The Future of Multimedia Databases and Multimedia Data Mining

Thomas Seidl
RWTH Aachen University, Germany

MMKM 2008
Milton Keynes, Feb. 14th, 2008

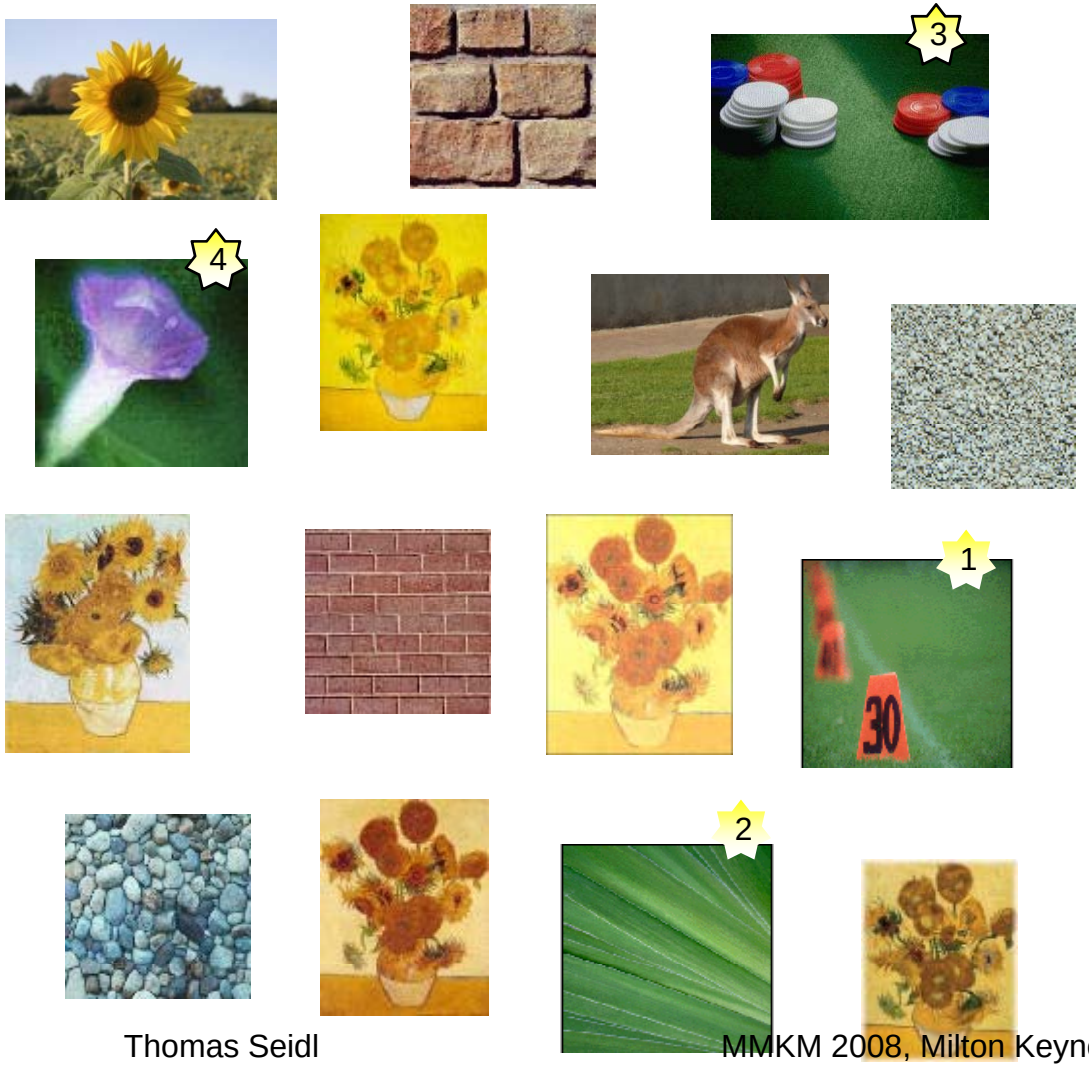
Overview

- Content-based similarity search
  - Complex models: quadratic forms, Earth Movers' distance
  - Efficient algorithms: approximations and indexing
- New interaction models (change of use)
  - Incremental search, relevance feedback, anytime querying
- From retrieval to new data mining tasks
  - Subspace clustering, outlier detection, stream data mining

Overview

- Content-based similarity search
  - Complex models: quadratic forms, Earth Movers' distance
  - Efficient algorithms: approximations and indexing
- New interaction models (change of use)
  - Incremental search, relevance feedback, anytime querying
- From retrieval to new data mining tasks
  - Subspace clustering, outlier detection, stream data mining

Content-based Similarity Search



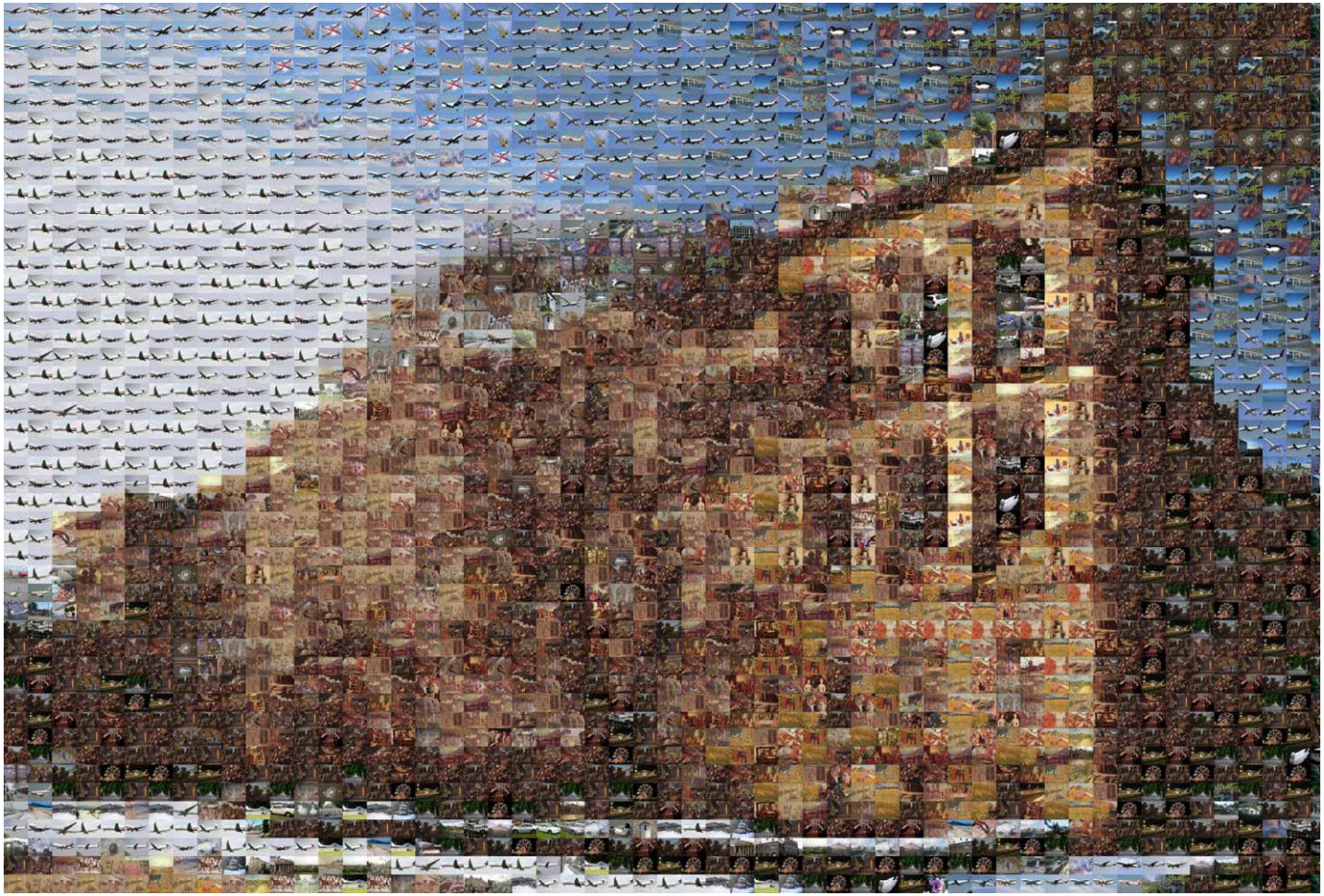
Retrieval task: Which images in the archive are similar to the example?

