Similarity Searching for Database Applications

Pavel Zezula

Masaryk University

Brno, Czech Republic

Outline of the talk

- On the importance of similarity and searching
- Principles of metric similarity searching
- Similarity search applications:
  – Searching in images of human faces
  – Searching for image annotation
  – Stream processing
  – Searching in motion capture data

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-Life Motivation

The social psychology view

- Any event in the history of organism is, in a sense, **unique**.
- *Recognition*, *learning*, and *judgment* presuppose an ability to categorize stimuli and classify situations by **similarity**.
- Similarity (*proximity*, *resemblance*, *communality*, *representativeness*, *psychological distance*, etc.) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.
- Similarity is **subjective** a **context-dependent**

Contemporary Networked Media

The digital data view

- Almost **everything** that we *see*, *read*, *hear*, *write*, *measure*, or *observe* can be **digital**.
- Users **autonomously** *contribute* to production of global media and the growth is **exponential**.
- Sites like Flickr, YouTube, Facebook host user contributed content for a variety of **events**.
- The elements of networked media are related by numerous multi-facet **links of similarity**.

Challenge

- Networked media database is getting close to the human "fact-bases"
  - the gap between physical and digital world has blurred

- **Similarity data management** is needed to *connect*, *search*, *filter*, *merge*, *relate*, *rank*, *cluster*, *classify*, *identify*, or *categorize* objects across various collections.

WHY?

It is the *similarity* which is in the world *revealing*.

Similarity and the Big Data

- ***Loads*** on a sharp ***rise*** – ***usage*** on ***decline***
- The (**3V**) problem of: ***Volume, Variety, Velocity***
- **Issues:**
  - **Acquisition**: what to keep and what to discard
  - **Unstructured data**: what content to extract
  - **Datafication**: render into data many new aspects
  - **Inaccuracy**: approximation, imprecision, noise

The Big Data problem

- **Shifts in thinking**:
  - from some to all (scalability)
  - from clean to messy (approximate)
- **Technological obstacles**: *heterogeneity*, *scale*, *timeliness*, *complexity*, and *privacy* aspects
- **Foundational challenges**: *scalable* and *secure* data *analysis*, *organization*, *retrieval*, and *modeling*

Search – the goals

1. We **search** to get results (papers, books, …)
2. We **ask** to find answers (what time …)
3. We use **filters** so that the right staff finds us
4. We **browse** while wandering and way-finding in typically restricted space

- In reality, we move fluidly between modes of

ask, ***browse***, ***filter***, and ***search***

Search – some quantitative facts

- 85% of all web traffic comes from search engines
- 450+ million searches/day are performed in North America alone
- 70%+ of all searches are done on Google sites

Search is the **most popular** application

(second to E-mail??)

Search – some experience

- 60% of searchers NEVER go past 1st page of search results
- The top three results draw 80% of the attention
- The first few results inordinately influence query reformulation.

Search - as an interaction

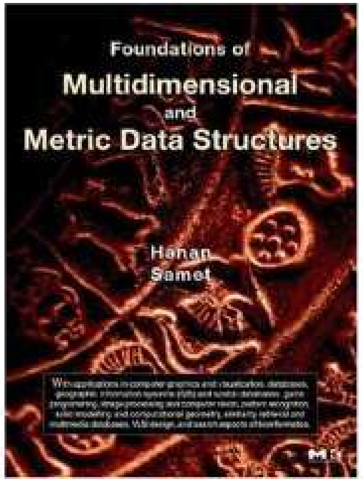
- When we search, our next actions are reactions to the stimuli of previous search results
- What we find is changing what we seek
- In any case, search must be:

fast, ***simple***, and ***relevant***

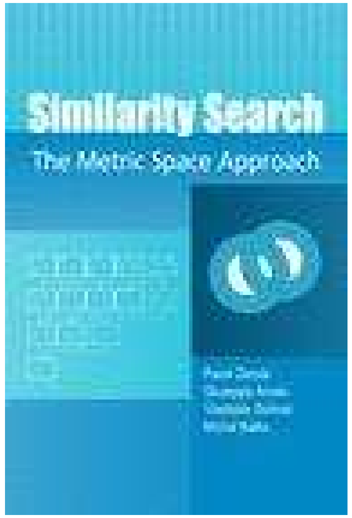
Search – changes our cognitive habits

1. We are increasingly handing off the job of remembering to search engines
2. When we expect information to be easily found again, we do not remember it well
3. Our original memory of facts is changing to a memory of ways to find the facts

State of the art in Metric Searching technology



Hanan Samet
Foundation of Multidimensional and Metric Data Structures
Morgan Kaufmann, 2006



P. Zezula, G. Amato, V. Dohnal, and M. Batko
Similarity Search: The Metric Space Approach
Springer, 2005

Teaching material:
http://www.nmis.isti.cnr.it/amato/similarity-search-book/

Similarity Search Conferences

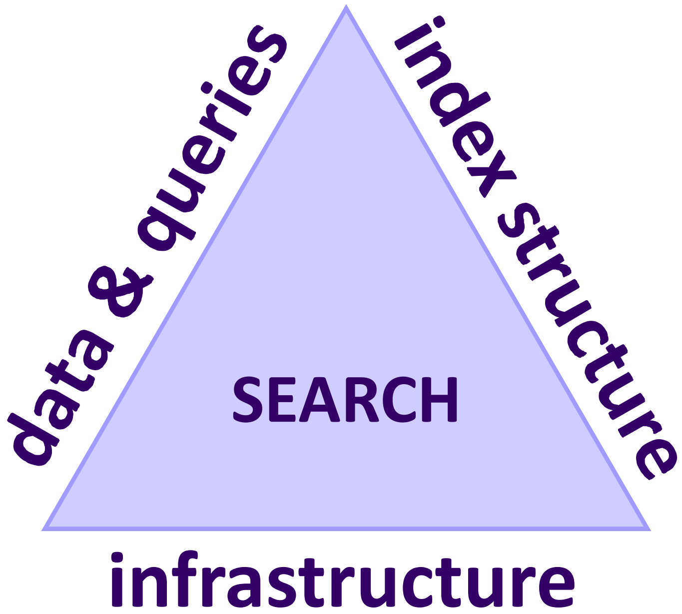
9^{th} SISAP 2016, October 24-26, Tokyo, Japan

The MUFIN Approach

MUFIN: MUlti-Feature Indexing Network

Extensibility
metric space

Scalability
P2P structure



Independence
Infrastructure as a service

Extensibility: Metric Abstraction of Similarity

- Metric space: $\mathcal{M} = (\mathcal{D}, d)$
  - $\mathcal{D}$ – domain
  - distance function $d(x,y)$

    $\forall x,y,z \in \mathcal{D}$
    - $d(x,y) > 0$ - *non-negativity*
    - $d(x,y) = 0 \Leftrightarrow x = y$ - *identity*
    - $d(x,y) = d(y,x)$ - *symmetry*
    - $d(x,y) \leq d(x,z) + d(z,y)$ - *triangle inequality*

Examples of Distance Functions

- L_p **Minkovski distance** (for vectors)
  - L_1 – city-block distance
  - L_2 – Euclidean distance
  - L_∞ – infinity

$$L_1(x, y) = \sum_{i=1}^{n} | x_i - y_i |$$

$$L_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$L_\infty(x, y) = \max_{i=1}^{n} |x_i - y_i|$$

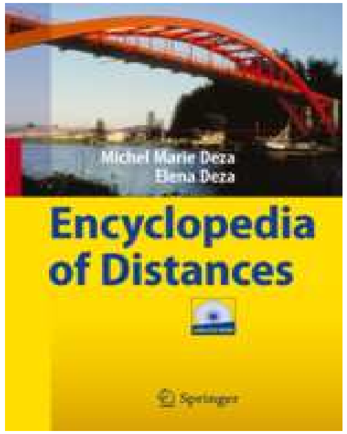
- **Edit distance** (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6

- **Jaccard's coefficient** (for sets A,B)

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Examples of Distance Functions

- **Mahalanobis distance**
  - for vectors with correlated dimensions
- **Hausdorff distance**
  - for sets with elements related by another distance
- **Earth movers distance**
  - primarily for histograms (sets of weighted features)
- and many others



Similarity Search Problem

- For $X \subseteq \mathcal{D}$ in metric space $\mathcal{M}$, pre-process X so that the similarity queries are executed efficiently.

