

---

# CoPhIR Image Collection under the Microscope

---

Michal Batko  
Petra Kohoutková  
David Novák

# The CoPhIR Dataset

- **Content-based Photo Image Retrieval Dataset**
  - Images from Flickr photo sharing system
  - Over 100 million images
  - Available freely for research: <http://cophir.isti.cnr.it/>
- **Five visual descriptors per image**
  - MPEG 7 descriptors
  - Color: Scalable Color, Color Layout, Color Structure
  - Texture: Edge Histogram, Homogeneous Texture
- **Used by the MUFIN search system**



---

# CoPhIR Similarity Search Effectiveness

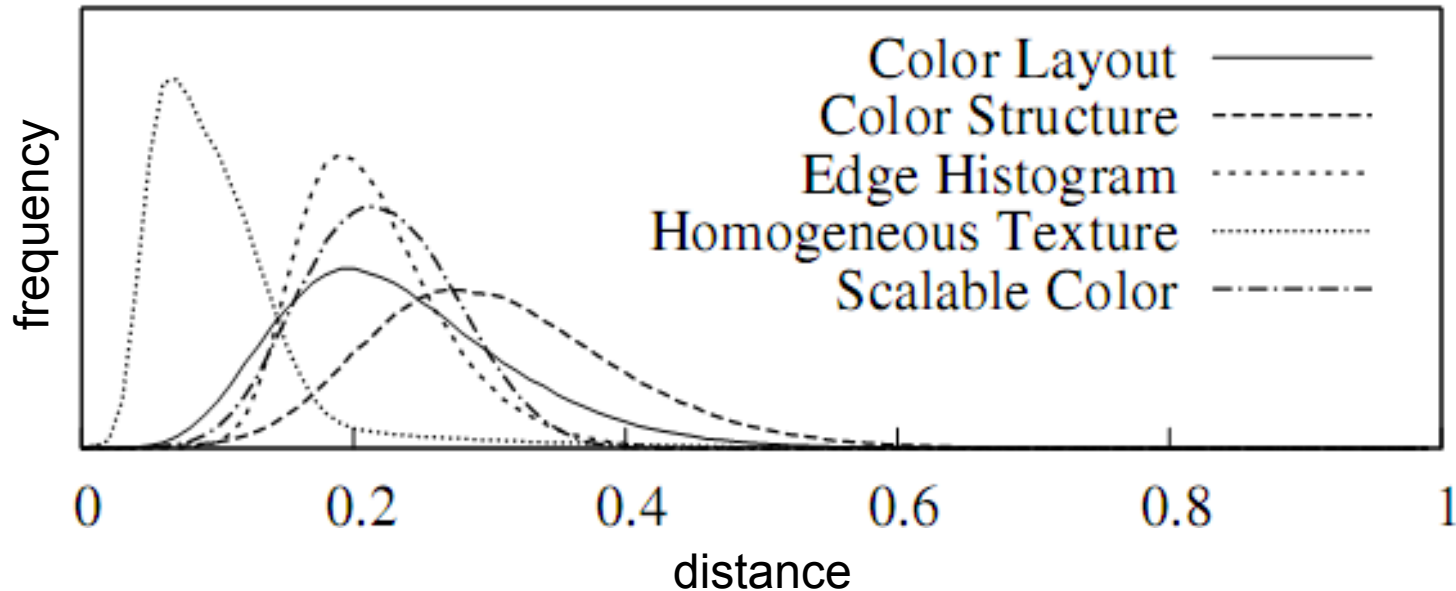
- Visual similarity
  - Five descriptors per image – how to combine?
  - Perception is subjective
  - Influenced by context
    - outdoor/indoor, people/animals/nature, etc.
- Indexing
  - Sophisticated combination is expensive
  - Complex user interface is not easy to use

---

# CoPhIR Similarity Search Effectiveness

- Insight into CoPhIR dataset
  - Metric properties
    - Distance histogram
    - Intrinsic dimensionality
    - Correlation of descriptors
  - Effectiveness of visual similarity
    - User satisfaction survey
    - Evaluation of retrieval accuracy