Justification for hits:
Similar color frequencies

Example: Mosaic Poster

Bild: Hendrik Brixius

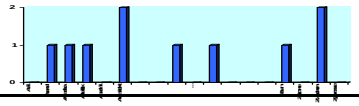
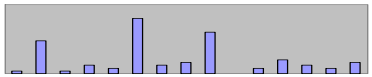
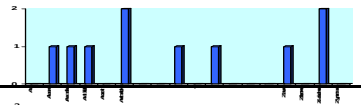
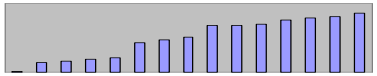
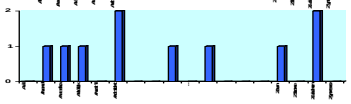
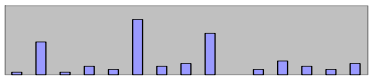
Result: Mosaic made from 2.500 Images



Result: Mosaic made from 2.500 Images

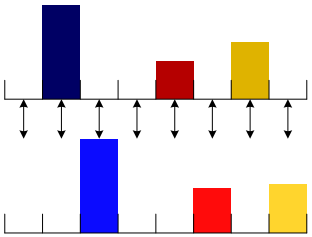
Description of Image Contents

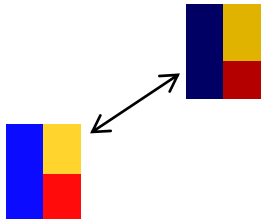
- Categories: *colors, textures, shapes,* etc.
- Description by histograms

Image	Color histogram	Texture histogram	...
#123673			...
#543643			...
#363273			...

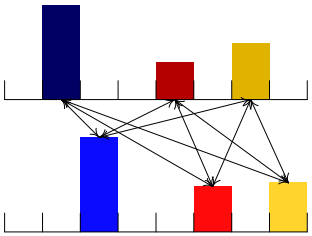
- Specification of similarity by formal models

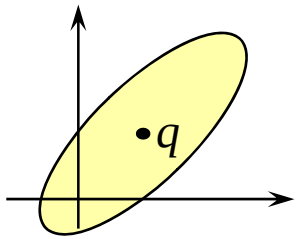
Adaptable Similarity Models





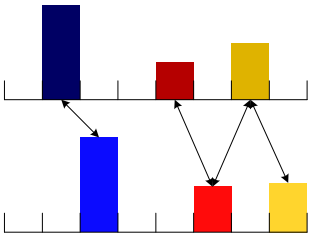
- Euclidean Distance
  - Neglects cross-bin similarities

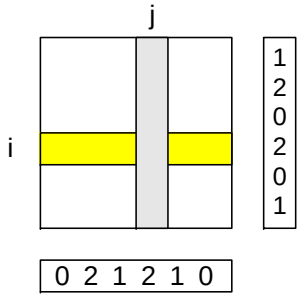




- Quadratic Forms
  - Linear Algebra

$$d_A(p,q) = \sqrt{(p-q)\cdot A\cdot (p-q)^t}$$

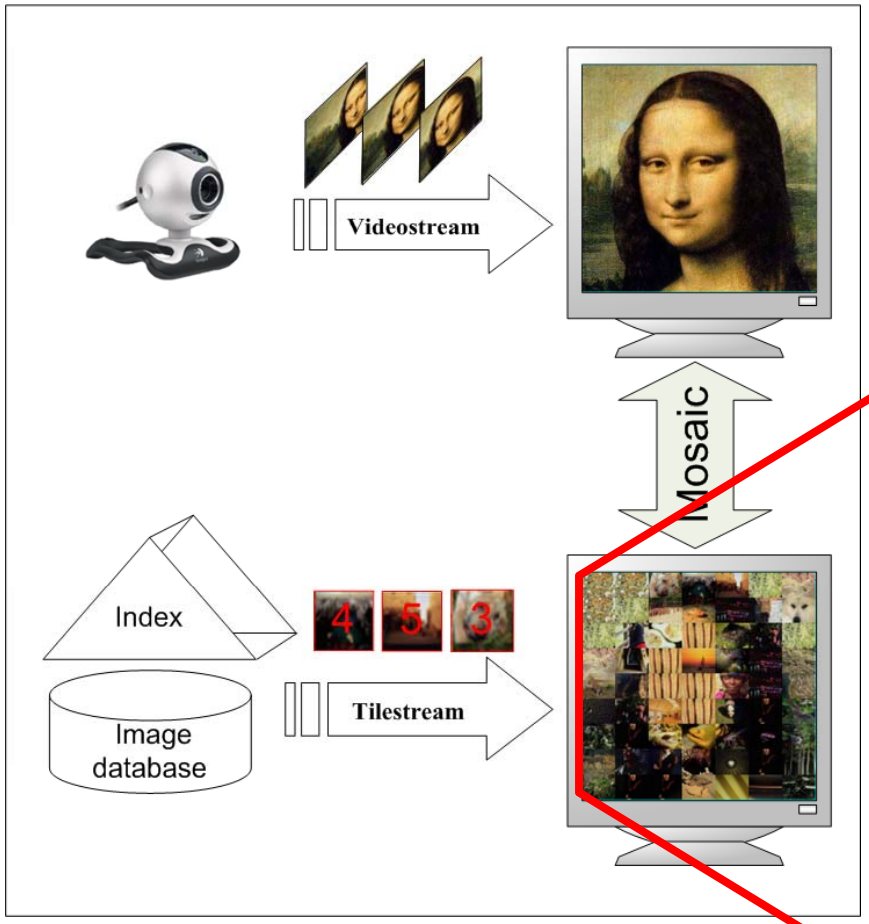




- Earth Mover's Distance
  - Linear Programming

$$EMD_C(p,q) = \min\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{c_{ij}}{m} f_{ij}, \quad \begin{array}{l} \forall i\,\forall j: f_{ij} \geq 0 \\ \forall i: \sum_{j=1}^{n} f_{ij} = p_i \\ \forall j: \sum_{i=1}^{n} f_{ij} = q_j \end{array} \right\}$$