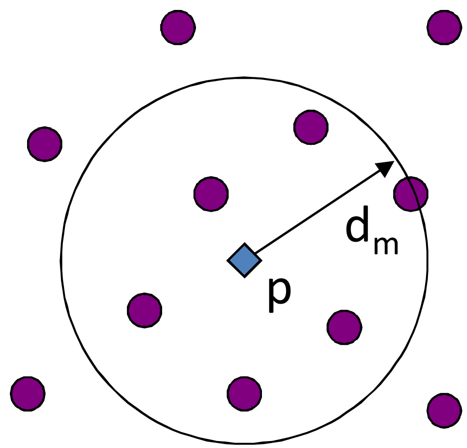
In metric space: no total ordering exists!

Basic Partitioning Principles

- Given a set $X \subseteq \mathcal{D}$ in $\mathcal{M}=(\mathcal{D},d)$, basic partitioning principles have been defined:
  - Ball partitioning
  - Generalized hyper-plane partitioning
  - Excluded middle partitioning

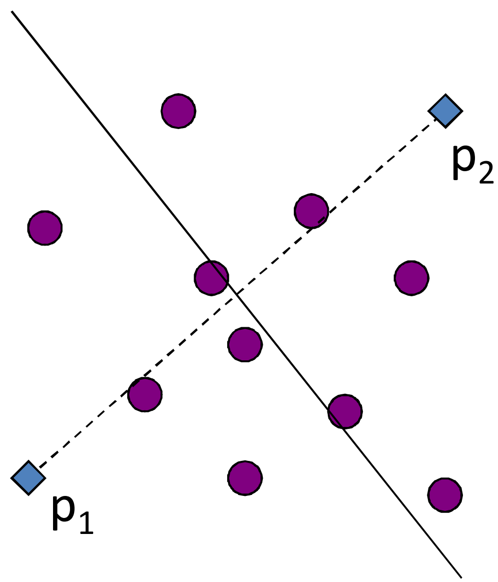
Ball Partitioning

- Inner set: $\{ x \in X \mid d(p,x) \leq d_m \}$
- Outer set: $\{ x \in X \mid d(p,x) > d_m \}$

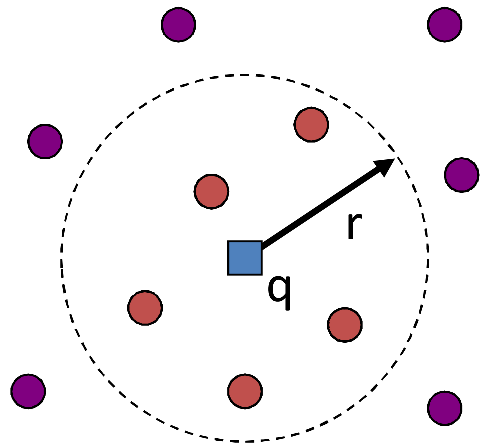


Generalized Hyper-plane

- $\{ x \in X \mid d(p_1,x) \leq d(p_2,x) \}$
- $\{ x \in X \mid d(p_1,x) > d(p_2,x) \}$



Similarity Range Query



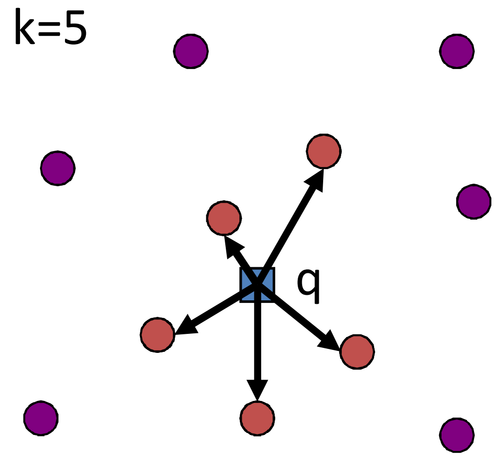
- range query
  - $R(q,r) = \{ x \in X \mid d(q,x) \le r \}$

… all museums up to 2km from my hotel …

Nearest Neighbor Query

- the nearest neighbor query
  - $NN(q) = x$
  - $x \in X, \forall y \in X, d(q,x) \leq d(q,y)$

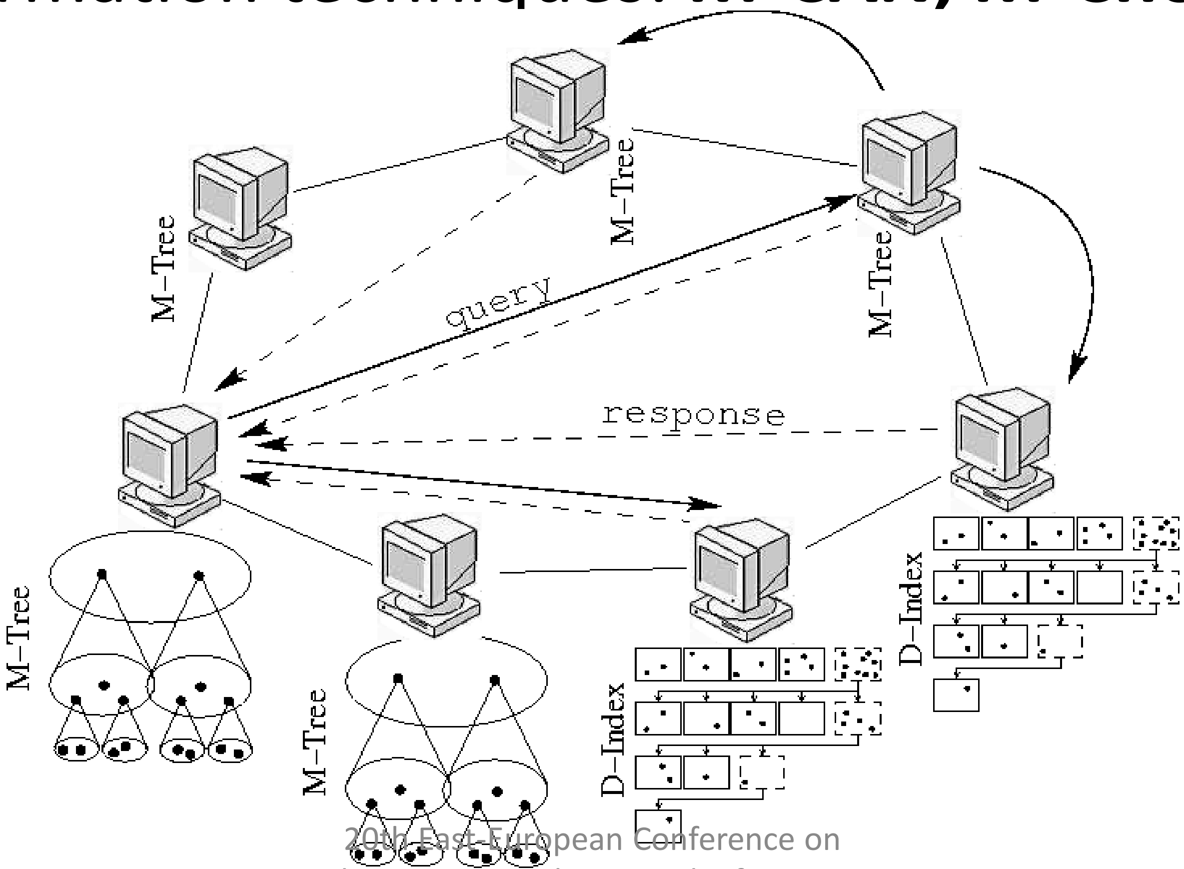
- k-nearest neighbor query
  - $k\text{-}NN(q,k) = A$
  - $A \subseteq X, |A| = k$
  - $\forall x \in A, y \in X - A, d(q,x) \leq d(q,y)$



... five closest museums to my hotel ...

