Wrocław University of Technology

Similaris

How do people perceive image similarity?

Krzysztof Michalak

Presentation plan

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

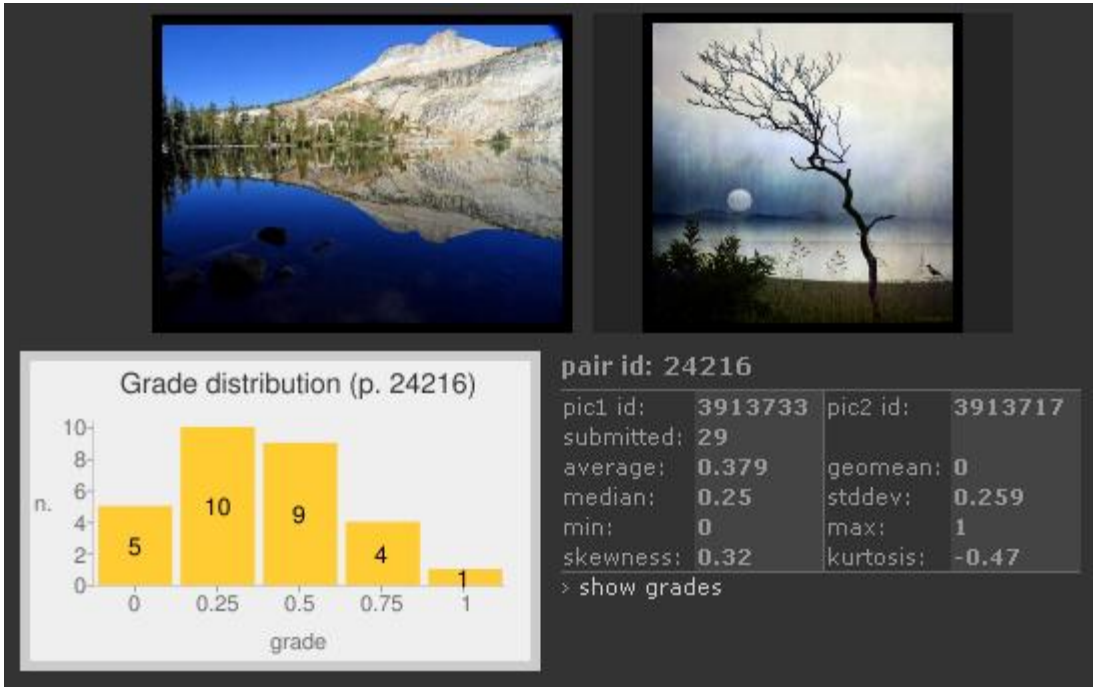
The experiment - overview

- 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

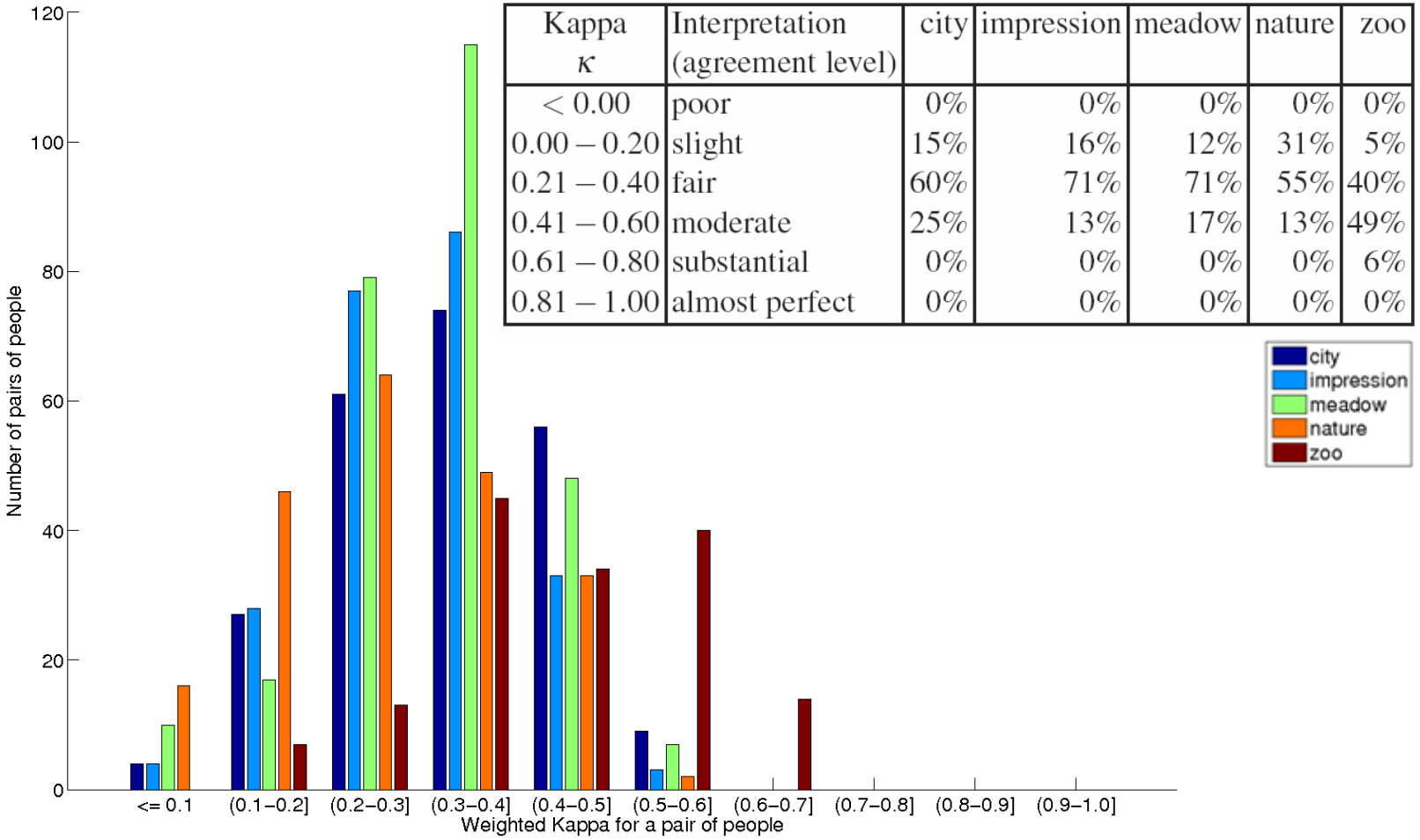
Nature image set
Grades distribution
[30, 0, 0, 0, 0]

