EFFICIENT SIMILARITY SEARCH AND ANALYSIS

Dr. Christian Beecks

Who am I?

- **Senior Researcher at Fraunhofer FIT**
- Academic career at RWTH Aachen University:
  - 2014 - 2017: Akademischer Rat
  - 2013 - 2014: Post-doctoral researcher
  - 2007 - 2013: Ph.D. in Computer Science
Thesis: *Distance-based Similarity Models for Content-based Multimedia Retrieval*

- Research: *How to access multimedia data efficiently?*
  - Adaptive similarity models
  - Efficient similarity search techniques
  - Scalable indexing and query processing algorithms

User-centered Ubiquitous Computing

Smart Cities

Industrie 4.0

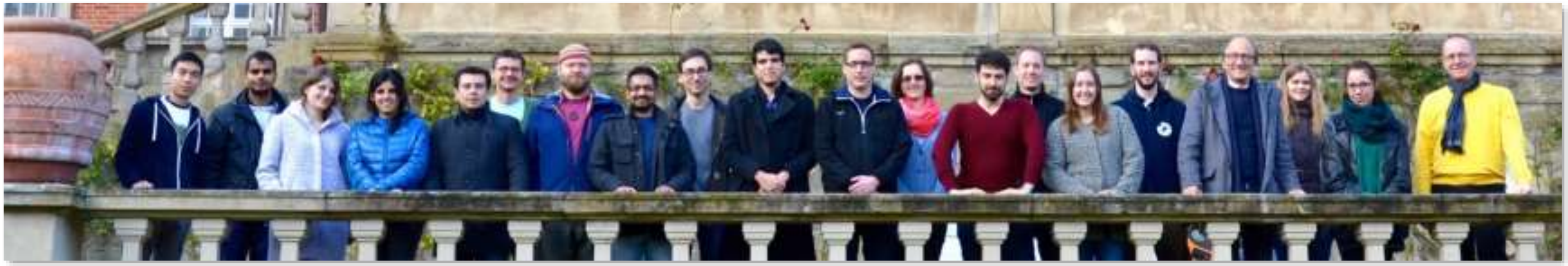
Energy Efficiency

Smart Grids

Smart Data

UCD

IoT Platforms



Fraunhofer FIT

AGENDA

1) Introduction
2) Smart Multimedia Data Representation
   - *How to model multimedia data?*
3) Adaptive Similarity Models
   - *How to compare multimedia data?*
4) Efficient Retrieval Approaches
   - *How to search multimedia data?*
5) Time for Discussion

AGENDA

1) Introduction
2) **Smart Multimedia Data Representation**
   - *How to model multimedia data?*
3) Adaptive Similarity Models
   - *How to compare multimedia data?*
4) Efficient Retrieval Approaches
   - *How to search multimedia data?*
5) Time for Discussion

Multimedia Data Never Sleeps

- **Multimedia** serves as **modern means of communication** and generates data at **billion-scale** every single day
- Every minute:
  - **WhatsApp** users share 347,222 photos
  - **Facebook Messenger** users share 216,302 photos
  - **Instagram** users post 216,000 photos
  - **Twitter** users send 347,222 tweets
  - **Snapchat** users share 284,722 snaps
  - **Vine** users play 1,041,666 videos
  - **Giphy** serves 596,217 animated images

➢ Models, methods, and algorithms for efficient similarity search and analysis

High Multi-modal Information Bandwidth

- **Content**: visual information, semantic concepts
- **Annotations**: labels, captions, tags
- **Metadata**: technical parameters (Exif, GPS location)

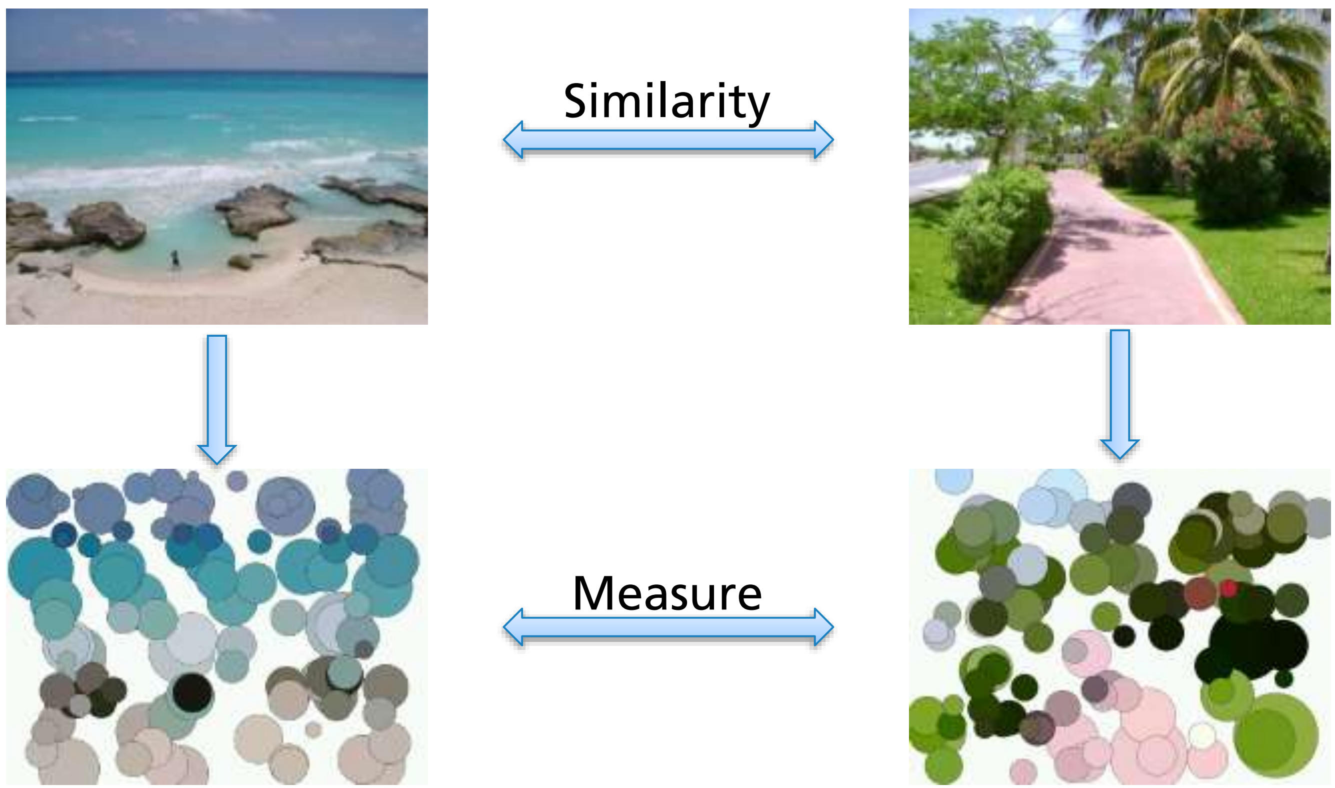
Cancún
Caribbean Sea
Mexico
ICDE 2008
sunny landscape
holidays

