Metric Query Processing

Petra Kohoutkova
pkoh@ics.muni.cz
Masaryk University, Brno, Czech Republic

Michal Batko
batko@fi.muni.cz
Masaryk University, Brno, Czech Republic

Pavel Zezula
zezula@fi.muni.cz
Masaryk University, Brno, Czech Republic

Similarity Searching in Metric Space

Similarity Paradigm

- content-based similarity search is suitable for any type of data (multimedia, DNA sequences, etc.)
- near objects are required rather than exact matches
- query-by-example: the query is defined by an object and
  - maximum distance of qualifying objects, or
  - number of nearest neigbors to be returned

Metric-Space Model

- treat the dataset $\mathcal{D}$ as unstructured objects together with a function $d : \mathcal{D} \times \mathcal{D} \longrightarrow \mathbb{R}$, which measures objects' proximity
- *metric space* is a pair $\mathcal{M} = (\mathcal{D}, d)$ and the *distance* function d must satisfy the following postulates $\forall x, y, z \in \mathcal{D}$:

$$\begin{array}{ll} d(x,y) \geq 0 \wedge (d(x,y) = 0 \Leftrightarrow x = y) & \textit{(non-negativity)}, \\ d(x,y) = d(y,x) & \textit{(symmetry)}, \\ d(x,z) \leq d(x,y) + d(y,z) & \textit{(triangle inequality)}. \end{array}$$

- the lower the distance the more similar the objects are

MUFIN Image Search

- prototype similarity search system MUFIN (developed at Masaryk University)
- distributed peer-to-peer architecture
- 100,000,000 images from the CoPhIR Test Collection[1]
- similarity of images measured by five MPEG-7 features:
  - Scalable Color (SC),
  - Color Structure (CS),
  - Color Layout (CL),
  - Edge Histogram (EH),
  - Homogeneous Texture (HT)
- the features are combined into a single metric space using the following aggregation function:

$$dist = 2 \cdot SC + 3 \cdot CS + 2 \cdot CL + 4 \cdot EH + 0.5 \cdot HT$$

- the weights have been determined experimentally
- focus on the *nearest neighbors query* $kNN(q, k)$, which returns k indexed objects with the smallest distances to q

