Utrecht University

Multimedia Retrieval Algorithmics

Remco Veltkamp
Utrecht University
The Netherlands

Sofsem 2007, January 22-26, Harrachov, Czech Republic

What is multimedia anyway?

- "The medium is the message"
  "The personal and social consequences of any medium result from the new scale that is introduced by any new technology"
- Popular misquotation (or not) of:
  "The medium is the massage" book by Marshall MacLuhan

What is multimedia anyway?

"Multimedia? As far as I'm concerned, it's reading with the radio on."

Rory Bremner, British comedian

End of slide show, click to exit.

Outline

- Multimedia retrieval
- Perceptual issues
- Algorithmic issues
- Shape based music retrieval
- Indexing

Multimedia

Definition:
Any combination of two or more media, represented in a digital form, sufficiently well integrated to be presented via a single interface, or manipulated by a single computer program

Loosely:
Multiple media: images, video, sound, 3D scenes

Multimedia aspects

- Production, authoring
- Delivery
- Storage, database
- Throughput, QoS
- Retrieval

Multimedia retrieval

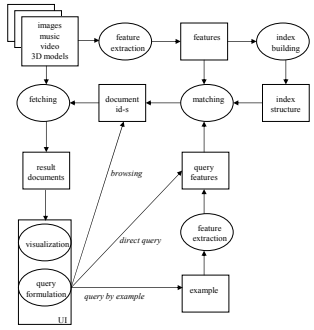
- Search, find, fetch, recover, restore, return
  $\Rightarrow$ getting back
- Traditional 'information retrieval': text
- MM retrieval: searching in large collections of images, video, sound, 3D scenes (the 5th wave in web searching)

Retrieval aspects

- Feature extraction
- Feature indexing
- Query formulation
- Feature matching
- Result visualization
- Feedback loop

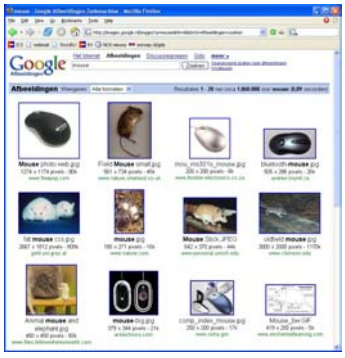
MM retrieval framework

- Media: images, music, video, 3D scenes
- Features: color, texture, shape
- Indexing: feature space, object space



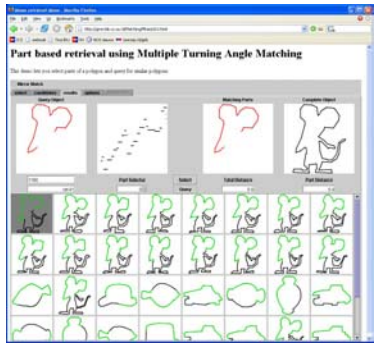
Current image search

- Based on file names, html tags, and surrounding text
- Ambiguous: synonyms and polysems



Content-based retrieval

- Based on analysis of images, music, video, 3D scenes



Applications

- Logo retrieval
- CAD searching
- Product catalogues
- Museum collections
- Photo archives
- Music selection
- Medical imaging
- Crime investigation, law enforcement
- Video searching
- Encyclopedia search
- Copyright protection

Example: Logo retrieval

- Services: search, watch
- Current practice: keyword based
- High level, but time consuming and error prone
- Keywords are Vienna classification codes

Vienna Classification Code

"castle":

category 7: constructions, structures for advertisement, gates or barriers

division 7.1: dwellings, buildings, advertisement hoardings or pillars, cages or kennels for animals

section: 7.1.1: castles, fortresses, crenellated walls, palaces

Example: Logo Retrieval

- Using Vienna classification code: up to 30.000 hits
- Visual inspection: 3000 per hour in morning 2000 per hour in afternoon
- $\Rightarrow$ automatic retrieval on the basis of shape and layout