Value	Tag
Manufacturer	Sony
Model	DSC-S500
Date and Time	01.04.2008, 18:42:41
x-Resolution	2,816
y-Resolution	2,112

Fraunhofer FIT

Measuring Similarity

- **Similarity model** formalizes the notion of (dis)similarity

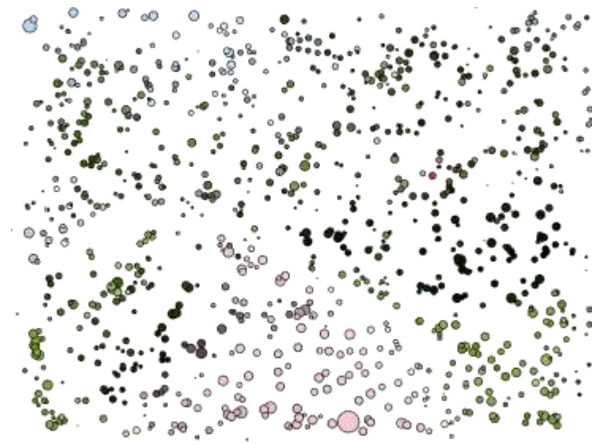


Feature Extraction

- Extraction and description of characteristic properties

multimedia object

feature extraction

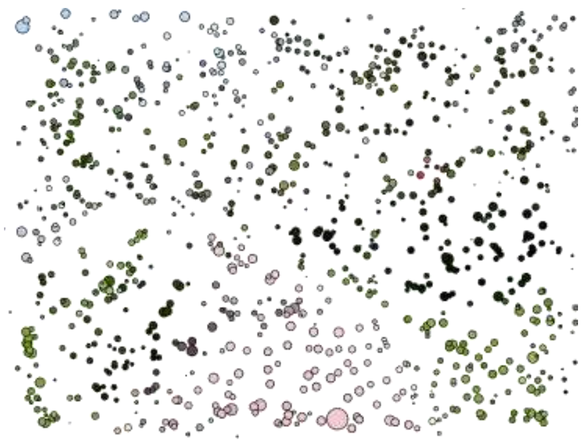


features $f_1, \dots, f_n \in \mathbb{F}$

- Each multimedia object is represented via features
  - Images: Color space $\mathbb{F} = \mathbb{R}^3$ or SIFT space $\mathbb{F} = \mathbb{R}^{128}$
  - Tweets: Term space $\mathbb{F} = \mathbb{D}$ or word embedding space $\mathbb{F} = \mathbb{R}^{300}$

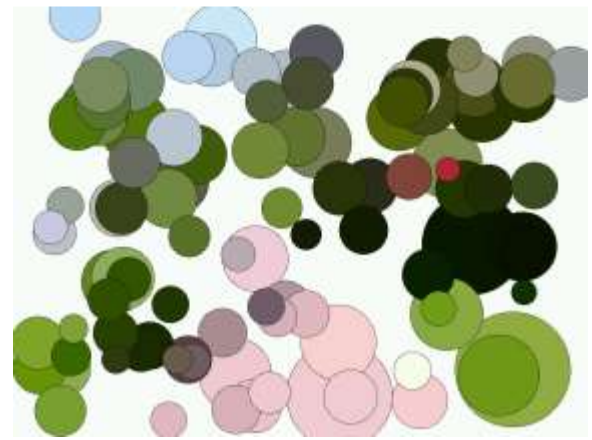
Feature Aggregation

- Aggregation and reduction of characteristic properties



features $f_1, \ldots, f_n \in \mathbb{F}$

feature aggregation

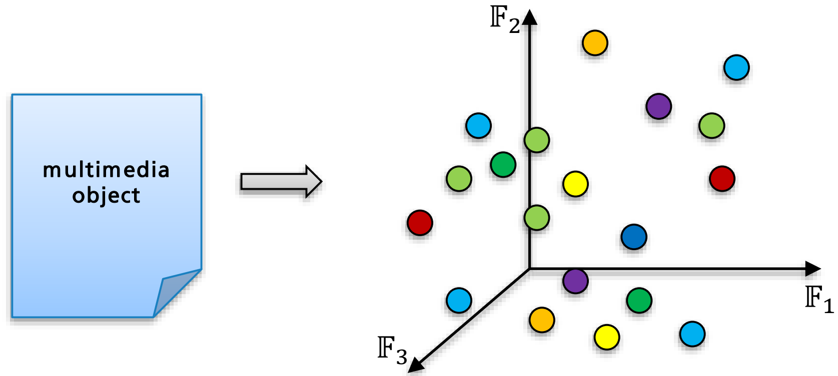


feature representation

- Features are summarized into a smart feature representation structure
  - Clustering algorithms: k-means, expectation maximization, …
  - Hashing: locality sensitive hashing, spectral hashing, …

Smart Multimedia Model: Feature Signature

- **Each multimedia object** is represented as an **individual distribution of features** in a feature space $\mathbb{F}$:

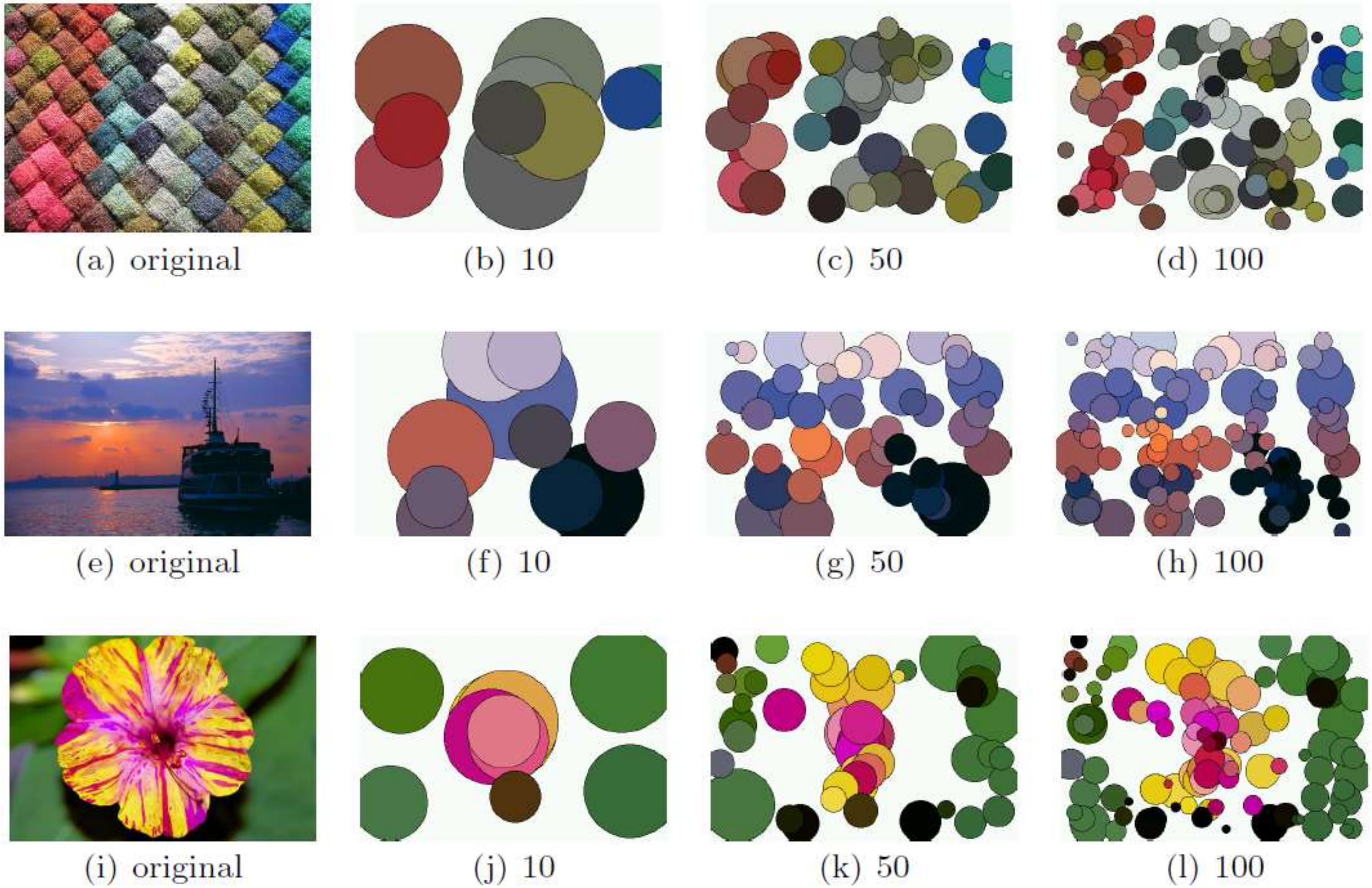


- These features are mathematically modelled by a **feature signature**

$$S\colon \mathbb{F} \to \mathbb{R} \text{ subject to } |\{f \in \mathbb{F} | S(f) \neq 0\}| < \infty$$

C. Beecks: *Distance-based similarity models for content-based multimedia retrieval*. PhD thesis, RWTH Aachen University, 2013.

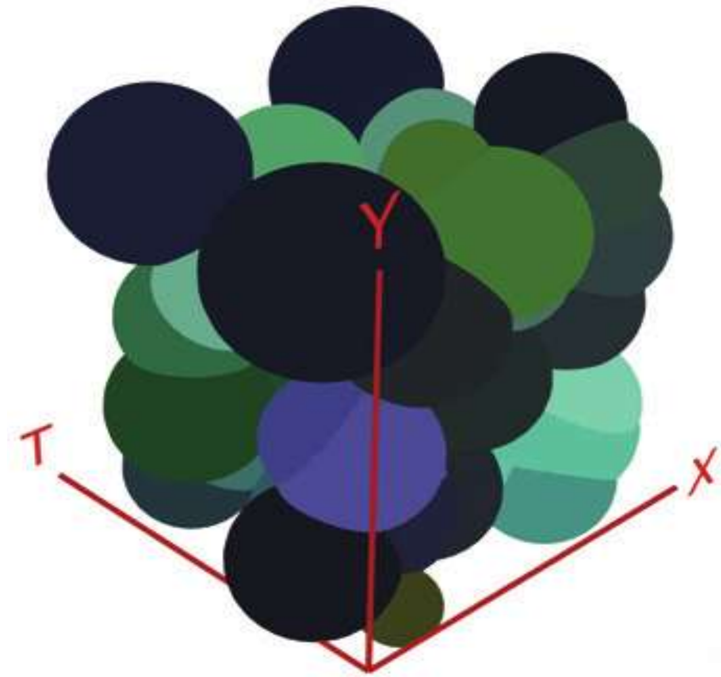
Image Signatures



C. Beecks, S. Kirchhoff, T. Seidl: *On Stability of Signature-based Similarity Measures for Content-based Image Retrieval*. Multimedia Tools Appl. 71(1): 349-362 (2014).