kNN Search: Response Example

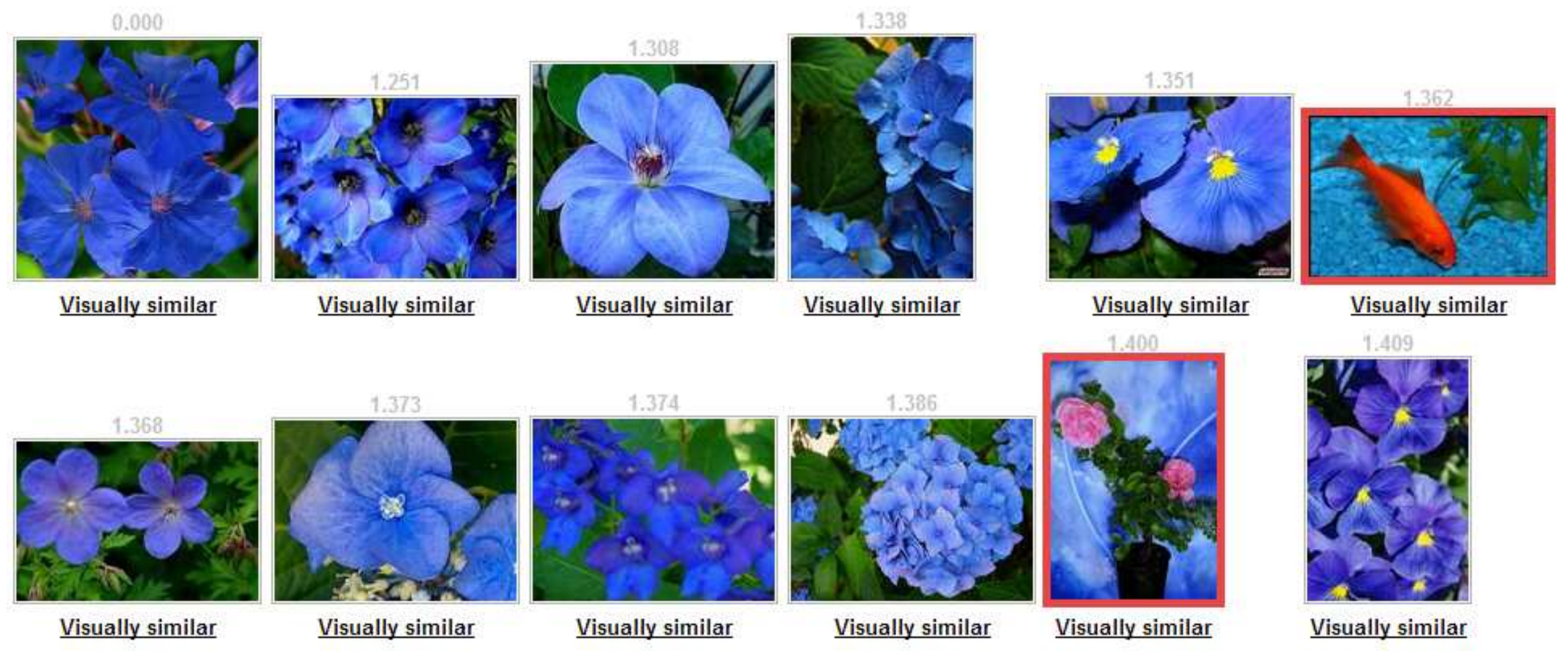

Figure 1 Image search screenshot.

Challenge

- the notion of similarity is subjective, cannot be exactly described by a single function
- what if we do not want the marked objects in the result?
- more complex query definition and execution strategies needed

More Sophisticated Query Evaluation Strategies

Dynamic Aggregation

- instead of having one combined metric space, use the five metric spaces defined by MPEG-7 descriptors
- the weights for overall distance computation are defined separately for each query
- the five metric spaces are searched using Threshold algorithm
- advantages: the search settings can be adjusted to user's preferences
- disadvantages: slow evaluation