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

City image set
Grades distribution
[6, 6, 6, 6, 6]

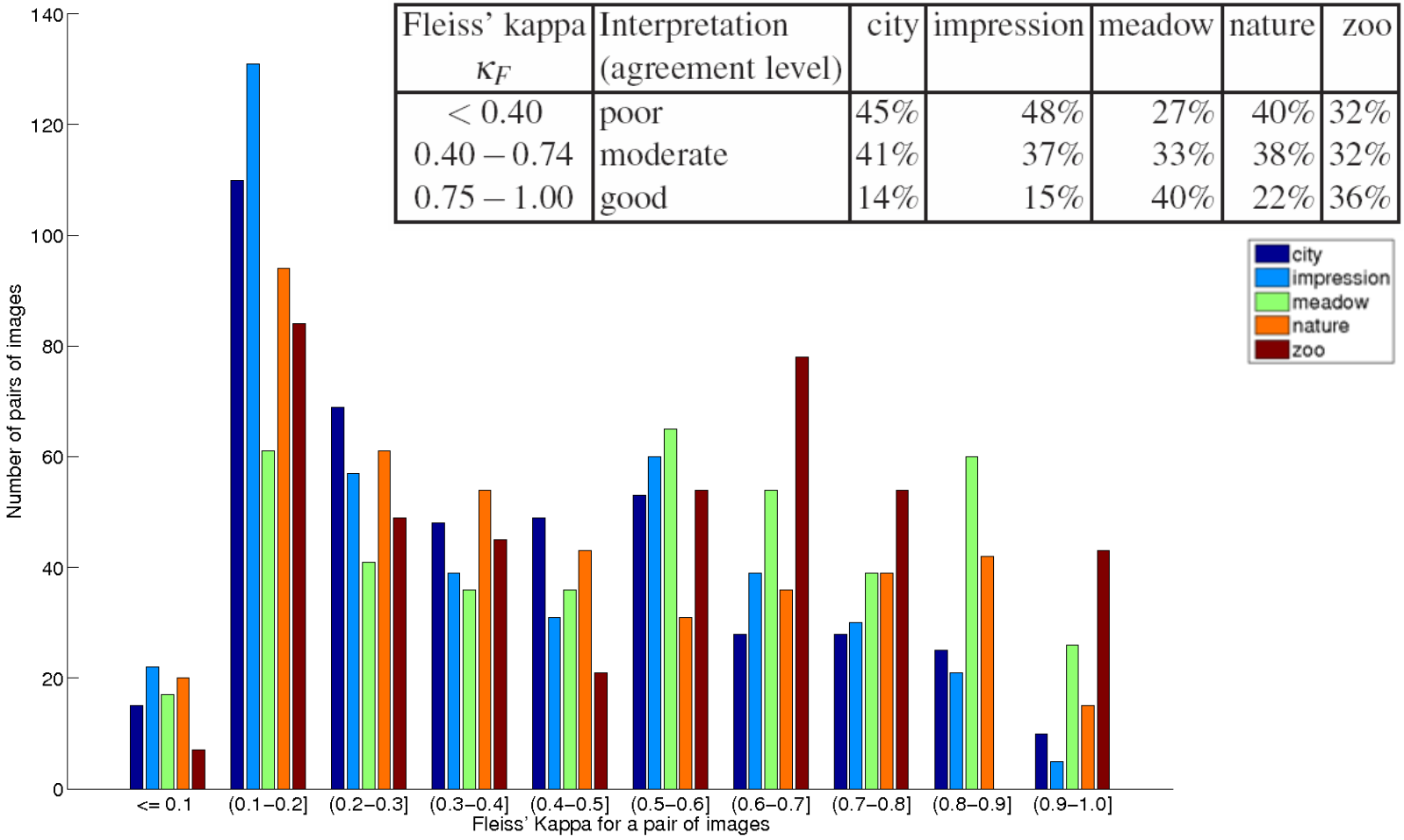
Results – varying level of agreement

Kappa κ	Interpretation (agreement level)	city	impression	meadow	nature	zoo
< 0.00	poor	0%	0%	0%	0%	0%
$0.00 - 0.20$	slight	15%	16%	12%	31%	5%
$0.21 - 0.40$	fair	60%	71%	71%	55%	40%
$0.41 - 0.60$	moderate	25%	13%	17%	13%	49%
$0.61 - 0.80$	substantial	0%	0%	0%	0%	6%
$0.81 - 1.00$	almost perfect	0%	0%	0%	0%	0%