Video Signatures

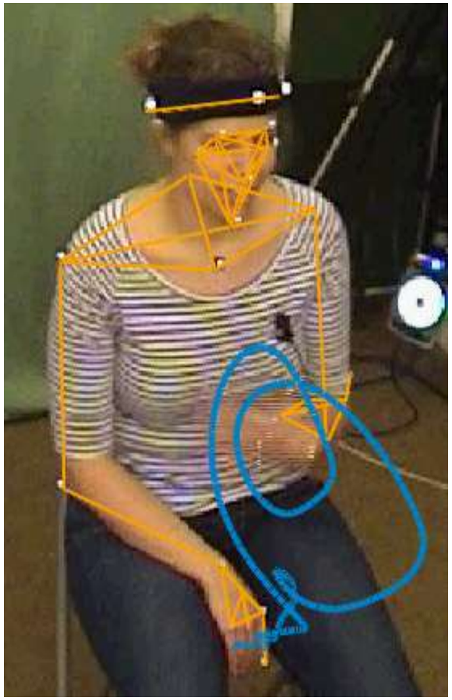
video

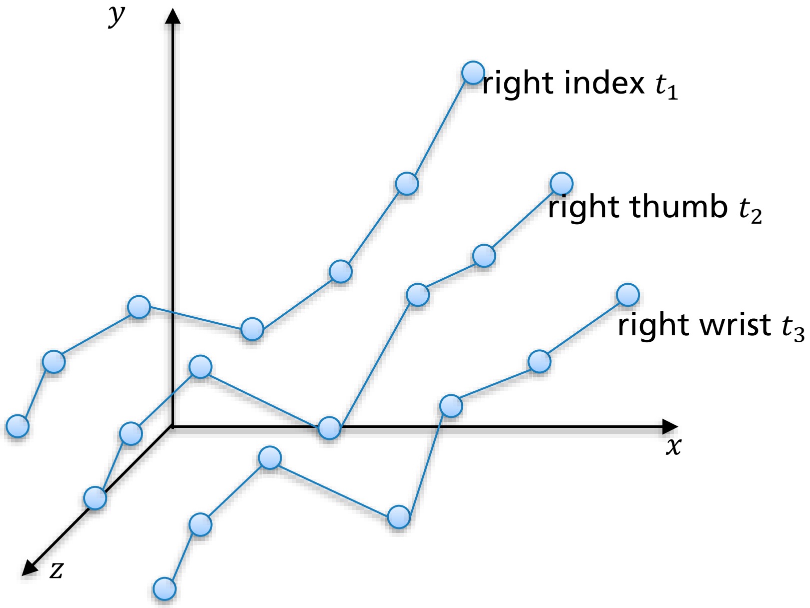


video signature

M. S. Uysal, C. Beecks, D. Sabinasz, J. Schmücking, T. Seidl: *Efficient Query Processing using the Earth's Mover Distance in Video Databases*. EDBT 2016: 389-400.

Gesture Signatures

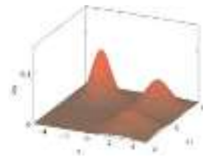




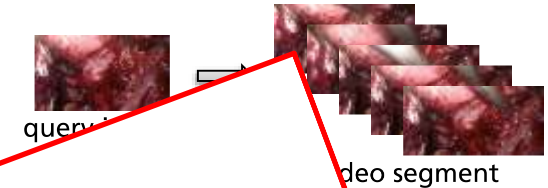
C. Beecks, M. Hassani, J. Hinnell, D. Schüller, B. Brenger, I. Mittelberg, T. Seidl: *Spatiotemporal Similarity Search in 3D Motion Capture Gesture Streams*. SSTD 2015: 355-372.

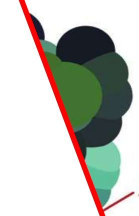
Application in Many Research-oriented Domains

(a) Picturesque sunset



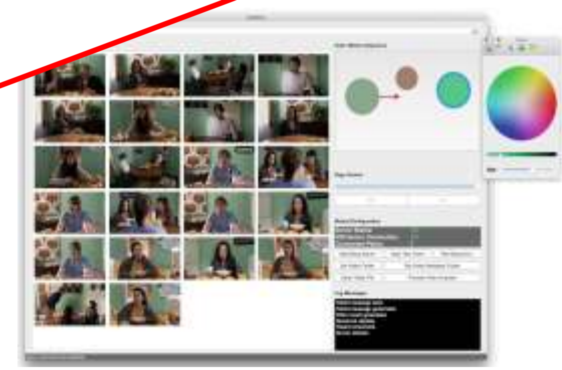
(b) Continuous distribution

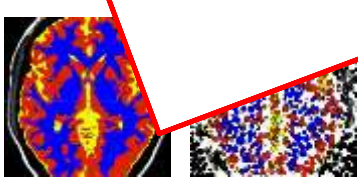


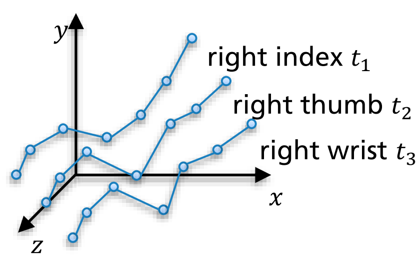


- Modelling Continuous Feature Distributi
- Image Exploration
- Endoscopic Image an
- Video Simil
- C
- G
- Sea
- etc.

Application in Physics?







Fraunhofer FIT

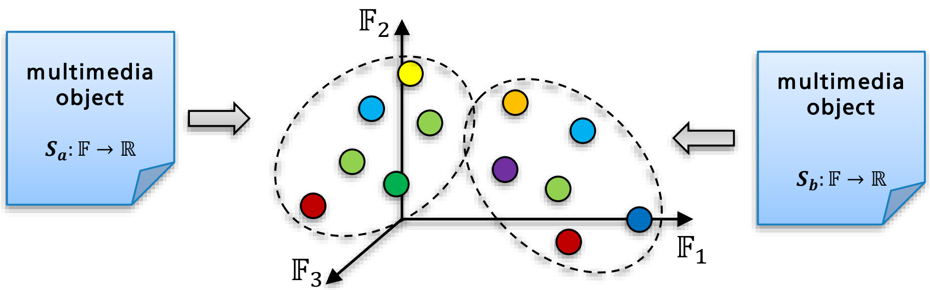
AGENDA

1) Introduction
2) Smart Multimedia Data Representation
   - *How to model multimedia data?*
3) **Adaptive Similarity Models**
   - *How to compare multimedia data?*
4) Efficient Retrieval Approaches
   - *How to search multimedia data?*
5) Time for Discussion

Fraunhofer FIT

Comparison of Feature Signatures

- Feature signatures adapt to multimedia objects
- How to quantify the degree of similarity between two feature signatures?