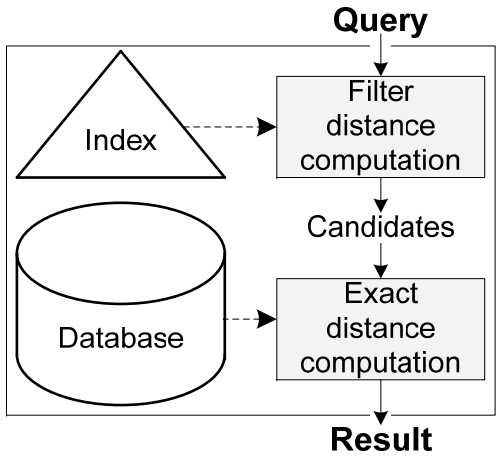
Mosaic Videos, Require Faster Retrieval …



Demo *AttentionAttractor*

[Assent, Krieger, Seidl: ICDE 2007]

Acceleration by Filter-Refine Architectures



Usage of filters

- Filter step for fast pruning
- Refinement step for exact result

[GEMINI: Faloutsos 1996; KNOP: Seidl&Kriegel 1998]

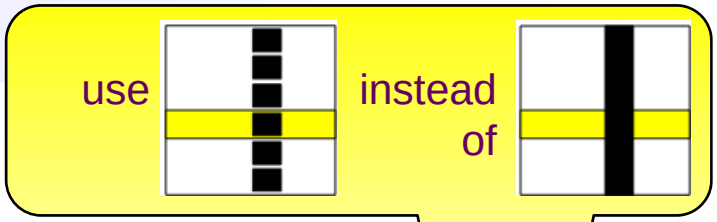
I ndex-enabled
C omplete
E fficient
S elective

Quality of filters: **ICES**

- ***I*** Filter is supported by index structures
- ***C*** Filter does not miss qualifying objects
- ***E*** Filter distance is calculated efficiently
- ***S*** Filter yields only small candidate set

[Assent, Wenning, Seidl: ICDE 2006]

A Filter for the EMD



$$EMD_C(x,y) = \min\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{c_{ij}}{m} f_{ij}, \quad \forall i\, \forall j : f_{ij} \geq 0, \ \forall i : \sum_{j=1}^{n} f_{ij} = x_i, \ \forall j : \sum_{i=1}^{n} f_{ij} = y_j \right\}$$

$\geq$ $\subseteq$

Constraint Relaxation

$$LB_{IM}(x,y) = \min\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{c_{ij}}{m} f_{ij}, \quad \forall i\, \forall j : f_{ij} \geq 0, \ \forall i : \sum_{j=1}^{n} f_{ij} = x_i, \ \forall j\, \forall i : f_{ij} \leq y_j \right\}$$

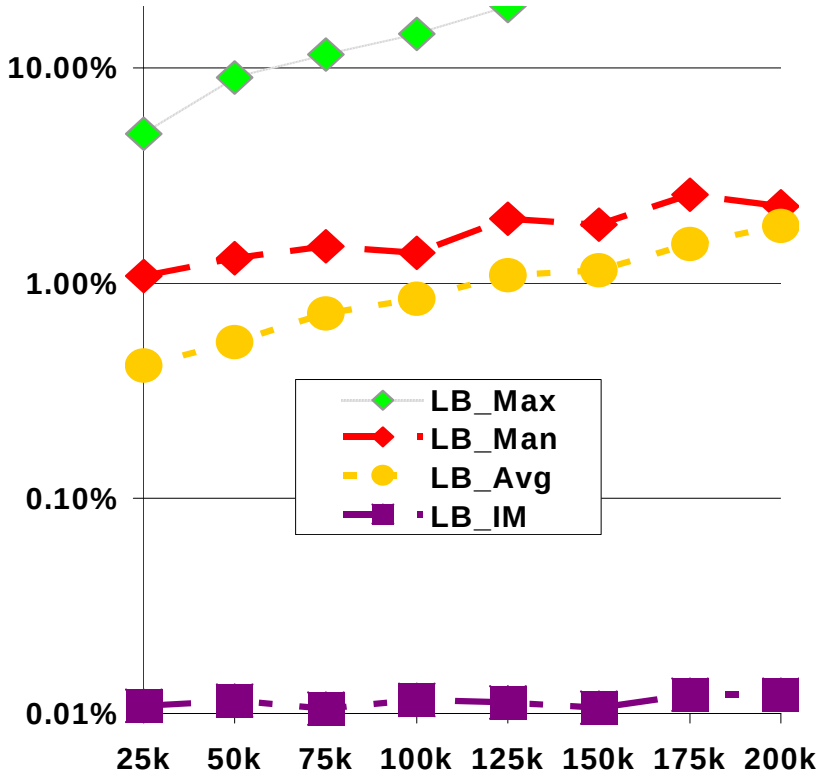
$$= \sum_{i=1}^{n} \min\left\{ \sum_{j=1}^{n} \frac{c_{ij}}{m} f_{ij}, \quad \forall j: f_{ij} \geq 0, \ \sum_{j=1}^{n} f_{ij} = x_i, \ \forall j: f_{ij} \leq y_j \right\}$$