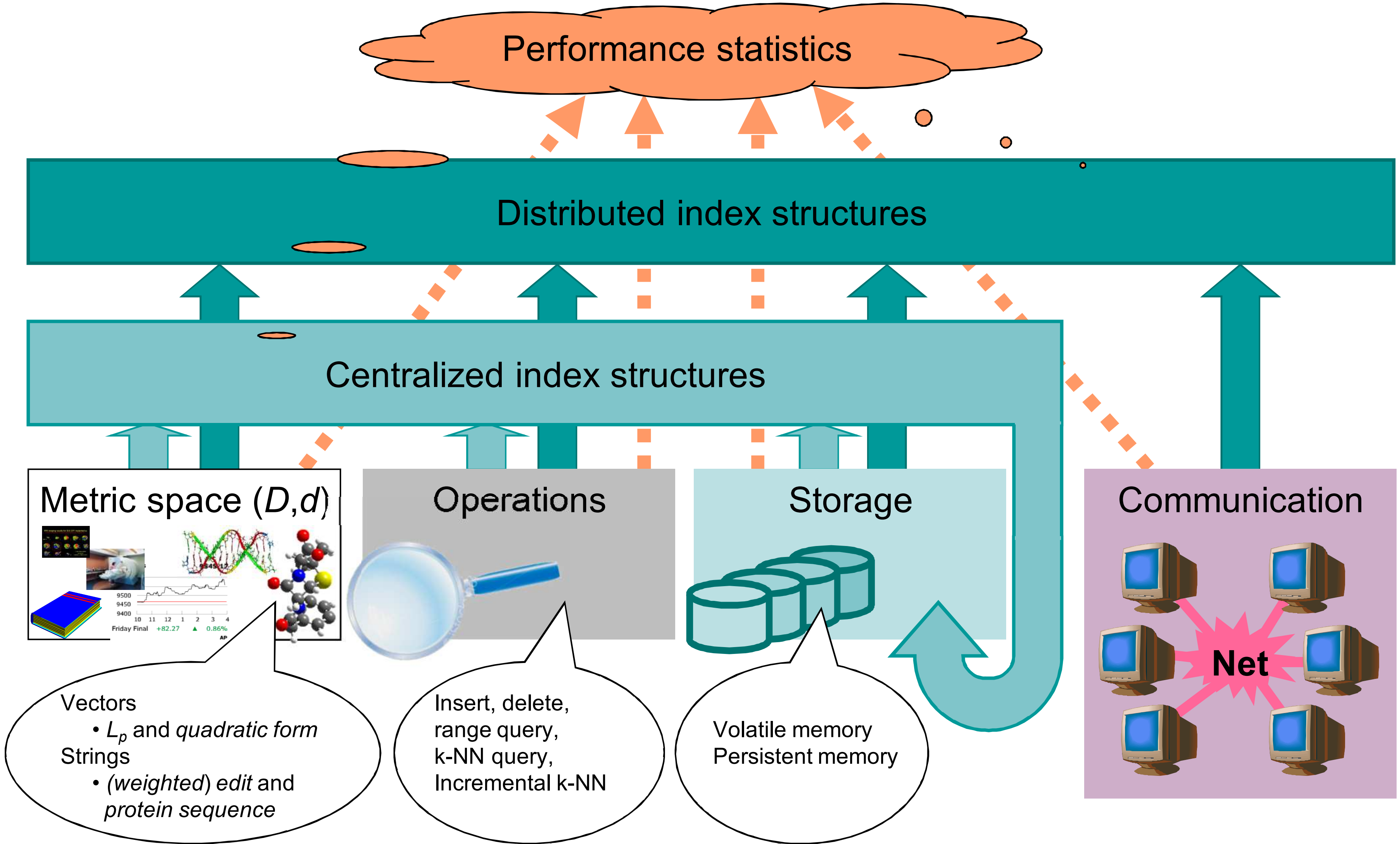
Scalability: Peer-to-Peer Indexing

- Local search: **Main memory structures**
- Native metric techniques: **GHT*, VPT***
- Transformation techniques: **M-CAN, M-Chord**



Infrastructure Independence: MESSIF

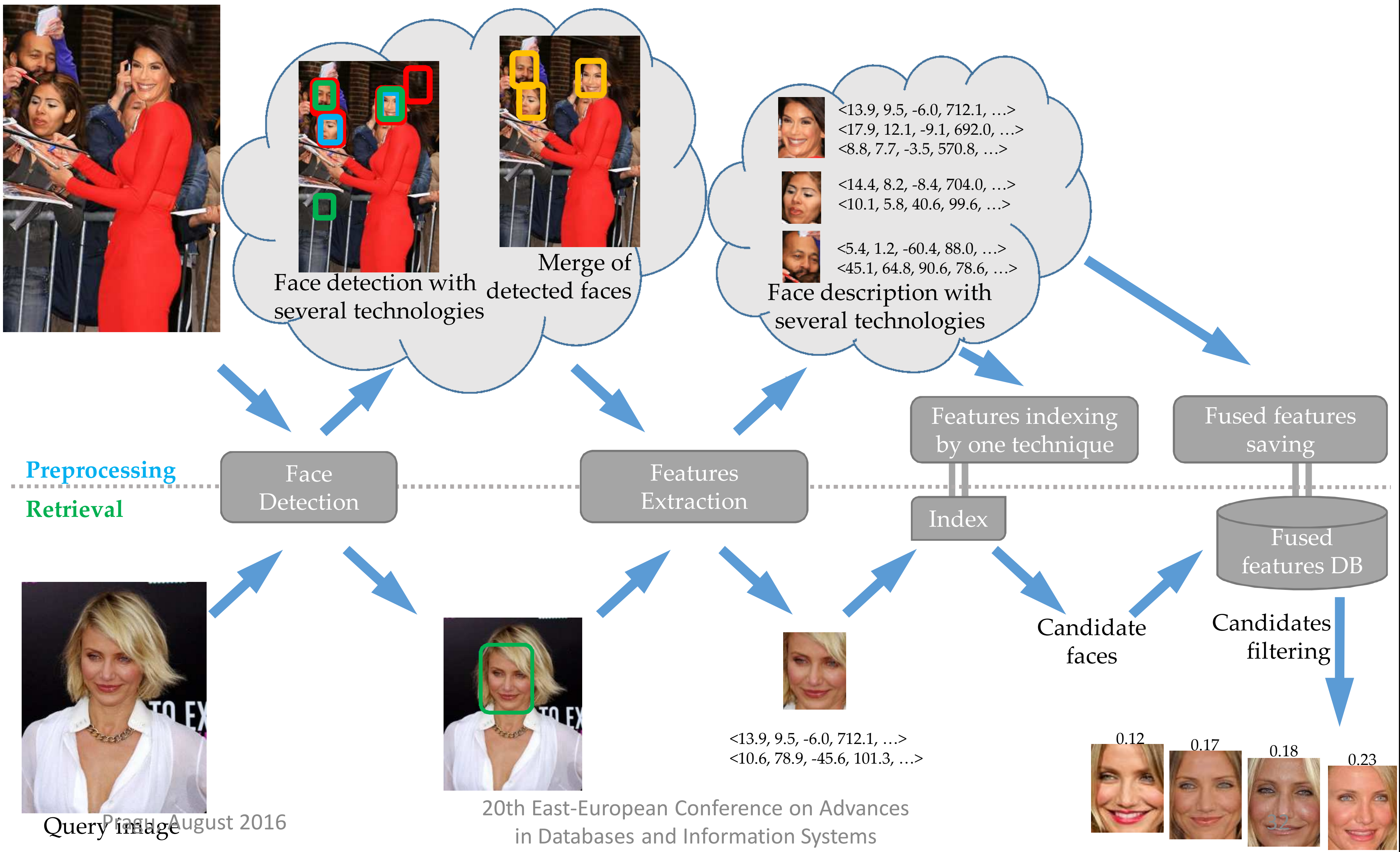
Metric Similarity Search Implementation Framework