- Restrict similarity computation to the representatives:

$$\mathrm{R}_{S_a} = \{f \in \mathbb{F} \mid S_a(f) \neq 0\} \text{ and } \mathrm{R}_{S_b} = \{f \in \mathbb{F} \mid S_b(f) \neq 0\}$$

- Necessity of a ground distance between features $f \in \mathbb{F}$

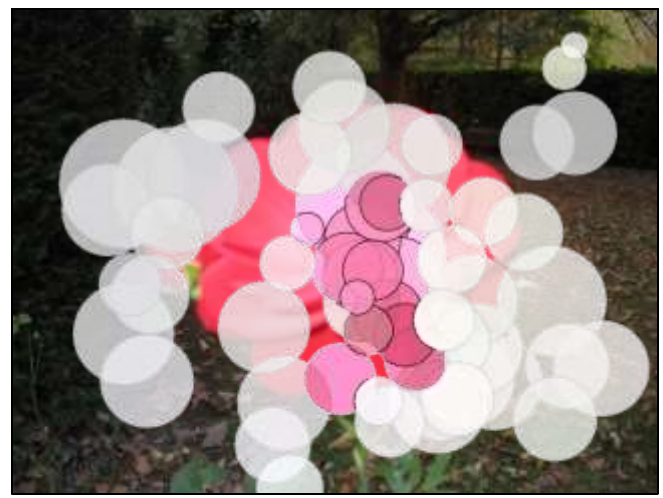
Fraunhofer FIT

Signature-based Similarity Measures

- **Matching-based measures**
  - Hausdorff Distance [Hausdorff, 1914]
  - Perceptually Modified Hausdorff Distance [Park et al., 2008]
  - Signature Matching Distance [Beecks et al., 2013a]
- **Transformation-based measures**
  - Earth Mover's Distance [Rubner et al., 2000]
- **Correlation-based measures**
  - Weighted Correlation Distance [Leow and Li, 2004]
  - Signature Quadratic Form Distance [Beecks et al., 2009, 2010a]

Signature Matching Distance

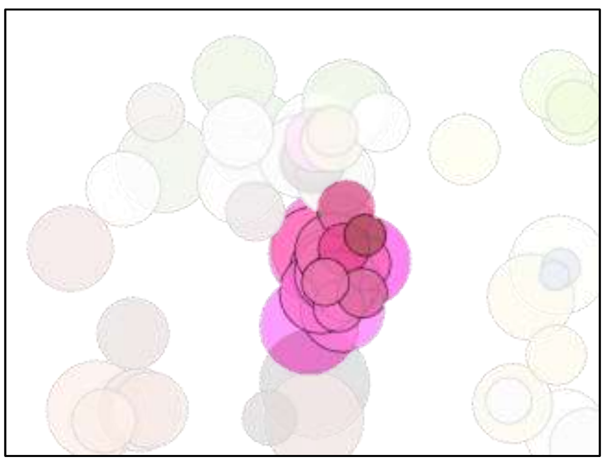
- **Idea:** Attribute distance definition to the most similar parts
- **Approach:**
  1. Computation of a **matching**
  2. Computation of a **cost function** that defines the distance

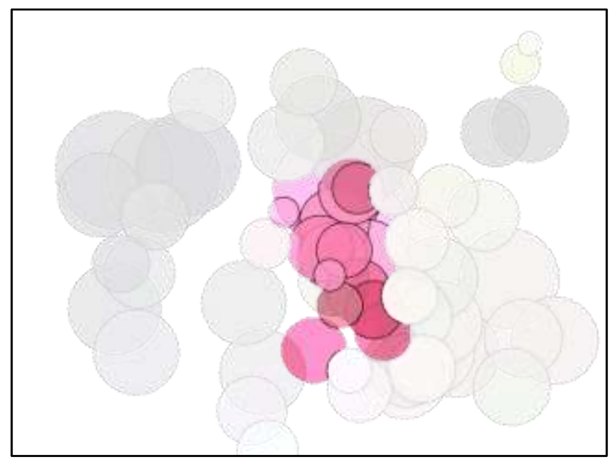


Matching

- A **matching** between two feature signatures X and Y is defined as a subset of the Cartesian product of the representatives R_X and R_Y:

$$m_{X\leftrightarrow Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$$



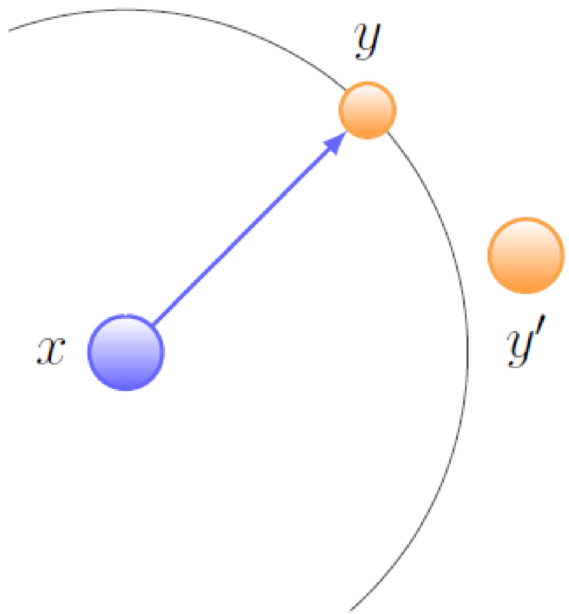


Fraunhofer FIT

Nearest Neighbor Matching

- Each representative is matched to its nearest neighbor

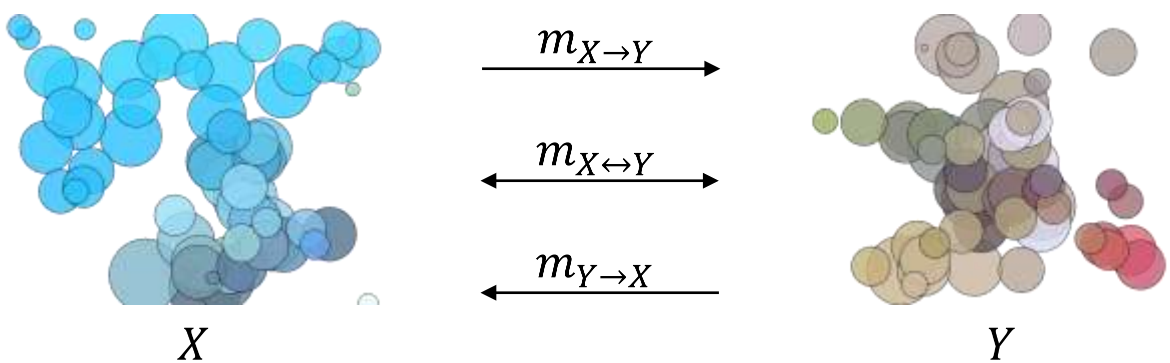
$$m^{\mathrm{NN}}_{X\to Y} = \left\{(x,y) \mid x \in \mathrm{R}_X \wedge y \in \mathrm{NN}_{\delta,\mathrm{R}_Y}(x)\right\}$$