Features: Color

Color signature:
count pixels of dominant colors

Features: Texture

Some pattern of color or intensity changes

Shape

Here: shape is geometry




Matching

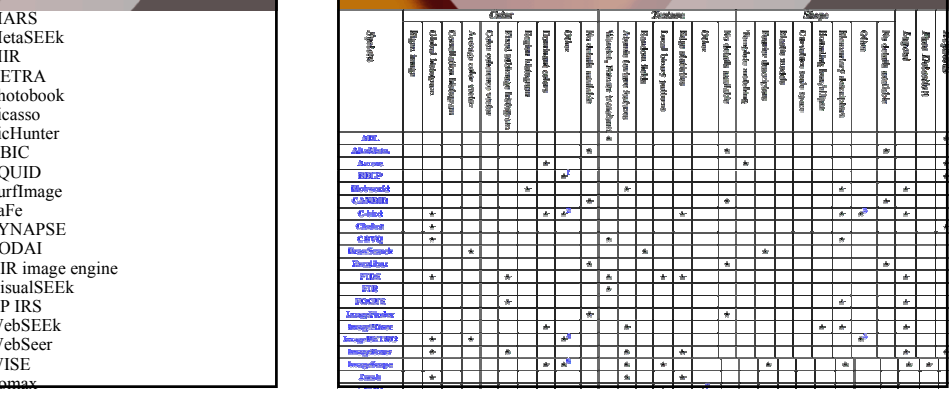
Given two images/objects/features A,B

- measure dissimilarity, distance *d(f(A),B)*
- using some distance function *d* (often called similarity rather than dissimilarity)
- under some transformation *f*

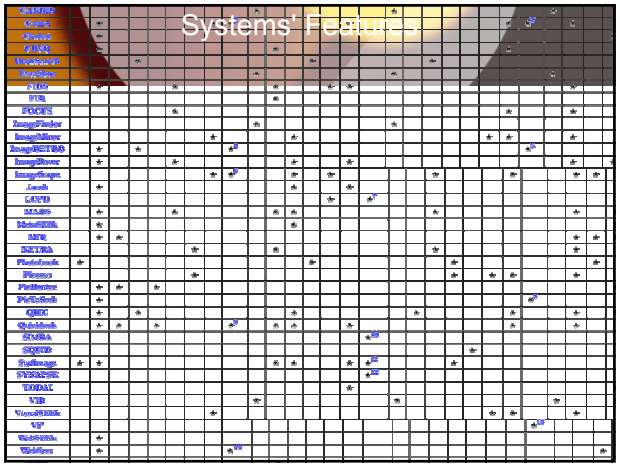
CBIR Systems

- ADL
- AltaVista Photofinder
- Amore
- Blobworld
- CANDID
- C-bird
- Chabot
- CBVQ
- Digital Library Project
- DrawSearch
- Excalibur
- FIR
- FOCUS
- ImageFinder
- ImageMiner
- ImageRETRO
- ImageRover
- ImageSearch
- Jacob
- LCPD
- MARS
- MetaSEEk
- MIR
- NETRA
- Photobook
- Picasso
- PicHunter
- QBIC
- SQUID
- SurfImage
- SaFe
- SYNAPSE
- TODAI
- VIR image engine
- VisualSEEk
- VP IRS
- WebSEEk
- WebSeer
- WISE
- Zomax

Systems' Features



Systems' Features



Systems' Features

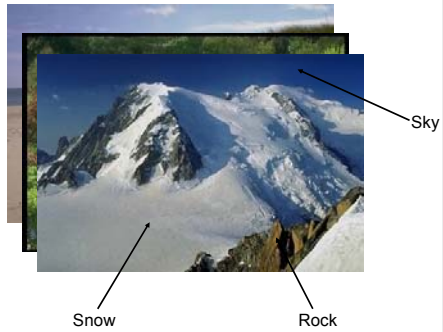
- 56 systems in the table
- 46 use any kind of color features
- 38 use texture
- 29 use shape
- 20 layout
- 5 use face detection.
- http://give-lab.cs.uu.nl/cbirsurvey

Level of Content

- Level 1: primitive features
  color, texture, shape, lay-out
- Level 2: objects, scenes
  table, mountain
- Level 3: abstract concepts
  dancing, democracy!