MUFIN demos

- http://disa.fi.muni.cz/imgsearch/similar
- http://www.pixmac.com/
- http://disa.fi.muni.cz/twenga/
- http://disa.fi.muni.cz/fingerprints/
- http://disa.fi.muni.cz/subseq/
- http://disa.fi.muni.cz/FaceMatch/
- http://disa.fi.muni.cz/annotation/
- http://disa.fi.muni.cz/motion-match/
- http://disa.fi.muni.cz/profimedia-neural_network-20M/

Similarity Search in Collections of Faces



Fused Face Detection and Face Matching

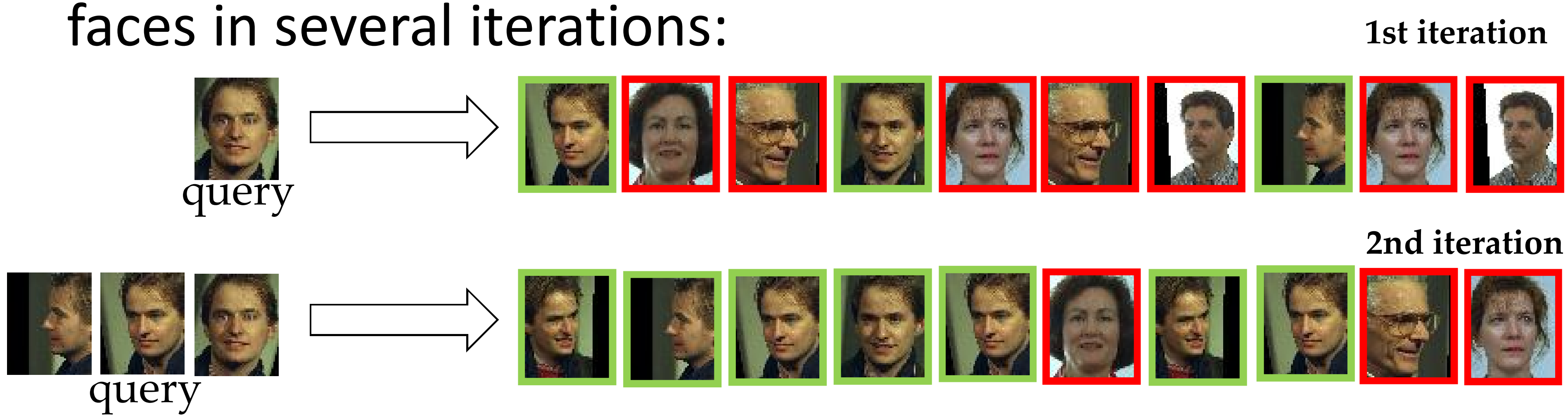
- Fused face detection:
  - Faces detected by more technologies are taken into account
  - Showcase: 3 technologies, compliance of at least two:

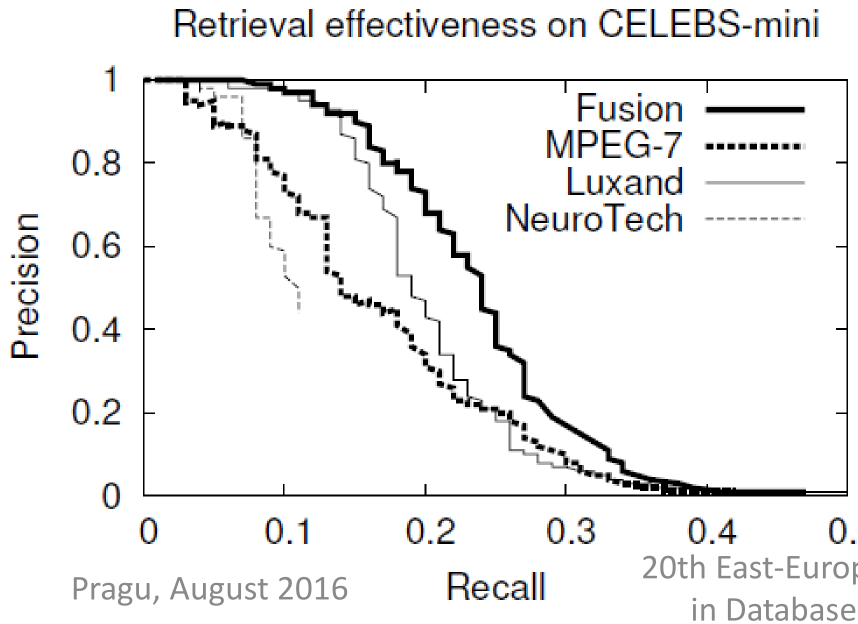
Software name	OpenCV	Luxand	Verilook	Compliance of at least 2
Recall / precision (%)	55 / 89	64 / 83	73 / 83	64 / 96

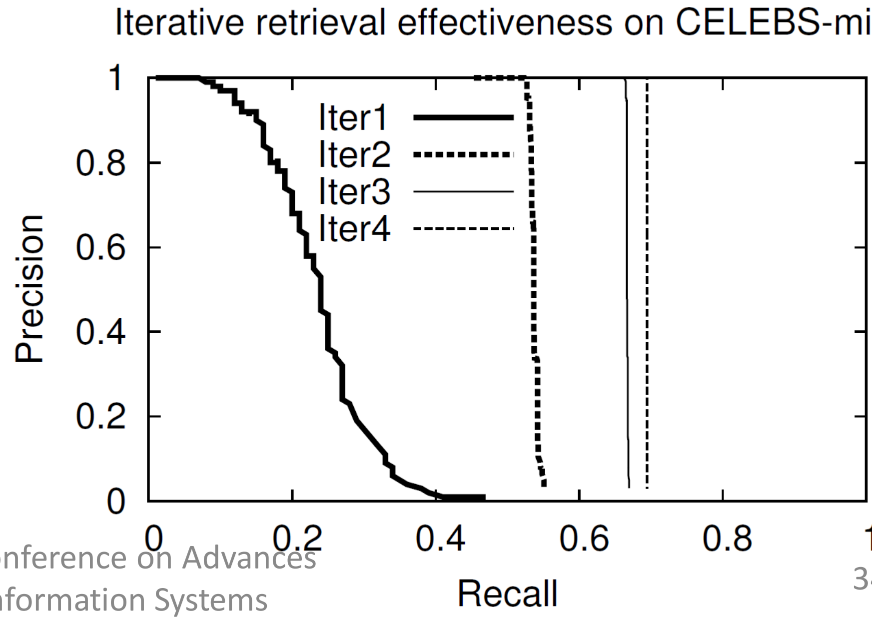
- Fused face matching:
  - Characteristic features from more technologies are available for each face
  - Similarity of two faces evaluated by each technology is normalized into interval [0, 1]
  - Normalized value expresses a probability that faces belong to the same person
  - Highest probability is used to determine the similarity of faces

Face Matching Results, Relevance Feedback

- User may improve results by marking correctly found faces in several iterations:







Search-based Image Annotation

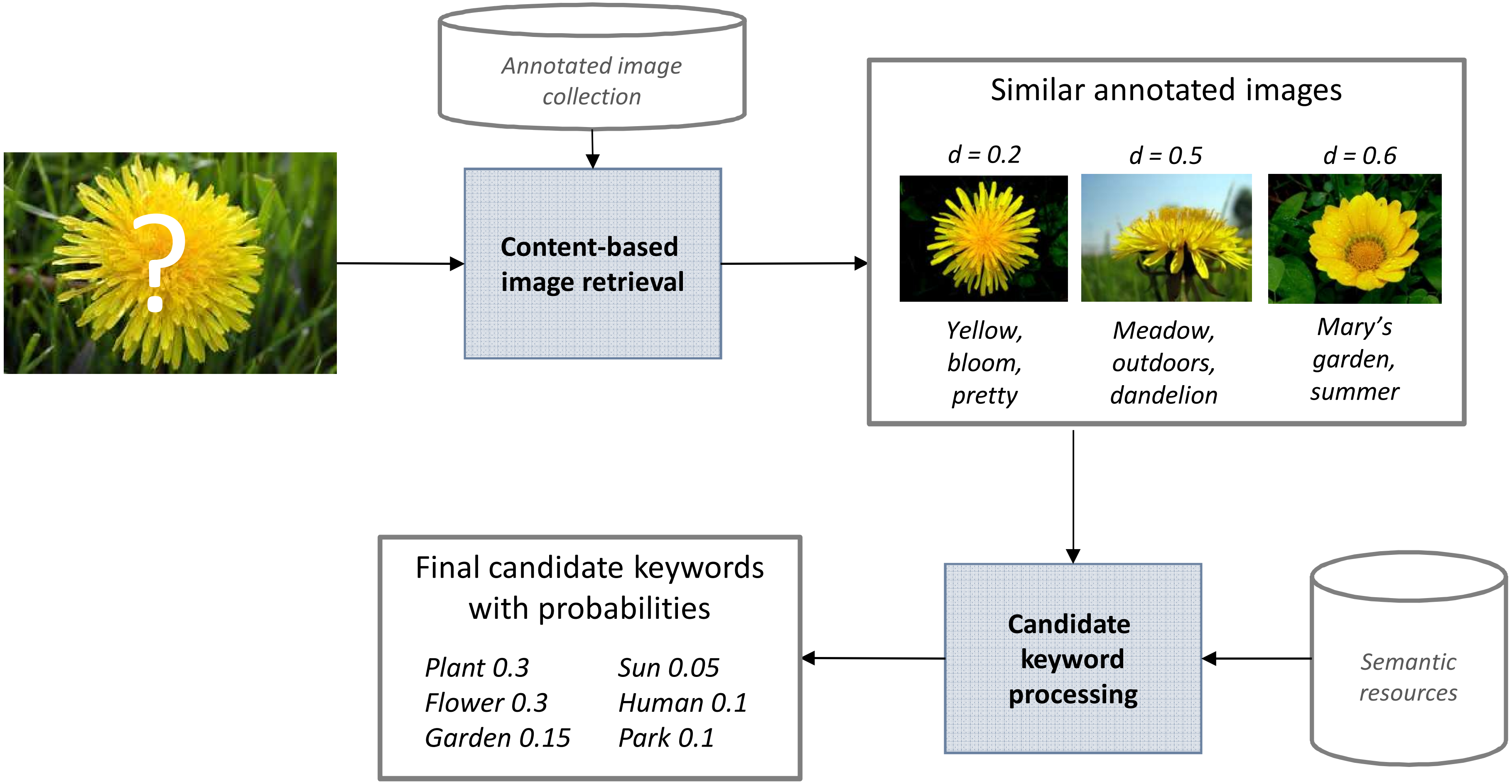
- Keyword-based image retrieval
  - Popular and intuitive
  - Needs pictures with text metadata
  - Manual annotation is expensive

Need for automatic image annotation

- We already have a strong tool – the similarity search
  - For any input image, we can retrieve visually similar images
  - Metadata of the similar images can be used to describe the original image

Search-based image annotation

Search-based annotation principles



Content-based retrieval for annotations

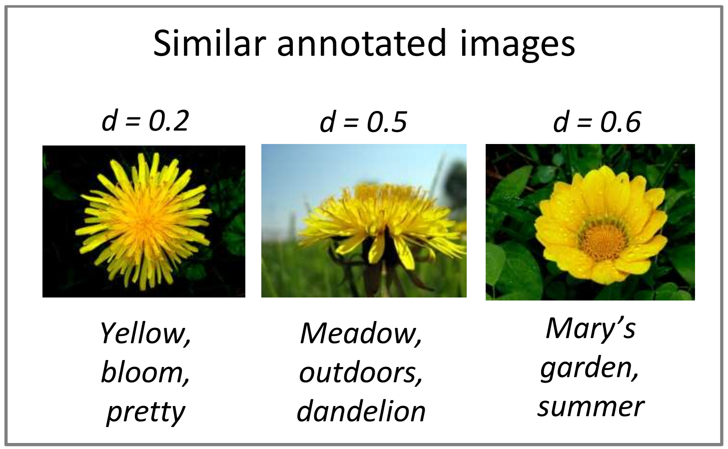
- What we need:

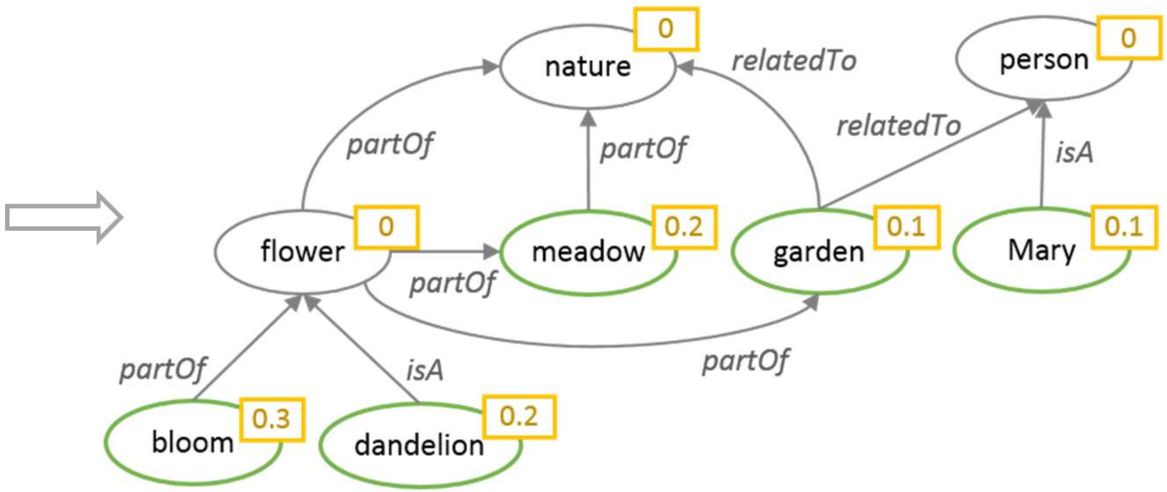
  - Large collection of reliably annotated images: Profiset
    - 20 million general-purpose photos from the Profimedia photostock company
    - Descriptive keywords for each photo provided by authors who want to sell the pictures → rich and reliable annotations
  - Efficient and effective search: DeCAF descriptors and PPP-codes
    - DeCAF: 4096-dimensional vector obtained from the last layer of a neural network image classifier
    - PPP-codes: effective permutation-based metric space indexing method

Profiset keywords: botany, close, closeup, color, daytime, detail, exterior, flower, germany, hepatica, horticulture, laughingstock, liverwort, lobed, mecklenburg, nature, nobilis, outdoor, outside, plant, pomerania, purple, round, western

ConceptRank

- Candidate keyword analysis inspired by Google PageRank
- Uses semantic connections between candidate keywords to determine the probability of individual candidates
- Main steps:
  - Construct a graph of candidate keywords related by WordNet semantic links
    - New candidates can be found during the WordNet exploration
  - Apply biased random walk with restarts to compute the score of each keyword
    - Keyword scores from the content-based search are included via the biased restart





Example

1. Retrieve 100 similar images from Profiset
2. Merge their keywords, compute frequencies
3. Build the semantic network using WordNet
4. Compute the ConceptRank
5. Apply postprocessing & return 20 most probable keywords