Signature Matching Distance: Definition

- **Signature Matching Distance** between two feature signatures X and Y is defined as:

$$\mathrm{SMD}_{\delta}(X,Y) = c(m_{X\to Y}) + c(m_{Y\to X}) - 2\cdot\lambda\cdot c(m_{X\leftrightarrow Y})$$



where the cost function c evaluates the dissimilarity of a matching

Fraunhofer FIT

Signature Quadratic Form Distance

- Instead of using a matching, we can also use the **similarity correlation** between two feature signatures:

$$\langle X, Y \rangle_s = \sum_{f \in \mathrm{R}_X} \sum_{g \in \mathrm{R}_Y} X(f) \cdot Y(g) \cdot s(f, g)$$

- **Signature Quadratic Form Distance** between two feature signatures:

$$\mathrm{SQFD}_s(X, Y) = \sqrt{\langle X - Y, X - Y \rangle_s}$$

C. Beecks, M. S. Uysal, T. Seidl: *Signature Quadratic Form Distance*. CIVR 2010: 438-445.

Earth Mover's Distance

- We could also model the similarity computation as a transportation problem and apply the **Earth Mover's Distance**

$$\mathrm{EMD}_\delta(X,Y) = \min_{\{f \mid f:\mathbb{F}\times\mathbb{F}\to\mathbb{R}\}} \left\{ \frac{\sum_{g\in \mathrm{R}_X} \sum_{h\in \mathrm{R}_Y} f(g,h)\cdot\delta(g,h)}{\min\{\sum_{g\in \mathrm{R}_X} X(g), \sum_{h\in \mathrm{R}_Y} Y(h)\}} \right\}$$

  subject to the constraints:

  - CNNeg: $\forall g \in \mathrm{R}_X, \forall h \in \mathrm{R}_Y : f(g,h) \geq 0$
  - CSource: $\forall g \in \mathrm{R}_X$: $\sum_{h\in \mathrm{R}_Y} f(g,h) \leq X(g)$
  - CTarget: $\forall h \in \mathrm{R}_Y$: $\sum_{g\in \mathrm{R}_X} f(g,h) \leq Y(h)$
  - CTotalFlow: $\sum_{\substack{g\in \mathrm{R}_X \\ h\in \mathrm{R}_Y}} f(g,h) = \min\{\sum_{g\in \mathrm{R}_X} X(g), \sum_{h\in \mathrm{R}_Y} Y(h)\}$

- Earth Mover's Distance [RTG98] is also known as first-degree Wasserstein or Mallows Distance [D70, LB01]

Fraunhofer FIT

Performance Evaluation

- Evaluation of feature signature approaches on the Holidays database [JDS08] and UKBench database [NS06]:

		Holidays		UKBench		
		MAP	size	MAP	score	size
EMD	PCT	0.720	90	0.741	2.78	50
	SIFT	0.678	70	0.536	2.05	90
	CSIFT	0.749	40	0.605	2.31	30
PMHD	PCT	0.804	80	**0.866**	3.30	90
	SIFT	0.673	70	0.531	2.03	90
	CSIFT	0.755	40	0.594	2.27	30
SQFD	PCT	0.761	40	0.766	2.86	60
	SIFT	0.690	80	0.585	2.23	100
	CSIFT	0.756	20	0.494	2.25	20
SMD	PCT	**0.810**	100	0.845	3.20	60
	SIFT	0.653	40	0.463	1.75	20
	CSIFT	0.735	20	0.531	2.00	20

AGENDA

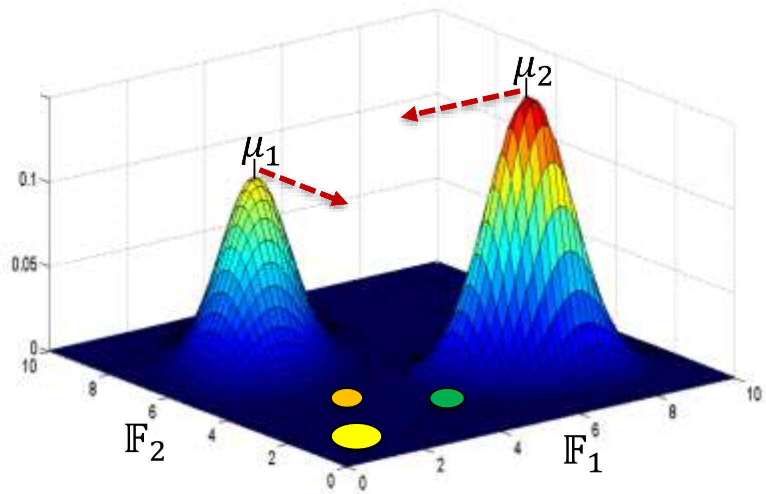
1) Introduction
2) Smart Multimedia Data Representation
   - *How to model multimedia data?*
3) Adaptive Similarity Models
   - *How to compare multimedia data?*
4) **Efficient Retrieval Approaches**
   - *How to search multimedia data?*
5) Time for Discussion

Potpourri

- **Model-specific approaches**
  - Signature Quadratic Form Distance
    - Maximum components approximation [ICDEW'10]
    - Similarity matrix compression [SISAP'10]
    - L_2– Signature Quadratic Form Distance [MMM'11]
    - GPU-based query processing [CIKM'11, JDPDB'12]
  - Earth Mover's Distance
    - IM-Sig constraint relaxation [CIKM'14, EDBT'16]
    - Dimensionality reduction [SSDBM'15]
- **Generic approaches**
  - Metric Indexing [ICMR'11]
  - Ptolemaic Indexing [SISAP'11, InfSyst'13]
  - Gradient-based Approximation [CIKM'15]
  - Multi-step Threshold Algorithm [IEEEBigData'16]

Gradient-based Approximation (1)

- **Parameter-based approximation** via a generative model



- **Gradient-based signature** $S_\nabla : \Theta \rightarrow \mathbb{R}$
  - representatives $\mathrm{R}_{S_\nabla} = \{\lambda \in \theta \mid S_\nabla(\lambda) \neq 0\}$
  - weights $S_\nabla(\lambda) = \nabla_\lambda \log \mathcal{L}(\theta|S) = \nabla_\lambda \log \prod_{f \in \mathbb{F}} p(f|\theta)^{S(f)}$

C. Beecks, M. S. Uysal, J. Hermanns, T. Seidl: *Gradient-based Signatures for Efficient Similarity Search in Large-scale Multimedia Databases*. CIKM 2015: 1241-1250.

