Interactive Image Search with Attributes

Adriana Kovashka
Department of Computer Science
January 13, 2015

Joint work with Kristen Grauman and Devi Parikh
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
Main Idea: Interactive Visual Search

**Problem:** User’s ability to communicate about visual content is severely limited

**Key idea:** New form of interaction using precise language-based feedback

“It was bigger and brighter and more beautiful...”
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

“Like this... but with curlier hair”

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

Kovashka, Parikh, and Grauman, CVPR 2012
Prior Work: Semantic Visual Attributes

- High-level descriptive properties shared by objects
- Human-understandable and machine-detectable
- Middle ground between user and system

Prior Work: Attributes for Recognition and Search

❖ Object recognition
  • Low-dimensional image representation for classification
  • Need not be semantic


❖ Image search
  • A form of keyword search

We need ability to compare images by attribute “strength”
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 : \{ (\ldots, \) \}, \ldots \} \]

\[ E_m : \{ (\ldots, \) \}, \ldots \} \]

Ordered pairs

Similar pairs

Parikh and Grauman, ICCV 2011
We need ability to compare images by attribute “strength”
WhittleSearch with Relative Attribute Feedback

Results Page 1

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.”

More open than

Selected feedback
More bright in color than
Less high at the heel than
Less ornaments than

More bright in color than
More formal than

Round 1

Round 2

Round 3

Match

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
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** converges on target faster than traditional binary feedback
Key Questions

1) How to open up communication?

2) How to account for ambiguity of search terms?
• 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

Imprecision of Attributes

Context

Is \textit{formal}?  

= \textit{formal} wear for a \textit{conference}? OR  

= \textit{formal} wear for a \textit{wedding}?
Imprecision of Attributes

Cultural

Is blue or green?

English: “blue”

Russian: “neither”
(“голубой” vs. “синий”)

Japanese: “both”
(“青” = blue and green)
Idea: Learn User-Specific Attributes

- 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 $w'_b$, learn adapted model $w_b$,

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

subject to $y_i x_i^T w_b \geq 1 - \xi_i$, $\xi_i \geq 0$, $\forall i$

J. Yang et al. ICDM 2007.
Learning Adapted Attributes

"formal"

"not formal"

Adapted boundary

Generic boundary
Learning Adapted Attributes

“formal”

User-exclusive boundary

Adapted boundary

Generic boundary

“not formal”
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

"white shiny heels" "shinier than"

Match rate

Shoes-Binary
- generic
- user-exclusive

SUN
- generic+
- user-adaptive
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”
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

<table>
<thead>
<tr>
<th>Annotator 1</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>
Results: Discovered Shades

COMFORTABLE

- Arch
- Sole
- Soft
- Cushion
Results: Discovered Shades

OPEN AREA

outsid

view

plenti

look

high

mane

structur

wall
Conclusion

• New language-based interaction using comparisons boosts search accuracy
• Informative and perceptually precise feedback
• Exploits interplay between vision and language