Candidate keywords after CBIR

church, architecture, travel, europe, building, religion, germany, buildings, north, churches, christianity, america, religious, exterior, st, historic, world, tourism, united, usa, …

Semantic network

4 relationships: hypernym *(dog → animal)*, hyponym *(animal → dog)*, meronym *(leaf → tree)*, holonym *(tree → leaf)*

270 network nodes, 471 edges

ConceptRank scores

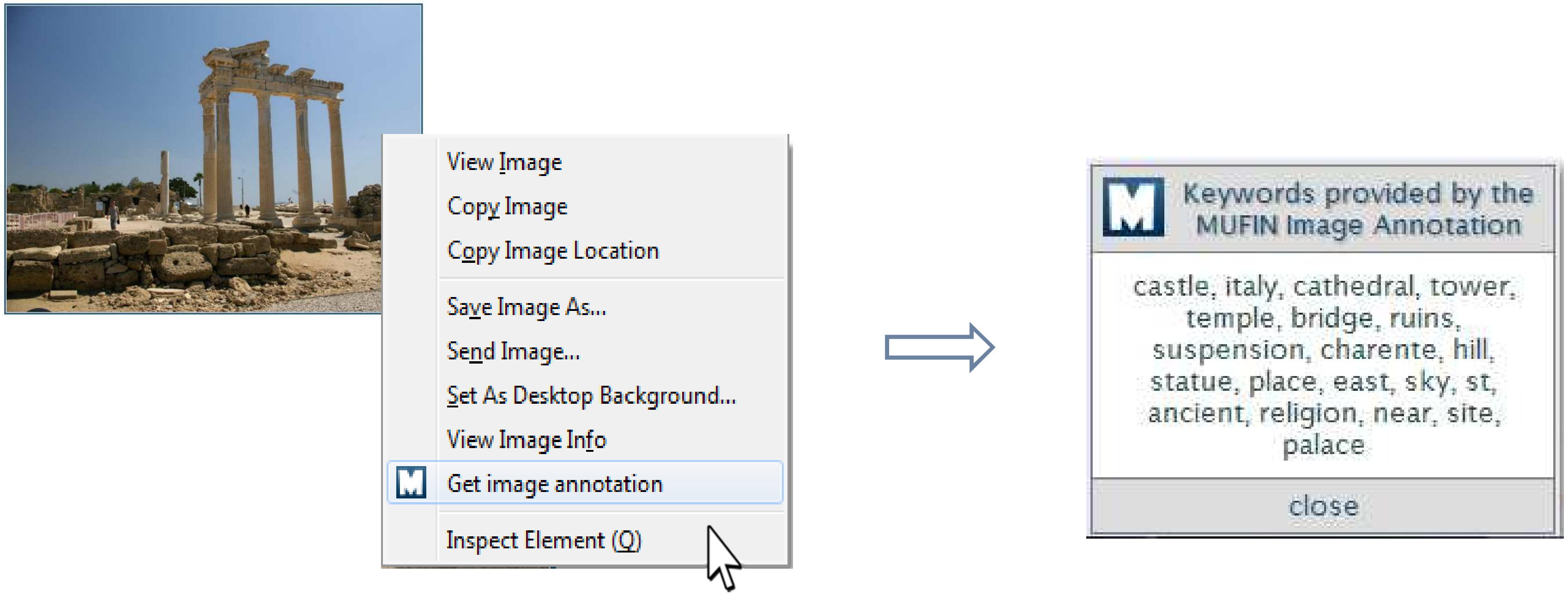
building (2.53), structure (2.41), LANDSCAPE (2.10), BUILDINGS (1.87), OBJECT (1.84), NATURE (1.78), place_of_worship (1.75), church (1.74), Europe (1.68), religion (1.64), continent (1.51), …

Final keywords

building, structure, church, religion, continent, group, travel, island, sky, architecture, tower, person, belief, locations, chapel, christianity, tourism, regions, country, district

Annotations in use

- Participation in the ImageCLEF 2014 Scalable Annotation Challenge
  - 2nd place, mean average precision of annotation approx. 60 %
- Web demo & Mozilla addon
  - http://disa.fi.muni.cz/prototype-applications/image-annotation



Similarity Search in Streams

- Two basic approaches to explore data:
  - Store, pre-process and search later, **database processing**
  - Process (filter) continuously, **stream processing**

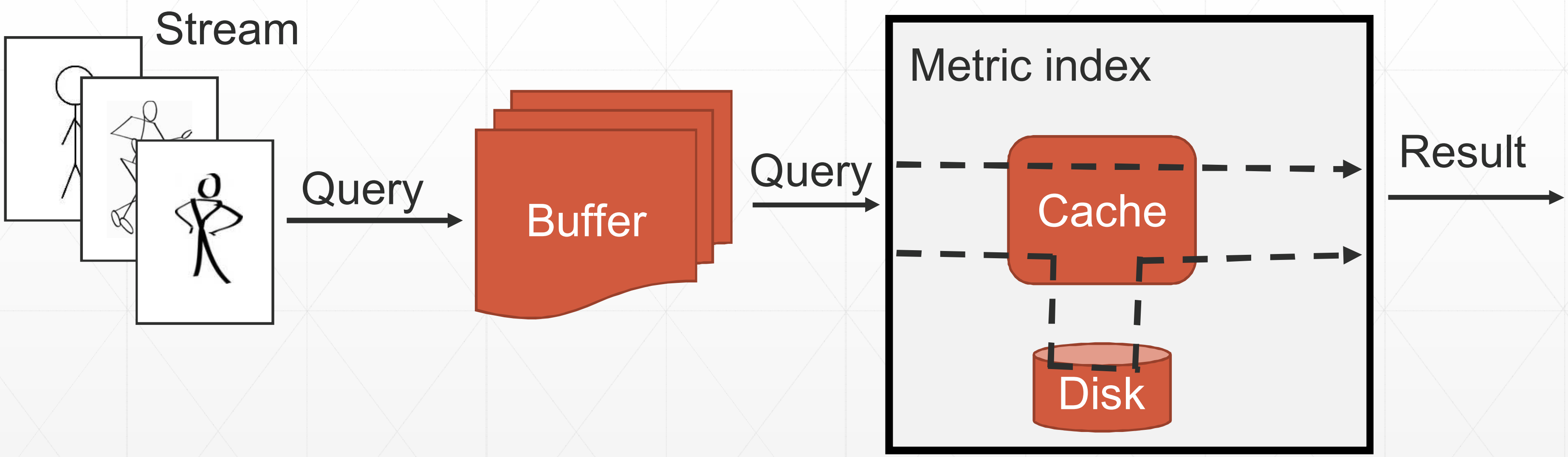
- Examples of stream processing applications:
  - Surveillance camera and event detection
  - Mail stream and spam filter
  - Publish/subscribe applications

Stream Processing Scenarios

- Stream: potentially infinite sequence of data items (d_1, d_2, …) – tuples, images, frames, etc.
- Basic scenarios:
  - Data items processed immediately, possible data item skipping → minimize delay - e.g., event detection
  - Process everything as fast as possible, delay possible to **maximize throughput** - our focus
- Motivating examples with similarity searching
  - Image annotation – annotate a stream of images collected by a web crawler
  - Publish/subscribe applications – categorize a stream of documents by similarity searching