Gradient-based Approximation (2)

- **Retrieval accuracy** [mean average precision]

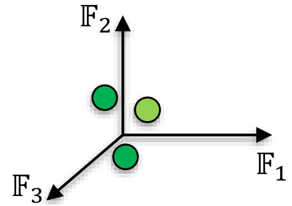
	Holidays	UKBench
Signature Matching Distance	0.81	0.85
Earth Mover's Distance	0.72	0.74
Signature Quadratic Form Distance	0.76	0.77
Gradient-based signatures + L_1	**0.78**	**0.82**
Binary Gradient-based signatures + XOR	**0.73**	**0.75**

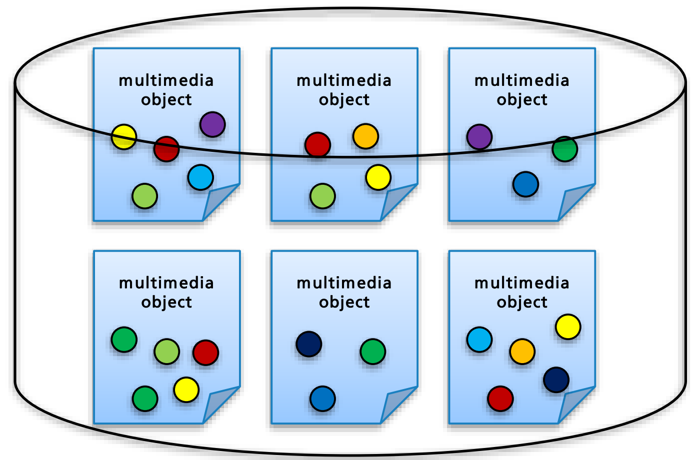
- **Retrieval efficiency** [query response time]
  - Gradient-based signatures: approximately **1.1 seconds**
  - Binary Gradient-based signatures: approximately **0.5 seconds**

Multi-step Threshold Algorithm (1)

- **Asymmetric** and **matching-based retrieval model:**

$$Q = \{q_1, q_2, q_3\}$$





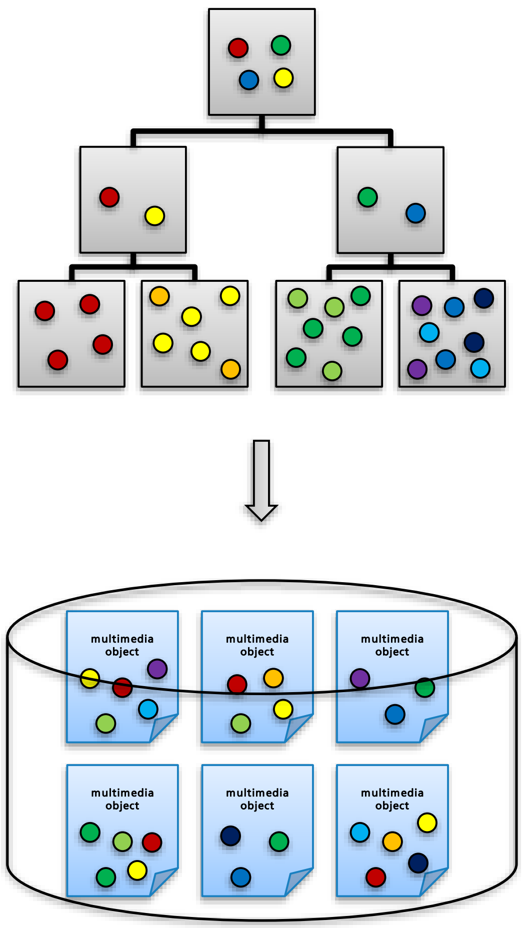
- **(Dis)similarity measure** between query Q and feature signature S based on feature distance δ:

$$D(Q,S) = \sum_{q \in Q} \min_{S(f \in \mathbb{F}) \neq 0} \delta(q,f)$$

Fraunhofer FIT

Multi-step Threshold Algorithm (2)

- **In-memory feature index**
  - Stores tuples of **features** and **signature IDs** $\langle f_1, ID_{S_1}\rangle, \langle f_2, ID_{S_1}\rangle, \ldots, \langle f_m, ID_{S_n}\rangle \in \mathbb{F} \times \mathbb{N}$
  - Feature space $\mathbb{F}$ is structured into nodes $\mathcal{N}$
  - Supported methods:
    - `node.getMinDist(Feature f)`
    - `idx.getNextNode(Feature f)`
    - `idx.getNodes(Identifier id)`
- **Multimedia database**
  - Stores **feature signatures** $S_1, \ldots, S_n \in \mathbb{R}^{\mathbb{F}}$
  - Supported methods:
    - `db.getSignature(Identifier id)`



Fraunhofer FIT

Multi-step Threshold Algorithm (3)

- **Feature-by-feature** query processing
  - **Candidate generation phase**
    - **Threshold algorithm** [FLN01] is used to generate candidates in parallel
    - Sorted and random access to the in-memory feature index
  - **Candidate refinement phase**
    - **Optimal multi-step algorithm** [SK98] is used to refine possible candidates
    - Sorted access to the underlying multimedia database
- **Optimizes both index and I/O access**

C. Beecks, A. Graß: *Multi-step Threshold Algorithm for Efficient Feature-based Query Processing in Large-scale Multimedia Databases*. IEEE BigData 2016.

