : \mathbf{w}_m^T x_i > \mathbf{w}_m^T x_j \]

\[ \forall (i, j) \in E_m : \mathbf{w}_m^T x_i = \mathbf{w}_m^T x_j \]
Learning Relative Attributes

Max-margin learning to rank formulation

$$\min \left( \frac{1}{2} \| \mathbf{w}_m^T \|_2^2 + C \left( \sum \xi_{ij}^2 + \sum \gamma_{ij}^2 \right) \right)$$

s.t. \( \mathbf{w}_m^T (\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \)

\( |\mathbf{w}_m^T (\mathbf{x}_i - \mathbf{x}_j)| \leq \gamma_{ij} \)

\( \xi_{ij} \geq 0; \gamma_{ij} \geq 0 \)

Image → Relative attribute score

Parikh and Grauman, ICCV 2011; Joachims, KDD 2002
We need ability to **compare** images by attribute “strength”
User: “I want something more natural than this.”

Update relevance scores

Kovashka, Parikh, and Grauman, CVPR 2012
WhittleSearch with Relative Attribute Feedback

“I want something more natural than this.”

“I want something less natural than this.”

“I want something with more perspective than this.”

Kovashka, Parikh, and Grauman, CVPR 2012
Query: “I want a bright, open shoe that is short on the leg.”

Round 1:

- More open than
- More bright in color than
- Less ornaments than

Selected feedback:
- Less high at the heel than

Round 2:

- More formal than
- More bright in color than

Round 3:

- Higher at the heel than
- More open than

Match:
- Pair of shoes with open design and bright color.
Datasets

Shoes [Berg10, Kovashka12]:
14,658 shoe images;
10 attributes:
“pointy”, “bright”, “high-heeled”, “feminine” etc.

OSR [Oliva01]:
2,688 scene images;
6 attributes:
“natural”, “perspective”, “open-air”, “close-depth” etc.

PubFig [Kumar08]:
772 face images;
11 attributes:
“masculine”, “young”, “smiling”, “round-face”, etc.

Data from 147 users
WhittleSearch Results (Summary)

- **Binary feedback** represents status quo [Rui et al. 1998, Cox et al. 2000, Ferecatu & Geman 2007, ...]
- **WhittleSearch** finds relevant results faster than traditional binary feedback
Find visually appealing shoes.

Our recommendations are based on appearance, not click-tracking.
Impact of **WhittleSearch**: Adobe Font Selection

- Users retrieve fonts that match requested attributes
- Fonts sorted by relative attribute scores

*O’Donovan et al., Exploratory Font Selection using Crowdsourced Attributes, SIGGRAPH 2014*
The most relevant images might not be most informative.

Existing active methods largely focus on binary relevance feedback and suffer from expensive selection procedures.

Idea: Attribute Pivots for Guiding Feedback

- **Relative** 20 questions game
- Select series of *most informative* visual comparisons that user should make to help deduce target

*Kovashka and Grauman, ICCV 2013*
Idea: Attribute Pivots for Guiding Feedback

Are the shoes you seek more or less feminine than [shoe image]?

**Traditional active approach**

- #1 pointy
- #2 open
- #3 bright
- ...
- #M feminine

**Using pivots**

- #1 \{ pointy, [image] \}
- #2 \{ open, [image] \}
- #3 \{ bright, [image] \}
- ...
- #M \{ feminine, [image] \}

**O(MN) comparisons total**

\(M: \sim 11\) \quad \(N: \sim 15k\)

**O(M) comparisons total**

\(M: \sim 11\)
Key Questions

1) How to open up communication?

Target is more smiling than

2) How to account for ambiguity of search terms?
Standard Attribute Learning: One Generic Model

- Learn a **generic** model by pooling training data regardless of the annotator identity
- Inter-annotator disagreement treated as noise
Problem: One Model *Does Not* Fit All

- There may be valid perceptual differences within an attribute, yet existing methods assume monolithic attribute sufficient

Personalization from Scratch

- Collect data from the user and train a classifier
Idea: Learn User-Specific Attributes

Standard approach:
- Treat learning perceived attributes as an adaptation problem
- Adapt generic attribute model with minimal user-specific labeled examples

Our idea:
- Treat learning perceived attributes as an adaptation problem
- Adapt generic attribute model with minimal user-specific labeled examples

Kovashka and Grauman, ICCV 2013
Learning Adapted Attributes

- Adapting binary attribute classifiers:

Given user-labeled data \( D_b = \{x_i, y_i\}_{i=1}^{N} \) and generic model \( \mathbf{w}'_b \), learn adapted model \( \mathbf{w}_b \),

\[
\min_{\mathbf{w}_b} \frac{1}{2} \| \mathbf{w}_b - \mathbf{w}'_b \|^2 + C \sum_{i=1}^{N} \xi_i,
\]

subject to \( y_i x_i^T \mathbf{w}_b \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \)

J. Yang et al. ICDM 2007
Learning Adapted Attributes

“formal”

Adapted boundary

Generic boundary

“not formal”
Datasets

**SUN Attributes**

[Patterson12]:
14,340 scene images
12 attributes:
“sailing”, “hiking”, “vacationing”, “open area”, “vegetation”, etc.

**Shoes [Berg10, Kovashka12]:**
14,658 shoe images;
10 attributes:
“pointy”, “bright”, “high-heeled”, “feminine” etc.
Adapted Attribute Accuracy

- Result over 3 datasets, 32 attributes, and 75 total users
- Our user-adaptive method most accurately captures perceived attributes
Personalizing Image Search with Adapted Attributes

- generic
- generic+
- user-exclusive
- user-adaptive

Match rate

Shoes-Binary

SUN

“white shiny heels”

“shinier than”
Which Images Most Influence Adaptation?

- pointy
- open
- bright
- ornamented
- shiny
- high-heeled
- long
- formal
- sporty
- feminine
- sailing
- vacationing
- hiking
- camping
- socializing
- shopping
- natural light
- clouds
- vegetation
- cold
- open area
- horizon far
Attribute Model Spectrum

**Generic**
- No personalization
- Assumes all users have the same notion of the attribute
- Robustness to noise via majority vote

**User-adaptive**
- Personalized to each user
- Assumes each user has a unique notion of the attribute
- No robustness to noise
Problem: User-Specific Extreme

- Different groups of users might subscribe to different shades of meaning of an attribute term
- How can we discover these shades automatically?
Our Idea: Discovering Shades of Attributes

- Discover “schools of thought” among users based on latent factors behind their use of attribute terms
- Allows discovery of the attribute’s “shades of meaning”

Kovashka and Grauman, IJCV 2015
## Approach: Collecting Crowd Labels

- Show users definitions but no example images
- Ask them to label images with presence of attribute
- Also ask for explanations of some responses

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointy</td>
<td>having a comparatively sharp point, or having numerous pointed parts</td>
</tr>
<tr>
<td>Open</td>
<td>having interspersed gaps, spaces, or intervals</td>
</tr>
<tr>
<td>Ornate</td>
<td>made in an intricate shape or decorated with complex patterns</td>
</tr>
<tr>
<td>Comfortable</td>
<td>providing physical comfort, ease and relaxation</td>
</tr>
<tr>
<td>Formal</td>
<td>designed for wear or use at elaborate ceremonial or social events</td>
</tr>
<tr>
<td>Brown</td>
<td>the color of, for example, chocolate and coffee</td>
</tr>
<tr>
<td>Fashionable</td>
<td>conforming to the current fashion; stylish; trendy; modern</td>
</tr>
<tr>
<td>To clutter</td>
<td>to make disorderly or hard to use by filling or covering with objects</td>
</tr>
<tr>
<td>To soothe</td>
<td>to bring comfort, composure, or relief</td>
</tr>
<tr>
<td>Open (area)</td>
<td>affording unobstructed passage or view</td>
</tr>
<tr>
<td>Modern</td>
<td>characteristic or expressive of recent times or the present; contemporary</td>
</tr>
<tr>
<td>Rustic</td>
<td>of, relating to, or typical of country life or country people</td>
</tr>
</tbody>
</table>
**Approach: Recovering Latent Factors**

Is the attribute “open” present in the image?

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Annotator 2</td>
<td>?</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Annotator 3</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Annotator 4</td>
<td>0</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>Annotator 5</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Annotator 6</td>
<td>?</td>
<td>0</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Annotator 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Factor 1 (toe?)</th>
<th>Factor 2 (heel?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1</td>
<td>0.85</td>
<td>0.12</td>
</tr>
<tr>
<td>Annotator 2</td>
<td>0.72</td>
<td>0.21</td>
</tr>
<tr>
<td>Annotator 3</td>
<td>0.91</td>
<td>0.17</td>
</tr>
<tr>
<td>Annotator 4</td>
<td>0.07</td>
<td>0.95</td>
</tr>
<tr>
<td>Annotator 5</td>
<td>0.50</td>
<td>0.92</td>
</tr>
<tr>
<td>Annotator 6</td>
<td>0.15</td>
<td>0.75</td>
</tr>
<tr>
<td>Annotator 7</td>
<td>0.45</td>
<td>0.50</td>
</tr>
</tbody>
</table>