# Distance Histogram



- Distances of 500 000 random pairs measured
- The higher and narrower the graph, the worse
  - many objects have nearly the same distance and all of them have to be visited during the search

# Intrinsic Dimensionality

- Depending on the distance distribution, the metric space can be more or less difficult to search. This is often denoted as the **intrinsic dimensionality** of the metric space.

$$\rho = \mu^2 / (2 \cdot \sigma^2)$$

- $\mu, \sigma^2$  ... mean and variance of the distance histogram
- Low dimensionality = easy to search

<b>MPEG-7 descriptor</b>	<b>Intrinsic dimensionality</b>
Color Structure	5.116
Color Layout	3.576
Edge Histogram	7.507
Homogeneous Texture	1.323
Scalable Color	7.144

# Descriptors Correlation

Descriptor	Color Layout	Color Structure	Edge Histogram	Hom. Texture	Scalable Color
Color Layout	1.00	<b>0.23</b>	0.10	0.06	<b>0.45</b>
Color Structure	<b>0.23</b>	1.00	<b>0.24</b>	0.09	<b>0.67</b>
Edge Histogram	0.10	<b>0.24</b>	1.00	<b>0.23</b>	0.18
Hom. Texture	0.06	0.09	<b>0.23</b>	1.00	0.09
Scalable Color	<b>0.45</b>	<b>0.67</b>	0.18	0.09	1.00

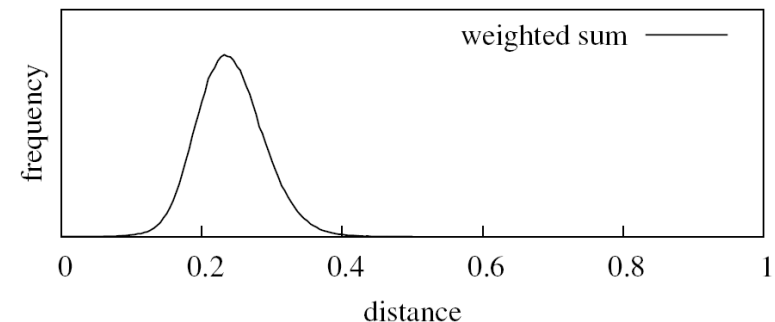
- As expected, color descriptors are correlated with each other more than with the texture descriptors and vice versa.
- The most correlated is Scalable Color
  - the first candidate for removal to save space or processing time

# Combining Descriptors

- Overall similarity
  - Weighted sum of all 5 descriptors
  - Weights provided by CoPhIR authors

MPEG-7 descriptor	Weight
Scalable Color	2.5
Color Structure	2.5
Color Layout	1.5
Edge Histogram	4.5
Homogeneous Texture	0.5

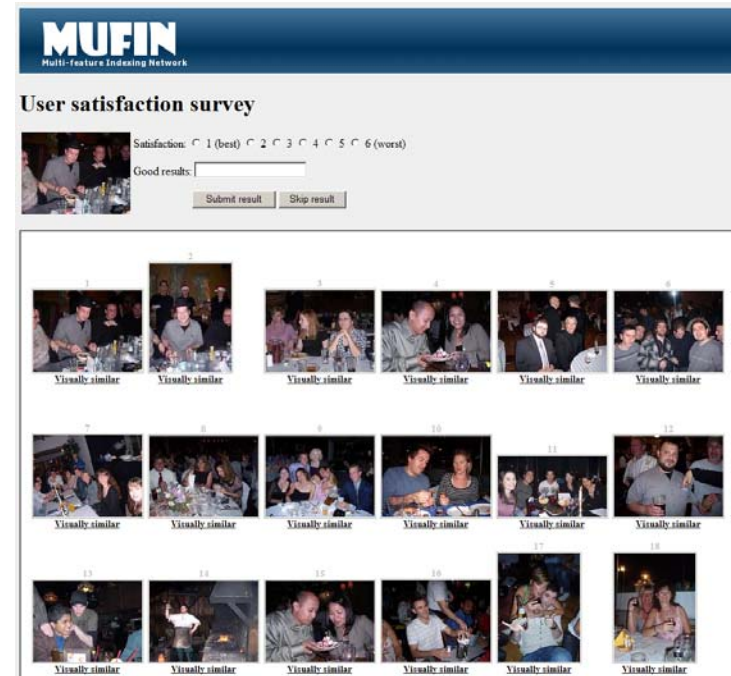
- Distance histogram
  - Intrinsic dimensionality = 12.9
  - Difficult to index