Algorithm 1: Multi-step Threshold Algorithm (MTA)

Input: database db with index structure idx, query features $Q = \{q_i\}_{i=1}^{l}$, number of nearest neighbors k

Output: k-nearest-neighbors $\mathrm{NN}_k \subseteq db$

$candidates \leftarrow$ newMinHeap($\langle id, \widetilde{D}_{id} \rangle$) ;
$results \leftarrow$ newMaxHeap($\langle S_{id}, D_{id} \rangle$) ;
while idx.hasNext() **do**
  /* candidate generation phase */
  $\theta \leftarrow 0$;
  foreach $q_i \in Q$ **do**
    $node \leftarrow idx$.getNextNode(q_i) ;
    $D_{\min} \leftarrow node$.getMinDist(q_i) ;
    foreach $id \in node$ **do**
      if !$candidates$.contains($\langle id, * \rangle$) **then**
        $nodes_{id} \leftarrow idx$.getNodes($id$) ;
        $\widetilde{D}_{id} \leftarrow$ computeLowerBound(Q,$nodes_{id}$) ;
        $candidates$.push($\langle id, \widetilde{D}_{id} \rangle$) ;
    $\theta \leftarrow \theta + D_{\min}$;
  /* candidate refinement phase */
  while $candidates$.peek().distance() $\leq \theta$ **do**
    $\langle id, \widetilde{D}_{id} \rangle \leftarrow candidates$.pop() ;
    if $results$.size() $< k$ **then**
      $S_{id} \leftarrow db$.getSignature(id) ;
      $D_{id} \leftarrow$ computeDissimilarity(Q, S_{id}) ;
      $results$.push($\langle S_{id}, D_{id} \rangle$) ;
    else
      if $results$.peek().distance() $\geq \widetilde{D}_{id}$ **then**
        $S_{id} \leftarrow db$.getSignature(id) ;
        $D_{id} \leftarrow$ computeDissimilarity(Q, S_{id}) ;
        $results$.push($\langle S_{id}, D_{id} \rangle$) ;
        $results$.pop() ;
      else
        return $results$;

Fraunhofer FIT

Multi-step Threshold Algorithm (4)

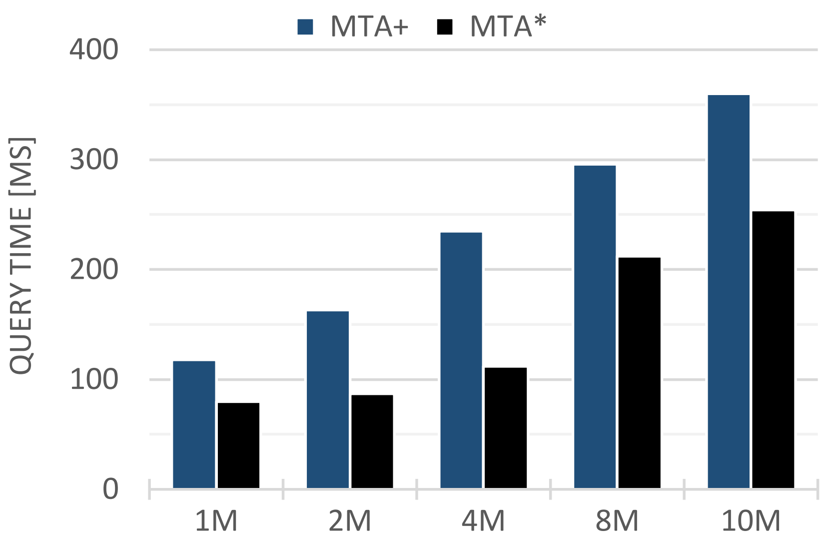
- **Retrieval accuracy** [mean average precision]

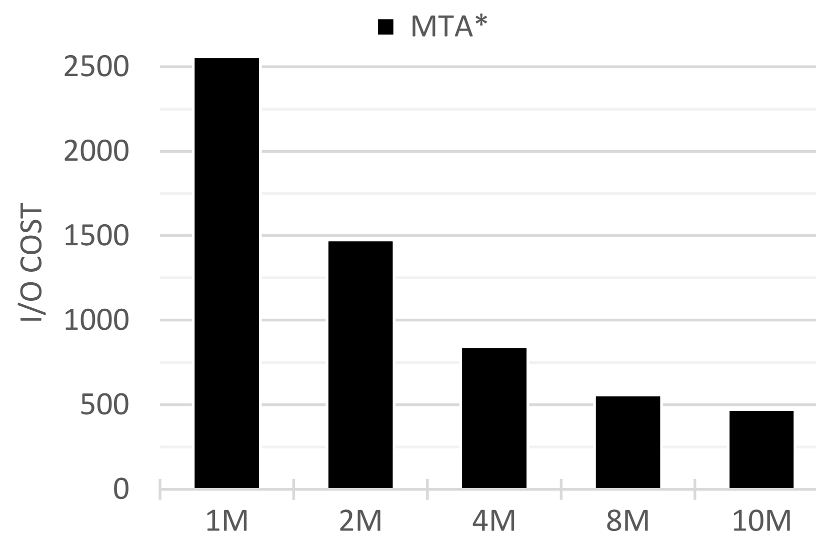
	Holidays	UKBench
Signature Matching Distance	0.81	0.85
Earth Mover's Distance	0.72	0.74
Signature Quadratic Form Distance	0.76	0.77
Gradient-based signatures + L_1	0.78	0.82
Binary Gradient-based signatures + XOR	0.73	0.75
Multi-step Threshold Algorithm	**0.83**	**0.89**

Fraunhofer FIT

Multi-step Threshold Algorithm (5)

- **Retrieval efficiency** [query response time]
  - **MTA⁺**: parallel variant
  - **MTA***: parallel variant + distance caching





Fraunhofer FIT

Conclusions

- **Efficient similarity search** requires **adaptive similarity models** and **scalable processing techniques**

- **Feature signature model** is a generic model that supports various data types and different features

- **Multi-step Threshold Algorithm** is an efficient and scalable solution for high-performance multimedia analysis

Fraunhofer FIT

AGENDA

1) Introduction
2) Smart Multimedia Data Representation
   - *How to model multimedia data?*
3) Adaptive Similarity Models
   - *How to compare multimedia data?*
4) Efficient Retrieval Approaches
   - *How to search multimedia data?*
5) **Time for Discussion**

Fraunhofer FIT

Thank you for attention!

Dr. Christian Beecks

Fraunhofer Institute for Applied Information Technology FIT

User-centered Ubiquitous Computing

http://www.fit.fraunhofer.de

christian.beecks@fit.fraunhofer.de

Fraunhofer FIT