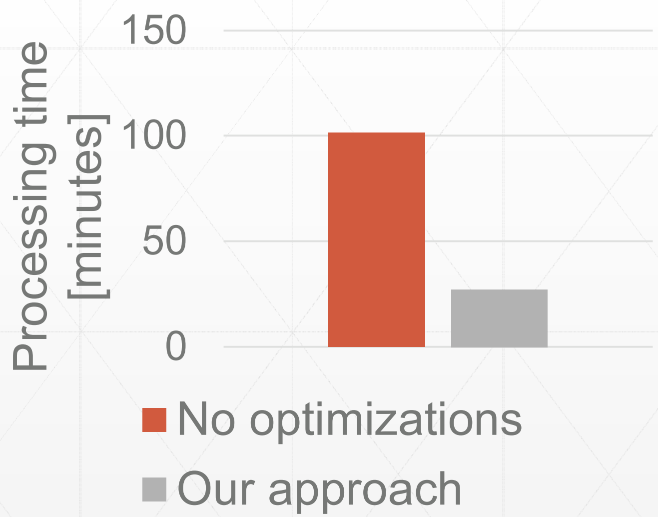
Processing Streams of Query Objects

- Typical large-scale similarity search approach:
  - partitioned data stored on a disk
  - partition reads from a disk form the bottleneck
- Idea: similar queries need similar sets of partitions → save accesses
- Buffer: memory used for reordering (clustering) queries
- Cache: memory containing previously read data partitions

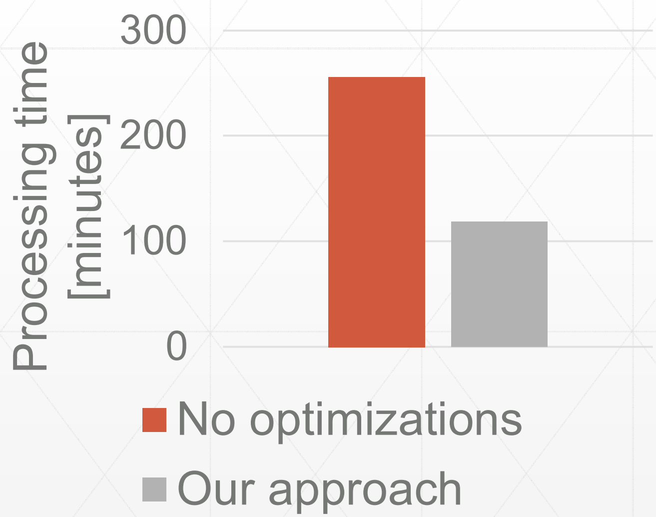


Experiment Results

- 100,000 processed 10-NN queries
- DB: 1 mil. MPEG-7 descriptors
- Buffer capacity: 8,000 queries
- Cache size: 40,000 objects (4% of the DB)

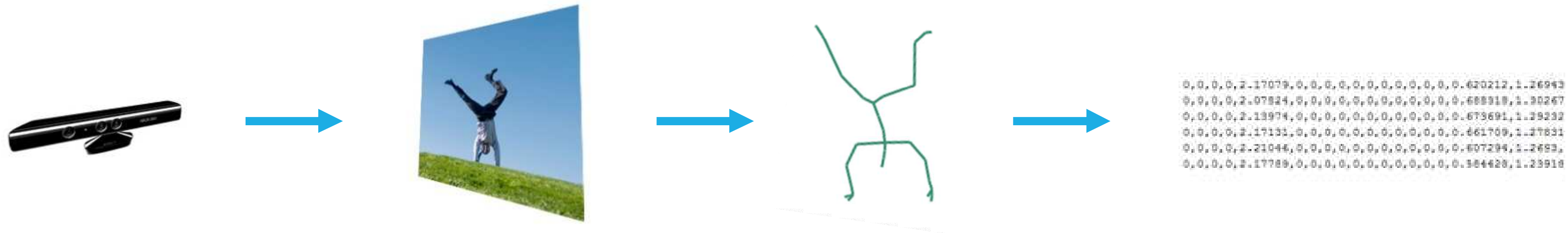


- 100,000 processed 10-NN queries
- DB: 10 mil. MPEG-7 descriptors
- Buffer capacity: 10,000 queries
- Cache size: 90,000 objects (0.9% of the DB)

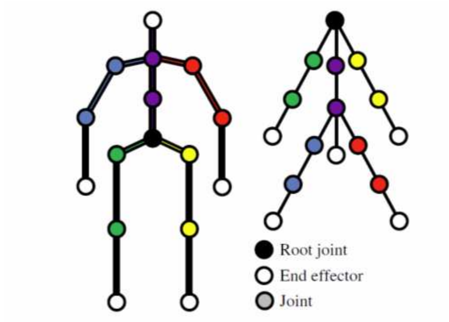


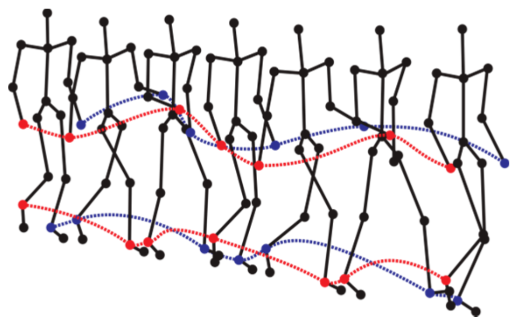
Similarity Search
in Motion Capture (Mocap) Data

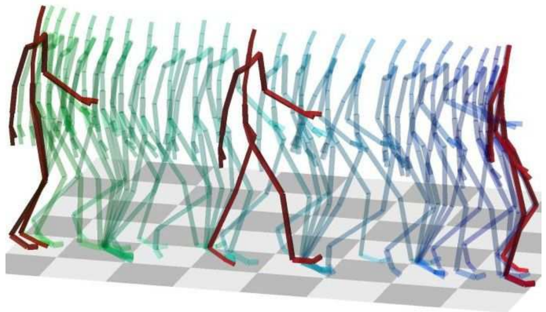
Digital representations of human motions, recorded by motion capturing devices for further use in a variety of applications.



What Is Mocap







- **Digital representation is** depicted by series of coordinates of body joints in space-time.
  - Complex multi-dimensional spatio-temporal data (3D space, 31 joints, 120 frames per second).
  - Visualized by simplified human skeleton (stick figure), coordinates of joints stored as float numbers.
  - **1 minute of such motion data ≈ 669,600 float numbers**.

The Need for a Similarity Measure

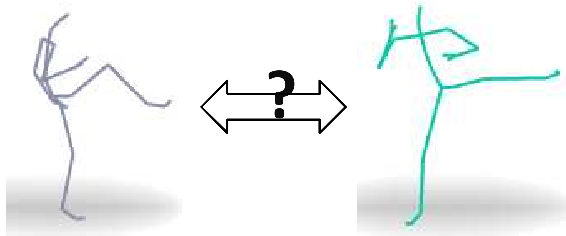
Almost every application of Mocap data

(analysis, searching, action recognition, detection, synthesis, clustering)

requires a pair-wise action comparison based on similarity.

The challenge:

- **Develop a measure** for content-based similarity comparison of Mocap data.



Motion Similarity Problems

The same action can be performed differently

- by different actors,
- in various styles,
- in various speed,
- or start at different body configurations.