- Given: a partially observed attribute-specific label matrix
- Want to recover its latent factors via matrix factorization
- Gives a representation of each user
Approach: Discovering Shades

• Cluster users in the space of latent factor representations
• Use $K$-means; select $K$ automatically
• Gives a representation of each shade for this attribute
Approach: Using Shades to Predict Attributes

Vote on labels

"open"

"not open"

Crowd

Adapt

Vote on labels

"open"

"not open"

Adapt

Vote on labels

"open"

"not open"

Adapt

Vote on labels

"open"

"not open"
## Results: Accuracy of Attribute Prediction

<table>
<thead>
<tr>
<th>Attribute</th>
<th>SHADES</th>
<th>GENERIC</th>
<th>User-exc</th>
<th>User-adp</th>
<th>Attr disc</th>
<th>Img clust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointy</td>
<td>76.3 (0.3)</td>
<td>74.0 (0.4)</td>
<td>67.8 (0.2)</td>
<td>74.8 (0.3)</td>
<td>74.5 (0.4)</td>
<td>74.3 (0.4)</td>
</tr>
<tr>
<td>Open</td>
<td>74.6 (0.4)</td>
<td>66.5 (0.5)</td>
<td>65.8 (0.2)</td>
<td>71.6 (0.3)</td>
<td>68.5 (0.4)</td>
<td>68.3 (0.4)</td>
</tr>
<tr>
<td>Ornate</td>
<td>62.8 (0.7)</td>
<td>56.4 (1.1)</td>
<td>59.6 (0.5)</td>
<td>61.1 (0.6)</td>
<td>58.3 (0.8)</td>
<td>58.6 (0.7)</td>
</tr>
<tr>
<td>Comfort.</td>
<td>77.3 (0.6)</td>
<td>75.0 (0.7)</td>
<td>68.7 (0.5)</td>
<td>75.5 (0.6)</td>
<td>76.0 (0.7)</td>
<td>75.4 (0.6)</td>
</tr>
<tr>
<td>Formal</td>
<td>78.8 (0.5)</td>
<td>76.2 (0.7)</td>
<td>69.6 (0.4)</td>
<td>77.1 (0.4)</td>
<td>77.4 (0.6)</td>
<td>77.0 (0.6)</td>
</tr>
<tr>
<td>Brown</td>
<td>70.9 (1.0)</td>
<td>69.5 (1.2)</td>
<td>61.9 (0.5)</td>
<td>68.5 (0.9)</td>
<td>69.3 (1.2)</td>
<td>69.8 (1.2)</td>
</tr>
<tr>
<td>Fashion.</td>
<td>62.2 (0.9)</td>
<td>58.5 (1.4)</td>
<td>60.5 (1.3)</td>
<td>62.0 (1.4)</td>
<td>61.2 (1.4)</td>
<td>61.5 (1.1)</td>
</tr>
<tr>
<td>Cluttered</td>
<td>64.5 (0.3)</td>
<td>60.5 (0.5)</td>
<td>58.8 (0.2)</td>
<td>63.1 (0.4)</td>
<td>60.4 (0.7)</td>
<td>60.8 (0.7)</td>
</tr>
<tr>
<td>Soothing</td>
<td>62.5 (0.4)</td>
<td>61.0 (0.5)</td>
<td>55.2 (0.2)</td>
<td>61.5 (0.4)</td>
<td>61.1 (0.4)</td>
<td>61.0 (0.5)</td>
</tr>
<tr>
<td>Open area</td>
<td>64.6 (0.6)</td>
<td>62.9 (1.0)</td>
<td>57.9 (0.4)</td>
<td>63.5 (0.5)</td>
<td>63.5 (0.8)</td>
<td>62.8 (0.9)</td>
</tr>
<tr>
<td>Modern</td>
<td>57.3 (0.8)</td>
<td>51.2 (0.9)</td>
<td>56.2 (0.7)</td>
<td>56.2 (1.1)</td>
<td>52.5 (0.9)</td>
<td>52.0 (1.1)</td>
</tr>
<tr>
<td>Rustic</td>
<td>67.4 (0.6)</td>
<td>66.7 (0.5)</td>
<td>63.4 (0.5)</td>
<td>67.0 (0.5)</td>
<td>67.2 (0.5)</td>
<td>67.2 (0.5)</td>
</tr>
</tbody>
</table>