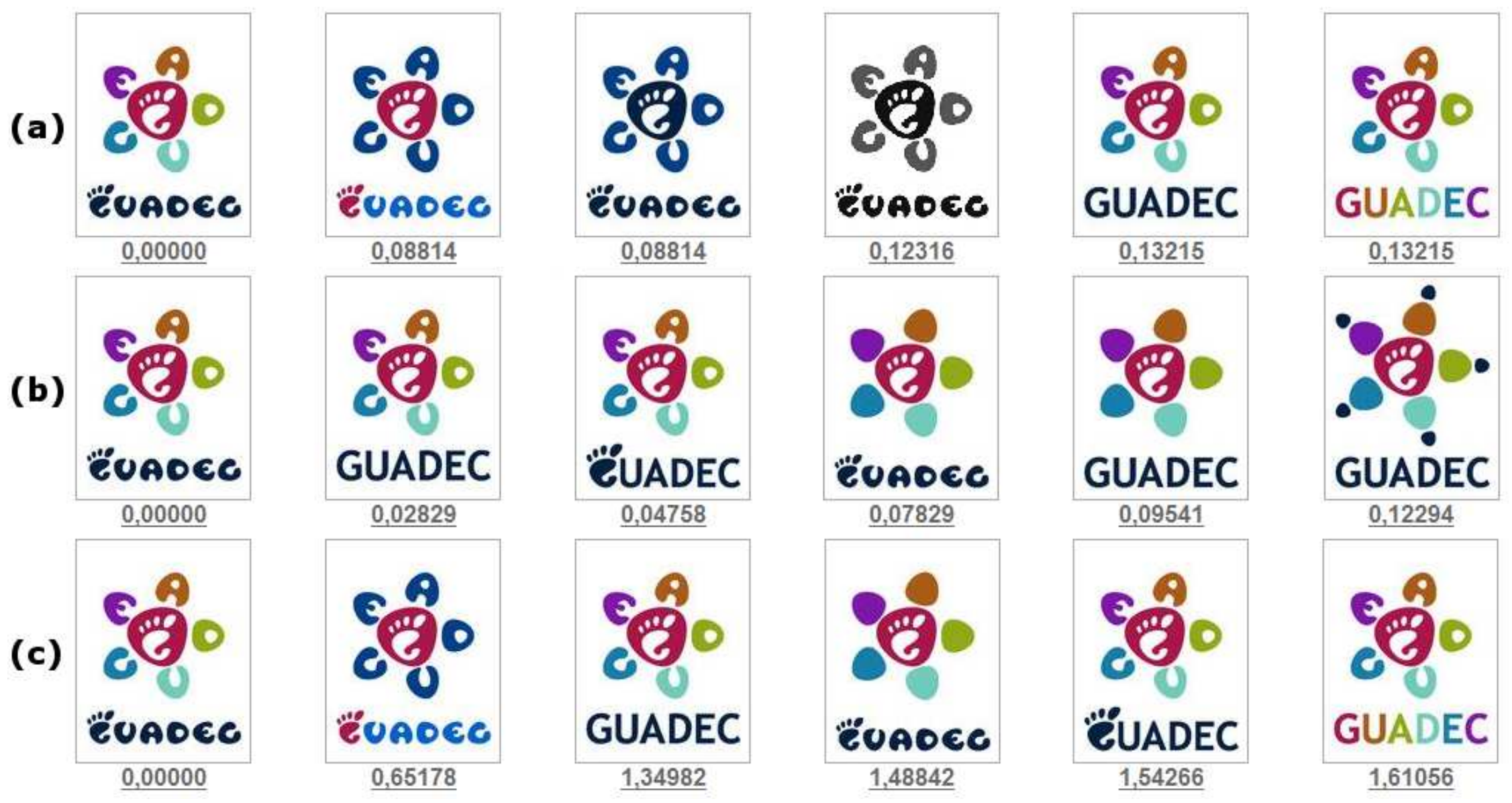

Figure 2 Search results for different aggregation settings.

In (a), zero weights for the three color overlays are used, thus focusing on image texture and edges only. On the other hand, just colors are queried in (b). The basic aggregation function is used in (c).

Multi-Object Query

- more than one object may be used to define the query
- execution similar to dynamic aggregation, uses Threshold algorithm to combine the results acquired for individual query objects
- disadvantages: slow evaluation

Filtering

- executed on individual peers during the query execution
- only objects satisfying the filtering condition may be added to the response
- possible filtering conditions:
  - keywords in image annotation
  - limits on individual feature distances
  - limit on overall distance
- advantages: flexible, quick (filtering executed on multiple peers in parallel), less data sent through the network

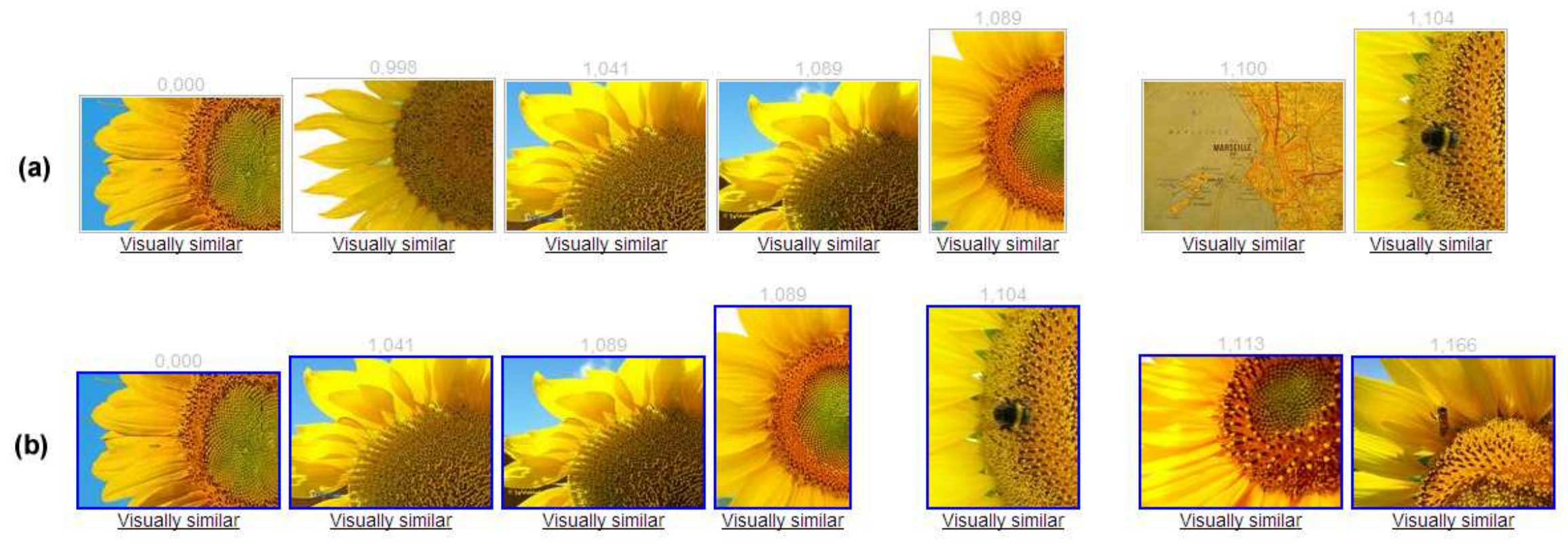

Figure 3 Filtering example.