I	+++
C	+++
E	+++
S	+++

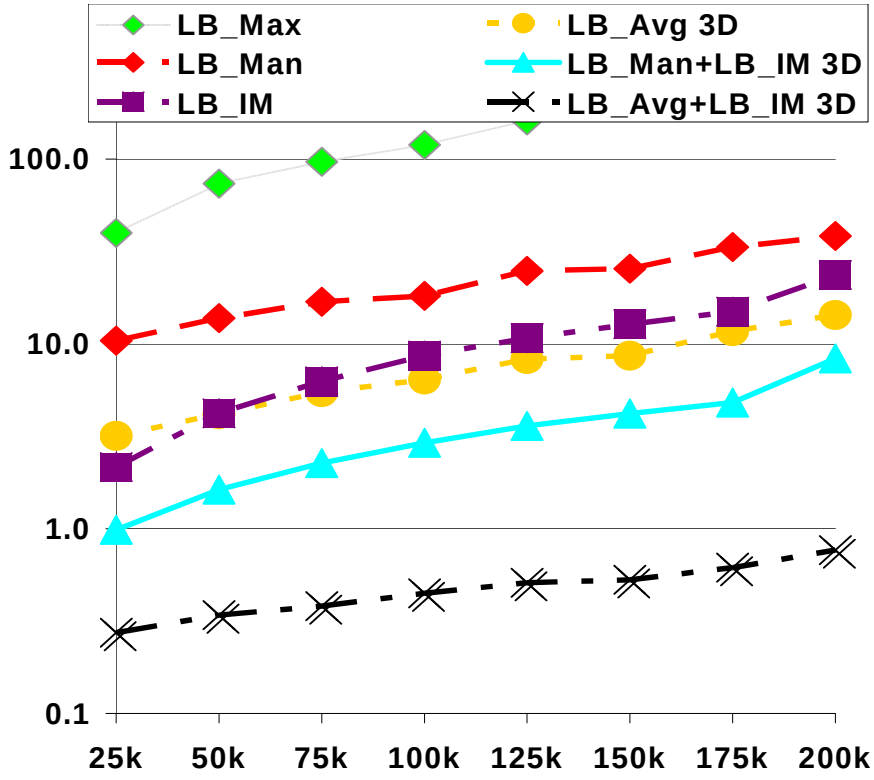
Filter: [Assent, Wenning, Seidl: ICDE 2006]
Index: [Assent, Wichterich, Meisen, Seidl: ICDE 2008]

Scalability w.r.t. Database Size

Selectivity



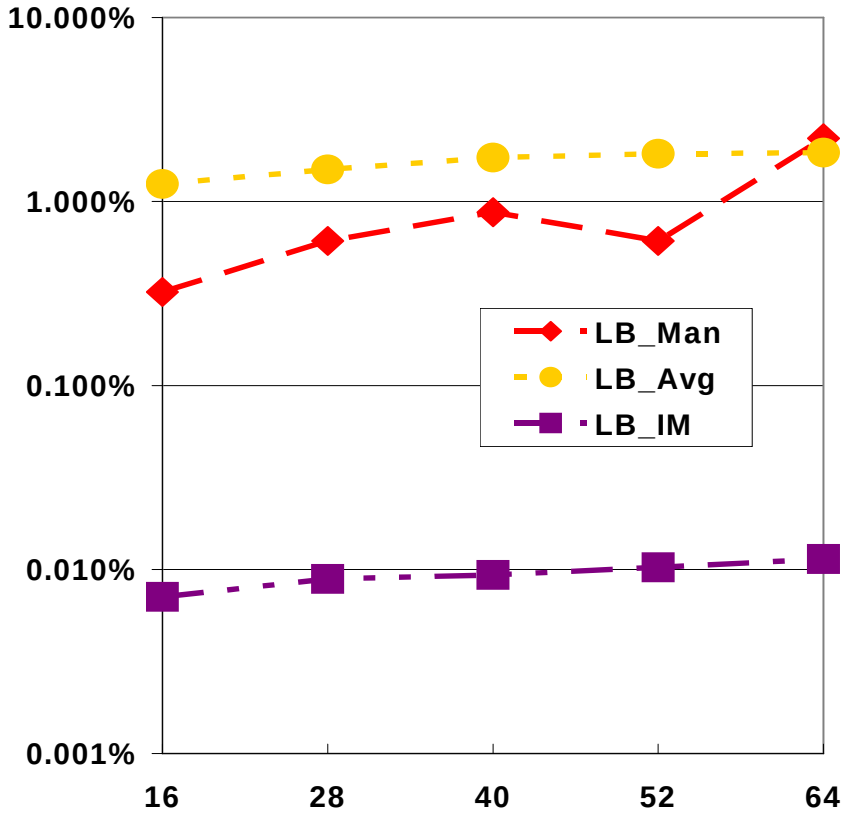
Response Time



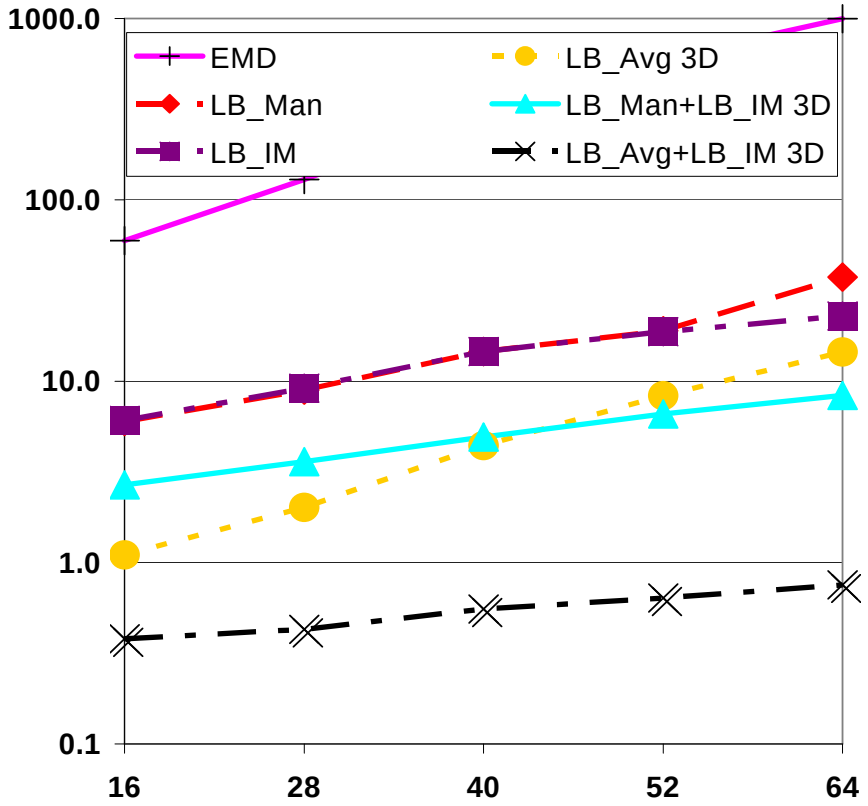
Data bases with 25,000 to 200,000 images, 64d color histograms, 10NN, log. scales

Scalability w.r.t. Dimensionality

Selectivity



Response Time



Data base with 200,000 images, color histograms of dimensionalities 16 to 64, 10NN, log. scales

Overview

- Content-based similarity search
  - Complex models: quadratic forms, Earth Movers' distance
  - Efficient algorithms: approximations and indexing
- New interaction models (change of use)
  - Incremental search, relevance feedback, anytime querying
- From retrieval to new data mining tasks
  - Subspace clustering, outlier detection, stream data mining

