How do people perceive image similarity?

Krzysztof Michalak
Presentation plan

• The experiment
• Data
• Measures
• Results
• People vs. algorithms
• Conclusion
The experiment - overview

• **Goal:**
  - Understand how people perceive image similarity
  - Improve algorithms

• **Approach:**
  - Prepare image sets (city, impression, meadow, nature, zoo)
  - Present pairs of images
  - Suggest 5 grades
  - Collect users’ votes
The experiment - tool
Data

- Raw grades: 0, 0.25, 0.5, 0.75, 1.0
- Aggregation:
  - per user
  - per image pair
- Statistical measures
- Inter-rater agreement
Results - typical cases

**Nature** image set
Grades distribution
[30, 0, 0, 0, 0]

**Zoo** image set
Grades distribution
[0, 0, 0, 4, 26]

**City** image set
Grades distribution
[6, 6, 6, 6, 6]
Results - varying level of agreement

<table>
<thead>
<tr>
<th>Kappa $\kappa$</th>
<th>Interpretation (agreement level)</th>
<th>city</th>
<th>impression</th>
<th>meadow</th>
<th>nature</th>
<th>zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>poor</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>0.00 – 0.20</td>
<td>slight</td>
<td>15%</td>
<td>16%</td>
<td>12%</td>
<td>31%</td>
<td>5%</td>
</tr>
<tr>
<td>0.21 – 0.40</td>
<td>fair</td>
<td>60%</td>
<td>71%</td>
<td>71%</td>
<td>55%</td>
<td>40%</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>moderate</td>
<td>25%</td>
<td>13%</td>
<td>17%</td>
<td>13%</td>
<td>49%</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>substantial</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
</tr>
<tr>
<td>0.81 – 1.00</td>
<td>almost perfect</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Number of pairs of people

Weighted Kappa for a pair of people

Wrocław University of Technology
Results - varying level of agreement

<table>
<thead>
<tr>
<th>Fleiss’ kappa $\kappa_F$</th>
<th>Interpretation (agreement level)</th>
<th>city</th>
<th>impression</th>
<th>meadow</th>
<th>nature</th>
<th>zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.40</td>
<td>poor</td>
<td>45%</td>
<td>48%</td>
<td>27%</td>
<td>40%</td>
<td>32%</td>
</tr>
<tr>
<td>0.40 – 0.74</td>
<td>moderate</td>
<td>41%</td>
<td>37%</td>
<td>33%</td>
<td>38%</td>
<td>32%</td>
</tr>
<tr>
<td>0.75 – 1.00</td>
<td>good</td>
<td>14%</td>
<td>15%</td>
<td>40%</td>
<td>22%</td>
<td>36%</td>
</tr>
</tbody>
</table>
Results - grade vs. agreement

City image set
435 pairs
30 raters
Fleiss’ Kappa
People vs. algorithms

• What to compare:
  - base image vs. all the others
  - ranking of image similarity using K-L measure
  - mean grade from Similararis

• How to compare:
  - visually
  - goodness-of-fit
People vs. algorithms - visually

- base image vs. all the others
- ranking of image similarity using K-L measure
- mean grade from Similaris

One of the best cases
Impression image set

Color = Fleiss Kappa
\([R; G; B] = [K_F; 0; 1 - K_F]\)
People vs. algorithms - best $R^2$ values

- **impression**
  - $R^2 = 0.6195$

- **meadow**
  - $R^2 = 0.5311$

- **impression**
  - $R^2 = 0.5204$

- **meadow**
  - $R^2 = 0.4092$

- **nature**
  - $R^2 = 0.3350$

- **meadow**
  - $R^2 = 0.3218$

- **impression**
  - $R^2 = 0.3018$

- **impression**
  - $R^2 = 0.3013$
People vs. algorithms

• What criteria are used by people?
  - objects
  - semantics
  - general properties (colour, texture) - mainly when other criteria unavailable

• When people compare „algorithmically”?
  - mostly: impression, meadow
  - rarely: city, zoo
Conclusion - image similarity

- Significant differences among people
- The level of agreement also varies
- Clearly dissimilar images - high agreement
- Extreme case - all possible grades
- Image similarity - not as well defined as it may seem
- Further work: ask people to assess various aspects (colour, composition, topic,...)
Conclusion - algorithms

• Objects - an important factor
• How to improve:
  - detect objects?
  - learn distance measures
  - use learning algorithms (e.g. neural networks)
Thank You!

Questions?