In (a), no filtering is used. In (b), only images with Color Structure distance lower than 0.8 may qualify for the result.

Postprocessing

- enables user to work with search results
- possible postprocessing operations:
  - reorder result set ignoring given descriptor
  - reorder result set ignoring discriptor with lowest/highest distance
  - reorder result set with respect to object distances to a selected object (other than query object)
- advantages: can work in iterations, improving the result
- disadvantages: only objects found in previous iterations are considered

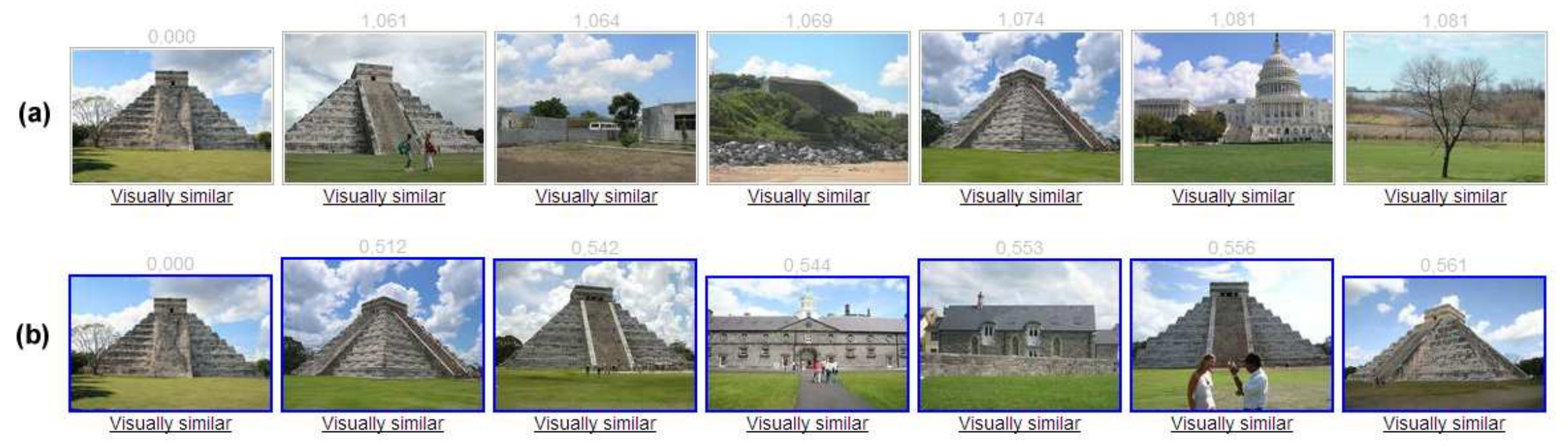

Figure 4 Postprocessing example.

In (a), the result of basic search is displayed. In (b), the result has been reordered ignoring descriptor with the highest distance.

Future Work

- propose and evaluate user satisfaction experiments for given search strategies
- propose a query language for similarity searching that would allow formulation of the similarity queries in a standardized uniform way and support advanced query types and features – combined queries with user-defined aggregation function, time/precision preferences, multi-object queries, advanced filtering and postprocessing

References

- P. Zezula, G. Amato, V. Dohnal and M. Batko. *Similarity Search: The Metric Space Approach*. Springer-Verlag, 2006.
- David Novak, Michal Batko and Pavel Zezula. *Web-scale System for Image Similarity Search: When the Dreams Are Coming True* in Proceedings of CBMI 2008, London, UK.
- Michal Batko, Petra Kohoutkova and Pavel Zezula. *Combining Metric Features in Large Collections.* In Proceedings of SISAP 2008, Cancun, Mexico.

MUFIN Project Web Page: http://mufin.fi.muni.cz/

[1] The Content-based Photo Image Retrieval (CoPhIR) Test-Collection (http://cophir.isti.cnr.it/). The images were crawled from Flickr http://www.flickr.com/ and the features extracted by ISTI-CNR, Pisa, Italy.