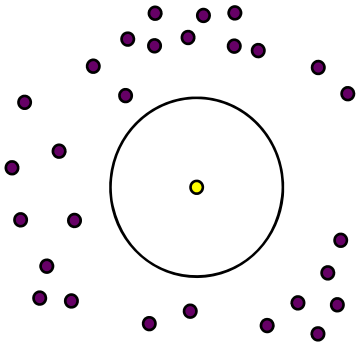
Change of Interaction Models

- Incremental Retrieval
  - From *ranges* over *nearest neighbors* to *incremental retrieval*
- Relevance Feedback
  - From *weights* over *covariances* to *ground distances*
- Anytime Search
  - From *blind waiting* over *progress estimation* to *progress monitoring*

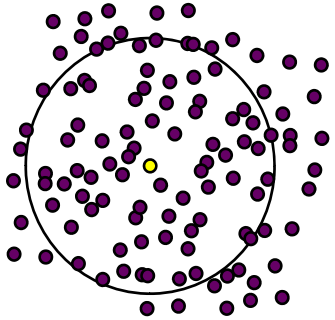
Incremental Nearest Neighbor Search

Range Queries

$$\{o \in DB \mid d(o,q) \leq \varepsilon\}$$



no results

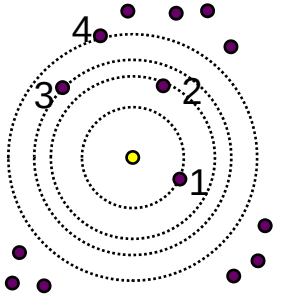


too many results

k-nn Queries

d_q - Ranking $r_q : \Im_{|DB|} \rightarrow DB$

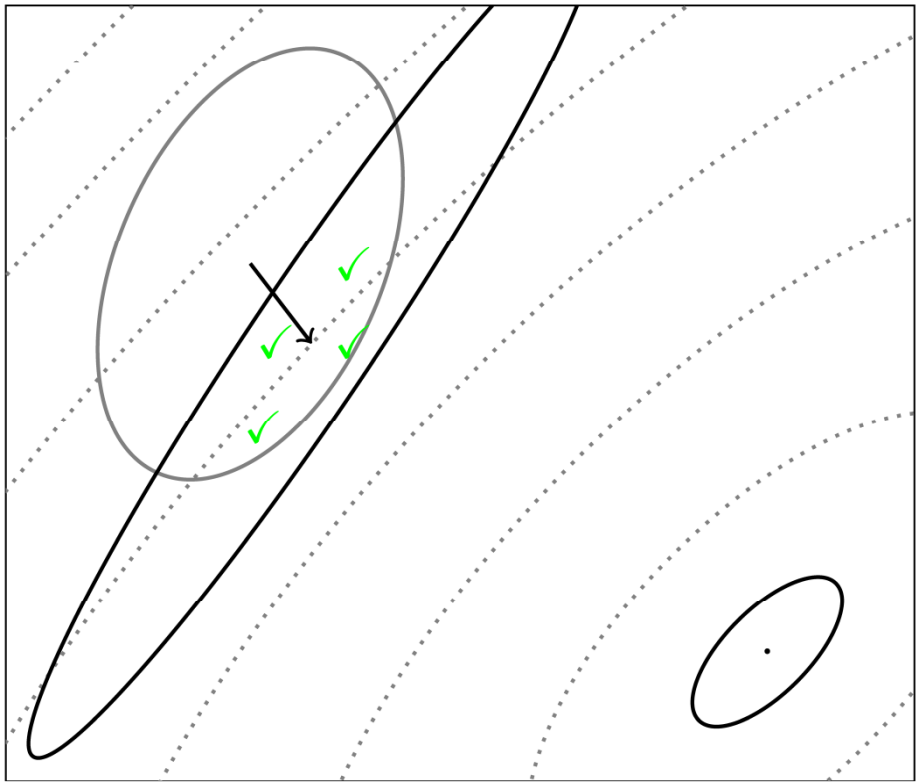
$i_1 \leq i_2 \Rightarrow d(r_q(i_1),q) \leq d(r_q(i_2),q)$



k-nearest neighbors: $r_q(\Im_k)$

Incremental Search: „Give-me-more“

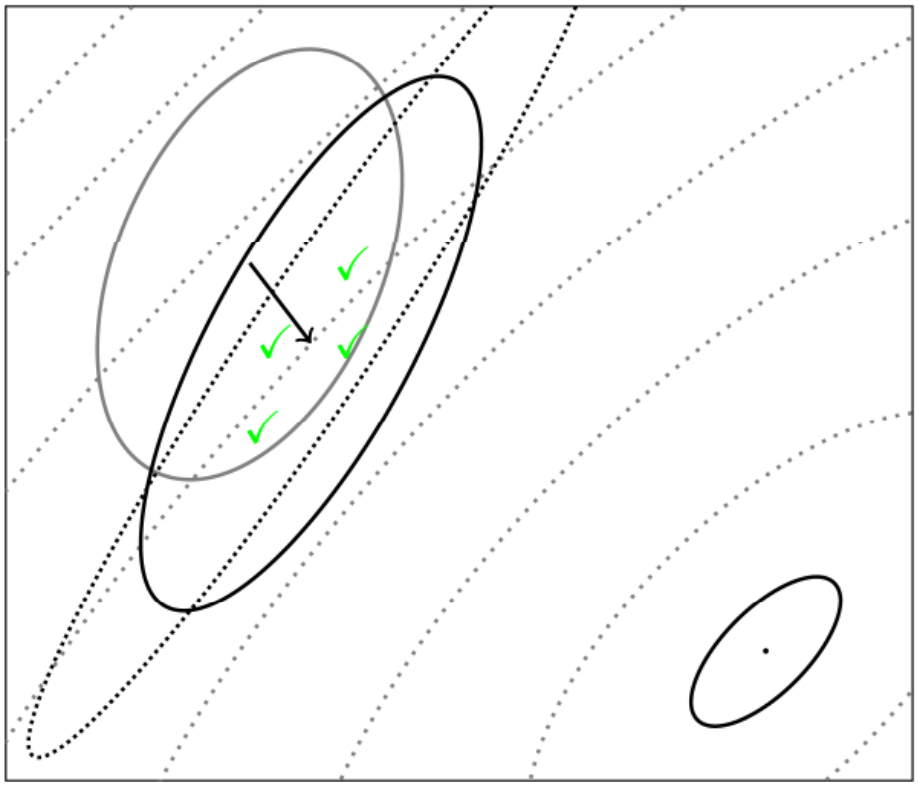
Relevance Feedback: MindReader



Regard cross-bin similarities (covariance matrix)

Y. Ishikawa, R. Subramanya, C. Faloutsos: *MindReader: Querying Databases Through Multiple Examples.* Proc. 24th Int. Conf. on Very Large Data Bases, 1998.