**User-adp:** [Kovashka and Grauman ICCV 2013]

**Attr disc:** [Rastegari et al. ECCV 2012]

- Shades make personalization more robust
- Shades’ advantage transfers to more accurate search
<table>
<thead>
<tr>
<th>Image</th>
<th>Attribute</th>
<th>Present?</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Boot 1" /></td>
<td>Ornate</td>
<td>No</td>
<td>&quot;Ornate means decorated with extra items not inherent in the making of the object. This boot has a <strong>camo print</strong> as part of the object, but <strong>no additional items</strong> put on it.&quot;</td>
</tr>
<tr>
<td><img src="image2.png" alt="Boot 2" /></td>
<td>Ornate</td>
<td>Yes</td>
<td>&quot;The <strong>flowerprint pattern is unorthodox</strong> for a rubber boot and really <strong>stands out</strong> against the jet black background.&quot;</td>
</tr>
</tbody>
</table>
### Collected Labels and Explanations

<table>
<thead>
<tr>
<th>Image</th>
<th>Attribute</th>
<th>Present?</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td>Ornate</td>
<td>No</td>
<td>&quot;Ornate means decorated with extra items not inherent in the making of the object. This boot has a <strong>camo print</strong> as part of the object, but <strong>no additional items</strong> put on it.&quot;</td>
</tr>
<tr>
<td><img src="image2.jpg" alt="Image" /></td>
<td>Ornate</td>
<td>Yes</td>
<td>&quot;The <strong>flowerprint pattern is unorthodox</strong> for a rubber boot and really <strong>stands out</strong> against the jet black background.&quot;</td>
</tr>
<tr>
<td><img src="image3.jpg" alt="Image" /></td>
<td>Open area</td>
<td>Yes</td>
<td>&quot;This is an <strong>enclosed area, but the room is very large</strong> and the ceiling is very high, giving a lot of room. I think that this makes it an enclosed area that is also an open area. &quot;</td>
</tr>
<tr>
<td><img src="image4.jpg" alt="Image" /></td>
<td>Open area</td>
<td>No</td>
<td>&quot;I do not consider the image to show an open area because the area shown is enclosed by walls. It is a larger space on the interior of the building so it does have some aspects of an open space.&quot;</td>
</tr>
</tbody>
</table>
## Collected Labels and Explanations

<table>
<thead>
<tr>
<th>Image</th>
<th>Attribute</th>
<th>Present?</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>Ornate</td>
<td>No</td>
<td>&quot;Ornate means decorated with extra items not inherent in the making of the object. This boot has a <strong>camo print</strong> as part of the object, but <strong>no additional items</strong> put on it.&quot;</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Ornate</td>
<td>Yes</td>
<td>&quot;The <strong>flowerprint pattern</strong> is unorthodox for a rubber boot and really <strong>stands out</strong> against the jet black background.&quot;</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Open area</td>
<td>Yes</td>
<td>&quot;This is an <strong>enclosed area, but the room is very large</strong> and the ceiling is very high, giving a lot of room. I think that this makes it an enclosed area that is also an open area. &quot;</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /></td>
<td>Open area</td>
<td>No</td>
<td>&quot;I do not consider the image to show an open area because the <strong>area shown is enclosed by walls</strong>. It is a larger space on the interior of the building so it does have some aspects of an open space.&quot;</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td>Comfortable</td>
<td>Yes</td>
<td>&quot;The <strong>heel is shorter and looks more sturdy</strong> with the thickness of the heel which would make it more comfortable then your typical heel.&quot;</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td>Formal</td>
<td>Yes</td>
<td>&quot;I believe the <strong>formal aspect of this should is the color and design</strong> of the fabrics on this shoe. I felt this shoe would be used by a person who wanted to be formal yet comfortable.&quot;</td>
</tr>
</tbody>
</table>
Results: Discovered Shades

COMFORTABLE
Results: Discovered Shades

OPEN AREA

outside

view
Results: Discovered Shades

SOOTHING
Adapting Attributes by Selecting Features Similar across Domains

Liu and Kovashka, WACV 2016
Adapting Attributes by Selecting Features Similar across Domains

- Idea: Learn model for new domain by adapting a model for an existing domain
- Idea: Rely on features that are similar between the two domains to make the adaptation possible

Liu and Kovashka, WACV 2016
A naïve adaptation method fails to outperform a method that learns from scratch, but feature selection makes our method most useful when data is scarce.