# User Satisfaction Survey

- Testing search effectiveness
  - Given set of 50 query images
  - Users rated the search result
    - Grades from 1 (best) to 6 (worst)
    - Selected best images (optional)
- Satisfaction Results
  - 10 million and 100 million images
  - MUFIN search engine



Dataset	Satisfaction	Queries	Users
10M	2.63	446	31
100M	2.17	401	28

# Evaluation of Retrieval Accuracy

- Average Normalized Modified Retrieval Rate (ANMRR)
  - Recommended for MPEG-7 visual descriptors
  - Based on *ground truth*
    - Images that are “correct” answer for a given query
  - Measures the position of ground-truth images in the *k*-NN results given by an index structure

$$\text{rank}(i) = \begin{cases} \text{position}(i), & \text{position}(i) \leq k \\ 1.25 \cdot k, & \text{otherwise.} \end{cases}$$

# Evaluation of Retrieval Accuracy Results

- **Modified ANMRR**
  - Definition of ground truth for millions of images
  - Fixed  $k = 30$ , ground truth from similarity result
- **Two datasets**
  - Average ANMMR from 50 query images
  - Effect of bigger dataset

<b>Dataset</b>	<b>ANMRR</b>	<b># improved</b>	<b># worsen</b>
10M	0.49	-	-
100M	0.41	62 %	28 %

# Logarithm Descriptor Combination

## ■ Motivation

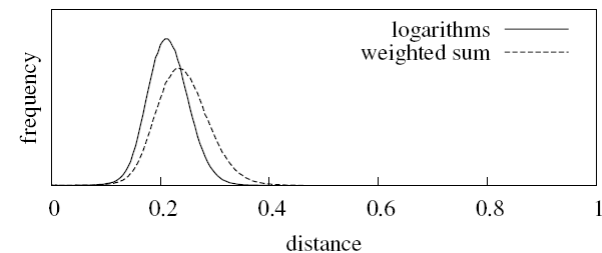
- Some user-preferred images received higher distances because of just one descriptor

## ■ Logarithms – smaller big-distance difference

- $2.5 \cdot \ln(SC) + 2.5 \cdot \ln(CS) + 1.5 \cdot \ln(CL) + 4.5 \cdot \ln(EH) + 0.5 \cdot \ln(HT)$
- Worsened intrinsic dimensionality: 16.2

## ■ Accuracy results

Dataset	ANMRR	# improved	# worsen
10M	0.49	-	-
logarithms	0.43	55 %	25 %



---

# Image Categorization

- Different approach to improve effectiveness
  - Define a set of categories
  - Tune aggregation function for the category
- Query Categorization
  - top-level categories from the Corel photo collection + additional category for drawings
- Aggregation function
  - weighted sum, weights tuned using half of the queries from given category

# Image Categorization (cont.)

- Tuned for 6 most frequent categories

category	SC	CS	CL	EH	HT
buildings	2.5	2.5	1.5	9	0.5
landscapes	2.5	2.5	3.5	4.5	0.5
parts	2.5	2.5	3.5	6	0.5
persons	2.5	2.5	0	6	0
vehicles	2.5	2.5	1.5	9	0
drawings	2.5	2.5	3	6	0

- Accuracy results

Dataset	ANMRR	# improved	# worsen
10M	0.49	-	-
categorized	0.43	67.5 %	10 %

---

# Conclusion

- Examined properties of individual descriptors
  - Metric space properties
  - Relationships between descriptor values
- User satisfaction survey
- ANMRR measurement
  
- Two ways of effectiveness improvement
  - Logarithms in the aggregation
  - Special metric functions for image categories