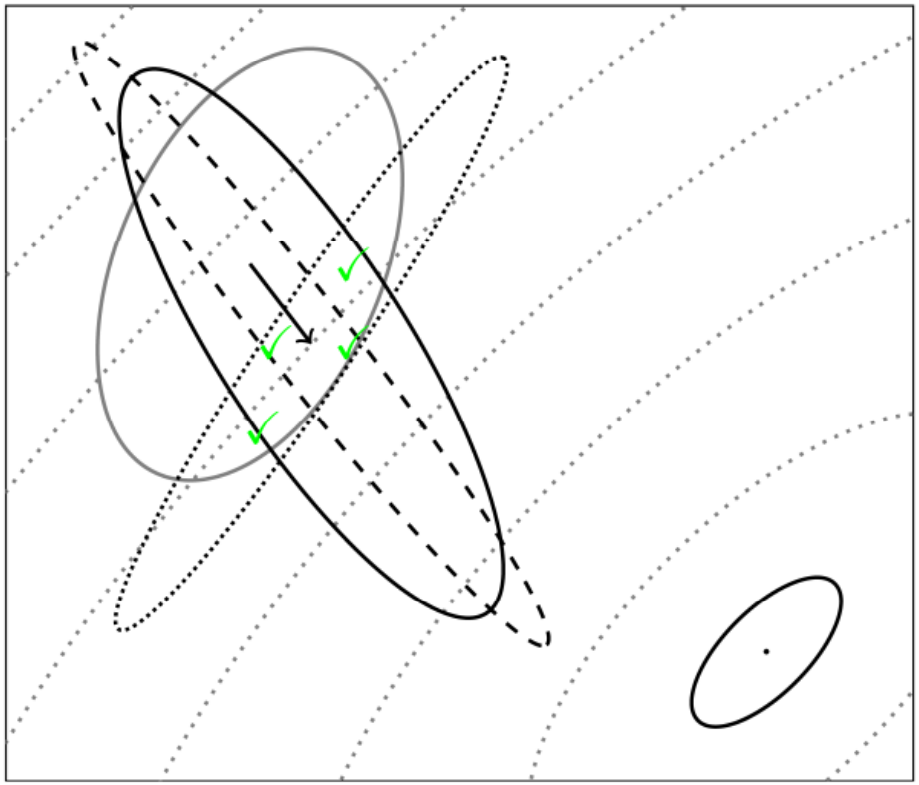
Relevance Feedback: History



Incorporate History of Feedback

- Exponential aging of former feedback influence

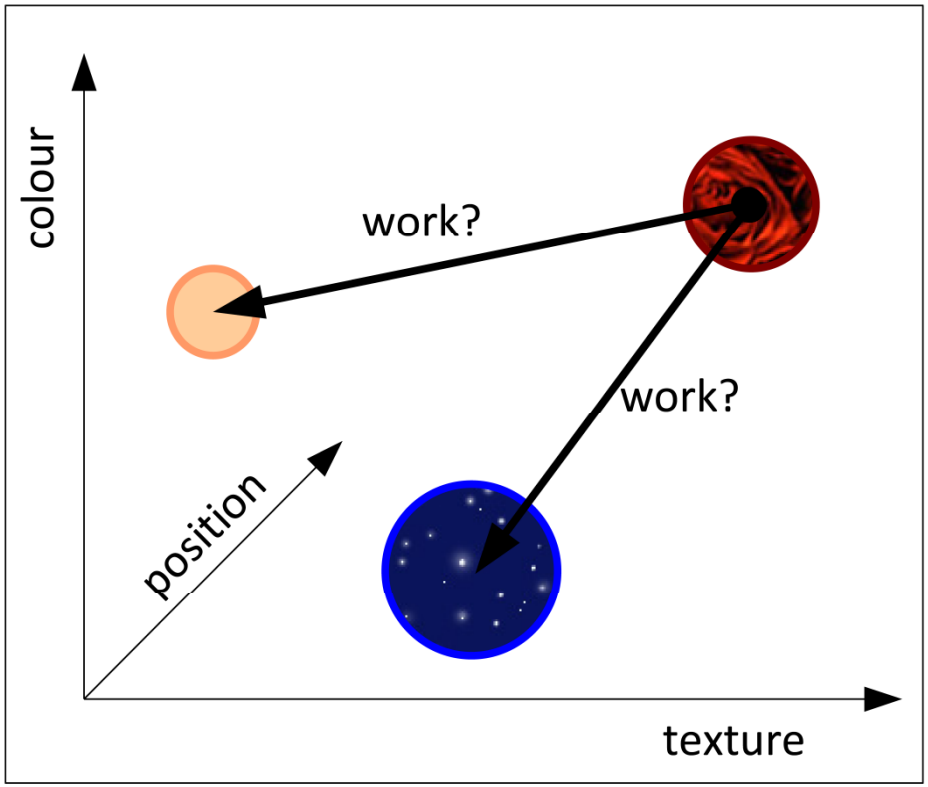
Relevance Feedback: Foresight



Two Phases

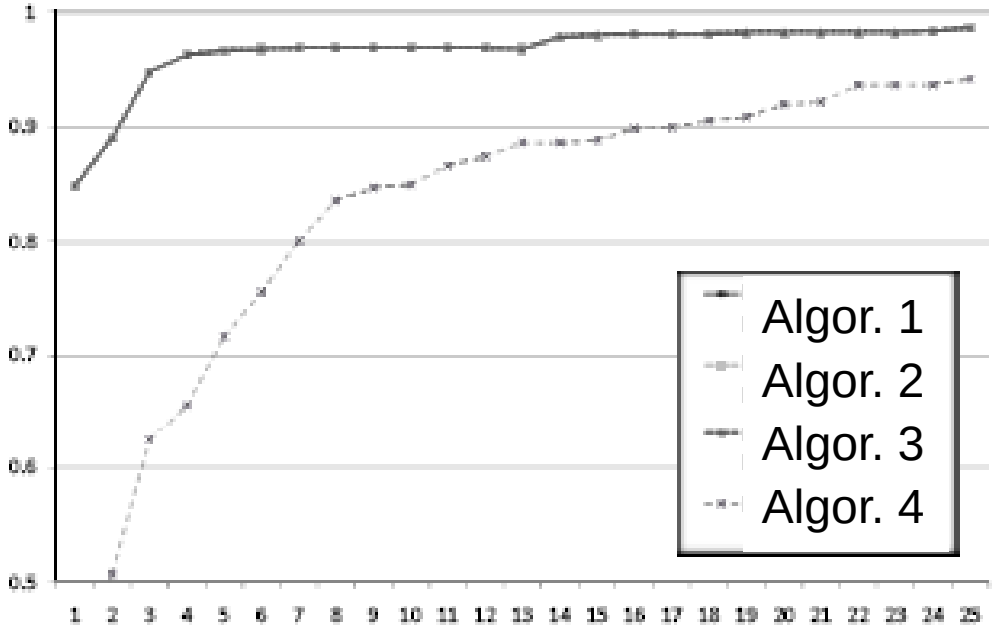
1. Approach fast to relevant area
2. Adjust similarity model at destination area

Relevance Feedback: Earth Mover's Distance



From *weights* over *covariances* (QF) to *ground distances* (EMD)

Anytime Data Mining (example: classification)



- No predefined budget but query can be interrupted at any time
- Expectation: increasing classification accuracy
  - The later the interruption, the better the classication accuracy should be

Yang, Y., G.I. Webb, K. Korb, and K-M. Ting (2007). Classifying under Computational Resource Constraints: Anytime Classification Using Probabilistic Estimators. Machine Learning 69(1). Netherlands: Springer, pp. 35-53.