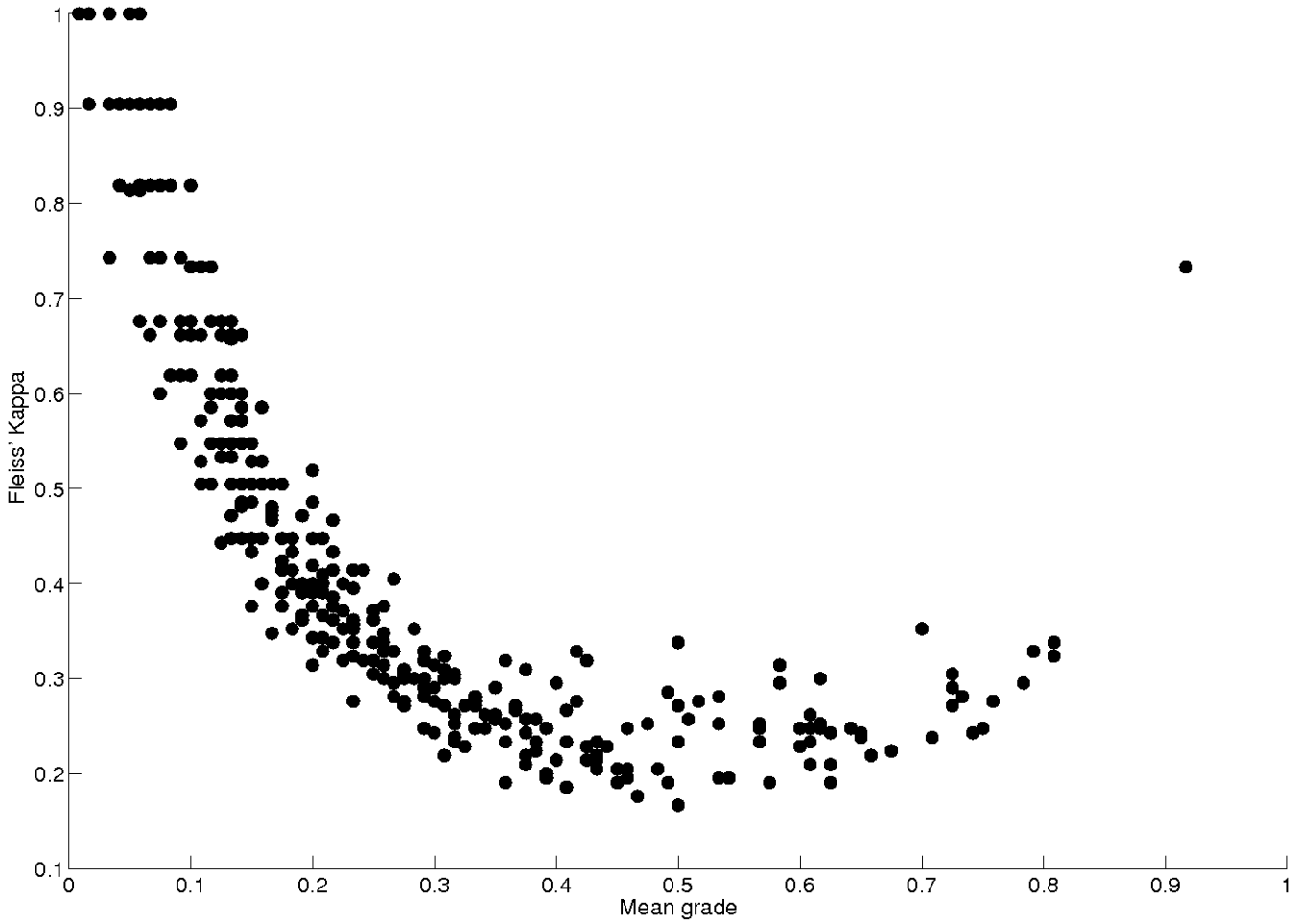


Results – varying level of agreement

Fleiss' kappa κ_F	Interpretation (agreement level)	city	impression	meadow	nature	zoo
< 0.40	poor	45%	48%	27%	40%	32%
$0.40 - 0.74$	moderate	41%	37%	33%	38%	32%
$0.75 - 1.00$	good	14%	15%	40%	22%	36%