Similarity of motions is application-dependent

- e.g., general action recognition vs.
  person-identification
- there is no universal similarity model

Comparing Similarity in Motions
General Overview

1. DATA REPRESENTATION

absolute coordinates, relative distances, joint rotation angles or velocities

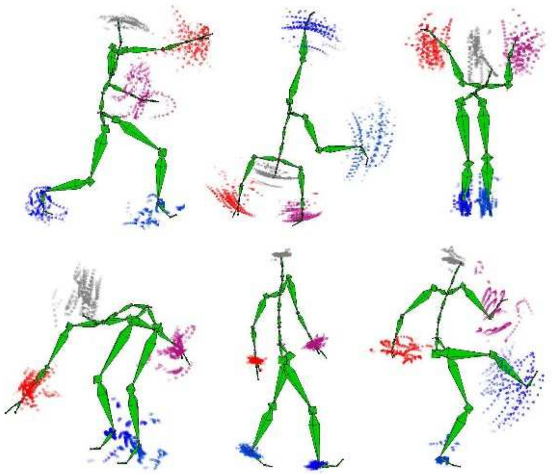
(quantization or dimensionality reduction might be applied)

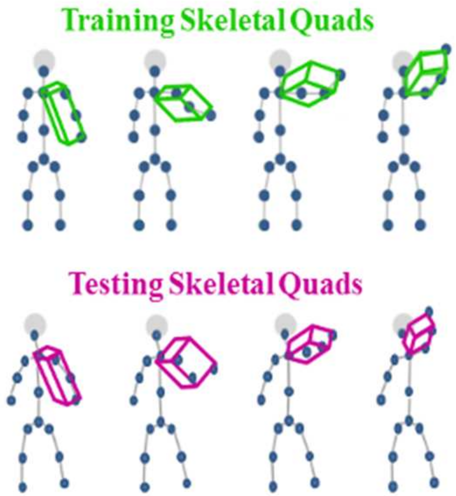
+

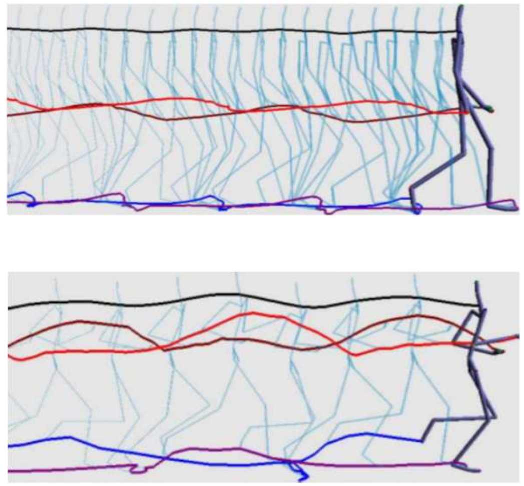
2. WAYS OF COMPARISON

Distance-Based functions	Machine Learning	Special structures
• Dynamic Time Warping • k-NN + L2	• Convolutional Neural Networks, Boltzmann M. • Support Vector Machines	• Motion and Action graphs • Temporal pyramids • Hidden Markov models

Comparing Similarity in Motions Examples





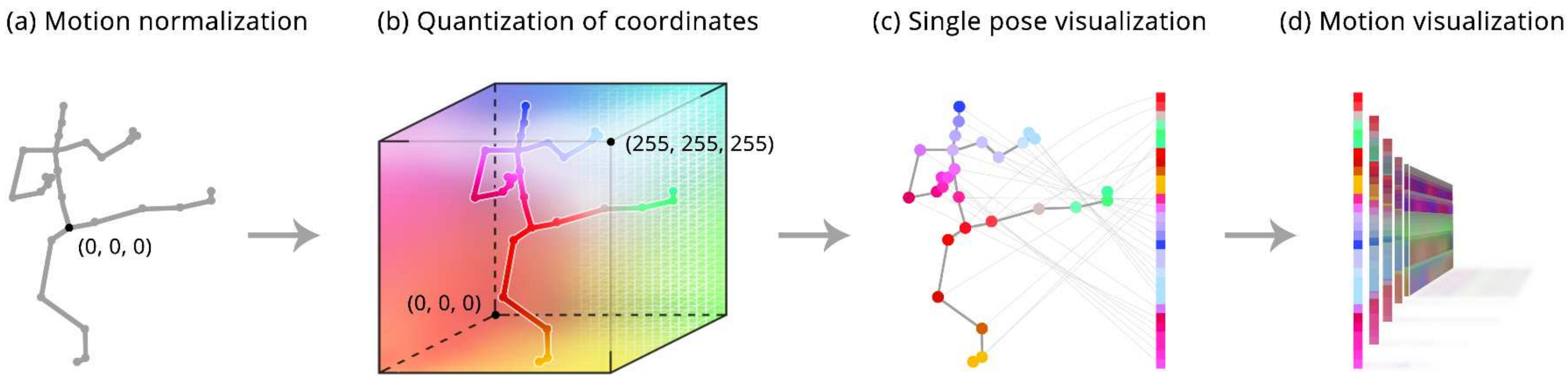


Joint positions features
+ Euclidean Distance
(Krüger 2010)

Fisher Vector
+ SVM classifier
(Evangelidis 2014)

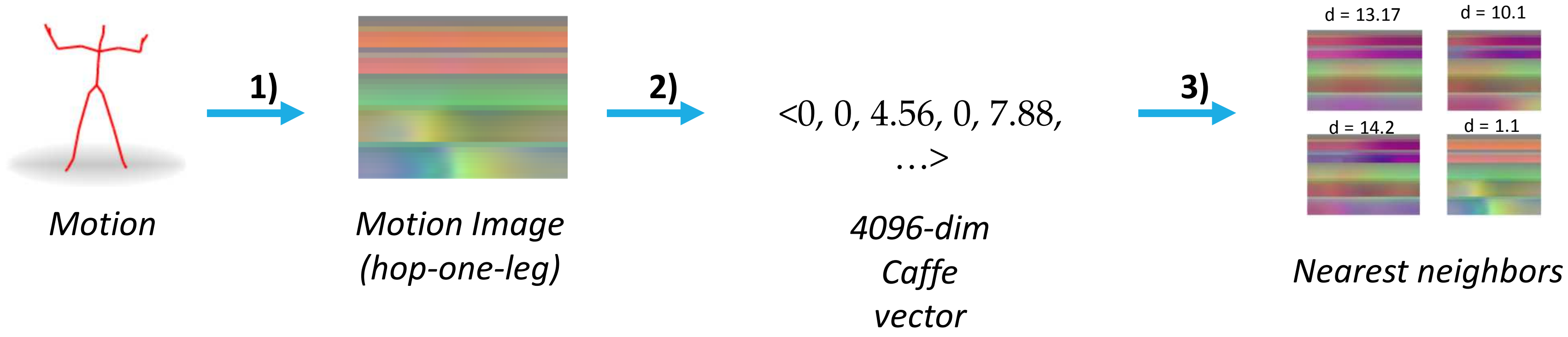
Time Series
+ Dynamic Time Warping
(Müller 2009)

Our Approach – Motion Images



Every single-frame joint configuration is normalized (by centering and rotating), then transformed into a RGB stripe image while fully preserving skeleton configuration.

Motion Images + Caffe



1) Effective transformation
from (dynamic) motion capture data into (static) images.

2) Extract fixed-size feature vector
using content-based image descriptors.

3) Index for fast and scalable search

Similarity model: Caffe + L_2

Motion Images

HDM05 dataset

2345 motions
120 fps, 31 joints
122 categories

$\xrightarrow{\text{transform}}$

Motion Images

1 motion = 1 image

Each image is rescaled to 256 x 256 px to fit as an input to the neural network.

$\xrightarrow{\text{extract}}$

4096-dimensional features

Caffe Descriptors

Extracted using Convolutional Neural Network trained on 1.3M photographs

Network can be fine-tuned to the domain of motion images

Output of 7th layer is a 4096-dimensional vector

Caffe descriptors are fixed sized vectors extracted from motion images. **They are compared for similarity by L_2 distance function.**

Motion Images – Properties

- **Pattern recognition is a mature concept nowadays**
  many highly accurate computer vision techniques might be employed.
- **The proposed similarity measure is robust and tolerant**
  towards inferior data quality, execution speed and imprecise segmentation.
- **Fixed-size feature vectors can be indexed in large scale**
  evaluate a query in one year long Mocap data in less than a second
- **Fixed-size feature vectors compress the original data**

Sizes Compared

- **5 seconds of mocap**
  - 3 x 31 x 120 x 5 = 57 600 floats
  - ≈ 460 KB
- **1 image 256 x 256 px in png format**
  - ≈ 5-10 KB
- **1 caffe descriptor**
  - 4096 floats ≈ 32 KB
  - 4096 bits ≈ 1 KB
- **1 mpeg7 descriptor**
  - 256 floats ≈ 2 KB

Laboratory of

Data Intensive Systems and Applications

disa.fi.muni.cz