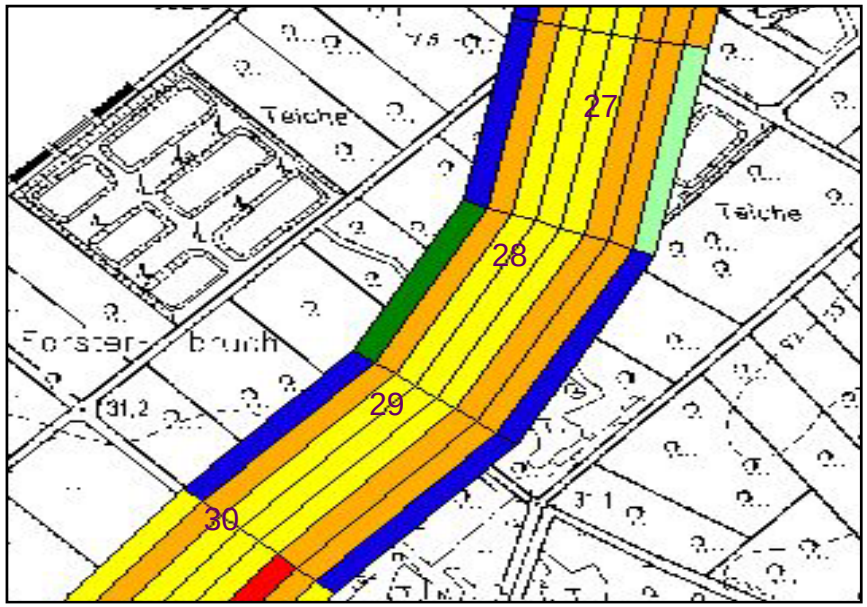
Overview

- Content-based similarity search
  - Complex models: quadratic forms, Earth Movers' distance
  - Efficient algorithms: approximations and indexing
- New interaction models (change of use)
  - Incremental search, relevance feedback, anytime querying
- From retrieval to new data mining tasks
  - Subspace clustering, outlier detection, stream data mining

From Information Retrieval to Data Mining

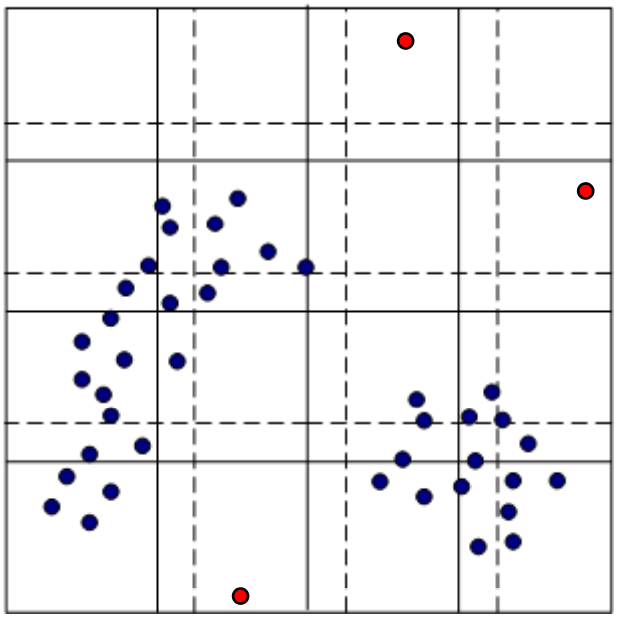
- Multimedia Information Retrieval
  - Driven by queries („query by example“)
  - Adaptable models for content-based similarity (QF, EMD)
  - Interactive usage: incremental search, relevance feedback
- Multimedia Data Mining
  - No query objects (Which to submit? Does browsing help?)
  - Reveal patterns which are hidden in vast amounts of data: regularities, irregularities
  - Typical tasks: clustering, mining association rules, aggregation / generalization of data

Subspace Clustering



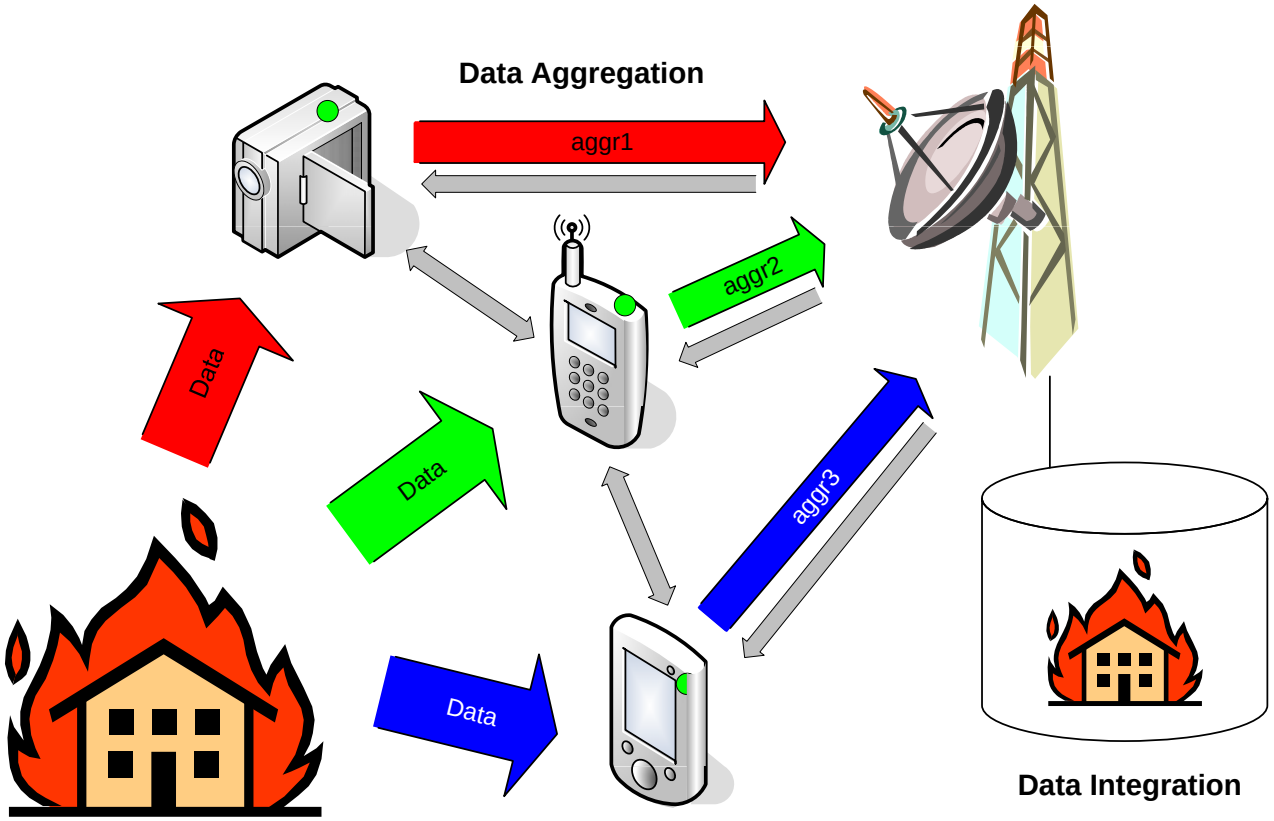
- Cluster structures are hidden by noisy dimensions
- Separate relevant and irrelevant dimensions locally
- Identify clusters and their respective subspaces