Results – grade vs. agreement



City image set

435 pairs

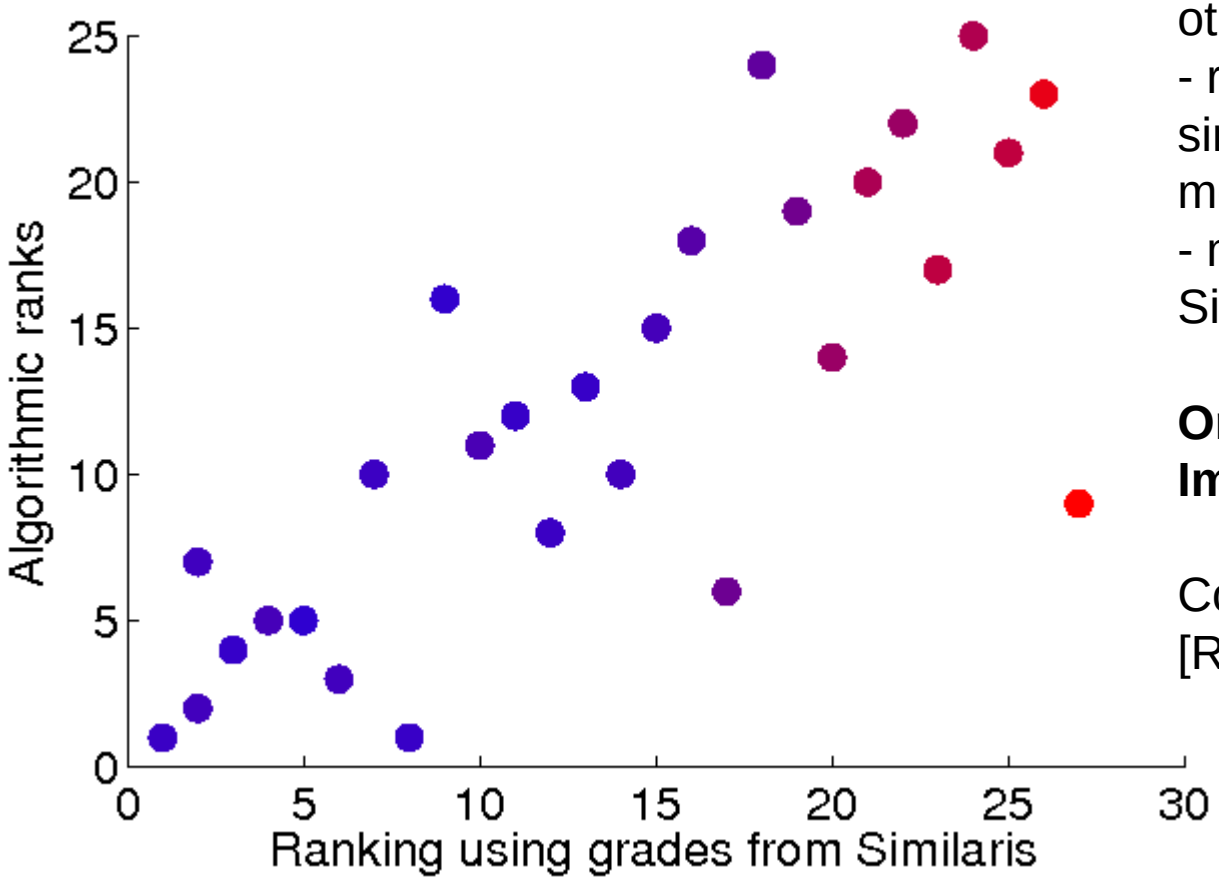
30 raters

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 Similaris
- 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 (eg. neural networks)

Thank You!

Questions?