Subspace Outlier Detection



- New challenge: how to define outliers wrt. subspace clusters?
- Definition: clusters = dense areas, outliers = singularities
- Question: Outliers = *noise* … or outliers = *objects of interest* ?
- Task: Outlier detection = complementary task to clustering

Stream Aggregation and Generalization



Limitations: battery power; bandwidth; monitoring attention

Conclusion

- Content-based similarity search
  - Complex models: quadratic forms, Earth Movers' distance
  - Efficient algorithms: approximations and indexing
- New interaction models (change of use)
  - Incremental search, relevance feedback, anytime querying
- From retrieval to new data mining tasks
  - Subspace clustering, outlier detection, stream data mining

Selected Publications

- Assent I., Krieger R., Glavic B., Seidl T.: *Clustering of Multidimensional Spatial Sequences.* In: Int. Journal on Knowledge and Information Systems (**KAIS**), 2008.
- Seidl T.: *Nearest Neighbor Classification*. In: Liu L., Özsu M. T. (Eds.): *Encyclopedia of Database Systems*. Springer, 2008. (to appear)
- Wichterich M., Assent I., Kranen P., Seidl T.: *Efficient EMD-based Similarity Search in Multimedia Databases via Flexible Dimensionality Reduction.* Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Vancouver, Canada, 2008.
- Assent I., Wichterich M., Meisen T., Seidl T.: *Efficient similarity search using the Earth Mover's Distance for large multimedia databases.* Proc. **IEEE** Int. Conf. on Data Engineering (**ICDE**), Cancun, Mexico, 2008.
- Assent I., Krieger R., Afschari F., Seidl T.: *The TS-Tree: Efficient Time Series Search and Retrieval*. Proc. Int. Conf. on Extending Data Base Technology (**EDBT**), Nantes, France, 2008.
- Assent I., Krieger R., Welter P., Herbers J., Seidl T.: *SubClass: Classification of Multidimensional Noisy Data Using Subspace Clusters*. Proc. Int. Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD**), Osaka, Japan, Springer **LNCS-LNAI**, 2008.
- Wichterich M., Beecks C., Seidl T.: *Ranking Multimedia Databases via Relevance Feedback with History and Foresight Support*. Proc. 2nd Int. Workshop on Ranking in Databases (**DBRank**) in conj. with IEEE 24th Int. Conf. on Data Engineering (ICDE), Cancun, Mexico, 2008.
- Müller E., Assent I., Steinhausen U., Seidl T.: *OutRank: ranking outliers in high dimensional data*. Proc. 2nd Int. Workshop on Ranking in Databases (**DBRank**) in conjunction with IEEE 24th Int. Conf. on Data Engineering (ICDE), Cancun, Mexico, 2008.
- Zaki M.J., Peters M., Assent I., Seidl T.: *Clicks: An effective algorithm for mining subspace clusters in categorical datasets.* Data & Knowledge Engineering (**DKE**) 60 (1): 51-70, 2007.
- Assent I., Krieger R., Müller E., Seidl T.: *VISA: Visual Subspace Clustering Analysis*. In ACM **SIGKDD Explorations** Special Issue on Visual Analytics, Vol. 9(2), Dec. 2007.
- Assent I., Krieger R., Müller E., Seidl T.: *DUSC: Dimensionality Unbiased Subspace Clustering*. Proc. **IEEE** Int. Conf. on Data Mining (**ICDM**), Omaha, Nebraska, USA, 2007.
- Assent I., Krieger R., Seidl T.: *AttentionAttractor: efficient video stream similarity query processing in real time* (Demo). Proc. **IEEE** 23rd Int. Conf. on Data Engineering (**ICDE**), Istanbul, Turkey, 2007.
- Assent I., Wenning A., Seidl T.: *Approximation Techniques for Indexing the Earth Mover's Distance in Multimedia Databases.* Proc. **IEEE** Int. Conf. on Data Engineering (**ICDE**), Atlanta, Georgia, USA, 2006.
- Assent I., Wichterich M., Seidl T.: *Adaptable Distance Functions for Similarity-based Multimedia Retrieval.* In: **Datenbank-Spektrum** Nr. 19: 23-31, 2006.

Abstract

Multimedia data archives, databases, and web services grow from day to day with a high speed. In order to cope with the huge amount of data, new exploration models and scalable algorithms need to be developed. The expected changes of use also demand changes on the technology level. Starting with content-based retrieval based on complex distance functions including quadratic forms and Earth Movers' distance, interaction models such as incremental search and relevance feedback are discussed. New data mining approaches including subspace clustering, outlier mining, stream data mining, and anytime classification also will be applied to multimedia databases in the future. Following this trend, the underlying retrieval and mining technologies are highly demanded to be extended. Future developments will yield novel approximations, indexing techniques, and multi-step query processing in order to provide efficiency and scalability.

Bio of Thomas Seidl

Thomas Seidl is a full professor for Computer Science and head of the Data Management and Data Exploration group at RWTH Aachen University, Germany, where he currently advises eight PhD students.

His research interests include data mining and data management in multimedia and spatio-temporal databases for applications from computational biology, medical imaging, mechanical engineering, computer graphics, etc. whith a focus on content, shape or structure of complex objects in large databases. Current projects aim at fast content-based multimedia retrieval, relevance feedback, subspace clustering, outlier detection, stream data minining, and anytime mining algorithms. His research in the field of relational indexing aims at exploiting the robustness and high performance of relational database systems for complex indexing tasks.

Having finished his MS in 1992 at the Technische Universität München, Thomas received his Ph.D. in 1997 and his venia legendi in 2001 from the University of Munich, Germany. In 2001, he was a guest lecturer at the University of Augsburg and from 2001 to 2002, he held a substitute professorship for Databases, Data Mining, and Visualization at the University of Constance, Germany.