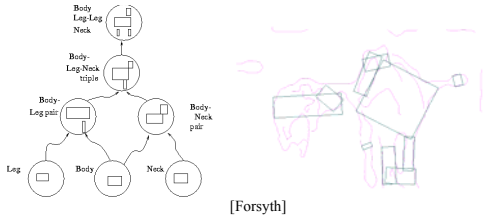
Level 2: scene classification

Classify scenes on the basis of material semantics



Level 2: object classification

Classify animals on the basis of body plans



[Forsyth]

Level 3: Abstract Concepts

- Manually semantic annotion of example documents
- Automatic transfer of semantic annotion, based on lower level features

Outline

- Multimedia retrieval
- Perceptual issues
- Algorithmic issues
- Shape based music retrieval
- Indexing

Optical truth ≠ perceptual truth

Human visual system favors ‘generic interpretation’ over nongeneric

- With dots: generic interpretation is blocked
- Without dots: generic ⇒ blue square

Gestalt Theory

- Initiated by Wertheimer (1923)
- "The whole is more than the sum of the parts"
- Goal: explain relation between patterns and their perceptual organization

Gestalt Principles/Laws

1 Figure and ground: elements are separated based on contrast

2 Similarity: similar elements seen as group

Gestalt Principles/Laws

3 Proximity/contiguity: elements close together seen as group

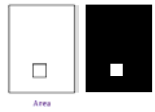
4 Continuity/continuation: preference for good continuation

5 Closure: tend to see complete figures

(Kanizsa triangle)

Gestalt Principles/Laws

6 Area: larger of two overlapping objects is seen as ground, smaller as figure



Gestalt Principles/Laws

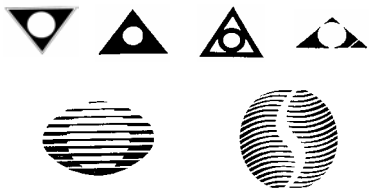
- No strenght ordering among different Gestalt laws: no unambiguous perceptual organization
- Koffka introduced law of Prägnanz: conveying the essence of something
- See a shape pattern as being as regular, simple, or symmetrical as possible
- Not part of the theory: measure for Prägnanz

Gestalt Principles/Laws

7 Prägnanz/Simplicity/Symmetry/Singularity: regions bound by symmetrical borders seen as coherent

Perceptual grouping

Identify which shape elements belong together, for example on the basis of Gestalt principles:



Perceptual grouping

original logo:

alternative human segmentations:

Perceptual matching

Two geometrical partial matches:

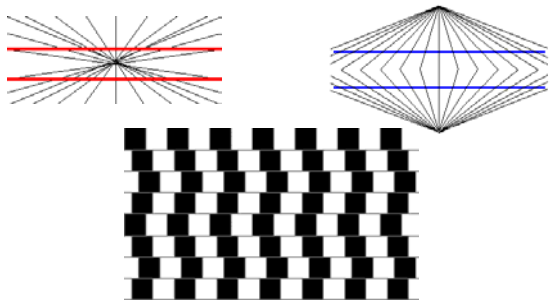
confusingly similar

not confusingly similar

Gestalt illusions

Hering illusion

Wundt illlusion



Perceptual features of sound events

Musical note:

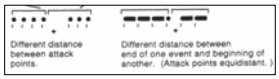
- Pitch
  - low-high: c. 90 categories
- Duration
  - long-short: multiples of 2 and 3
  - 'quantizing' into categories
- Loudness
  - soft-loud; non-categorical
- Timbre, tone quality
  - categorical? (voice and instrument recognition)

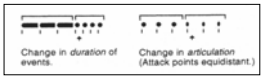
Perceptual grouping

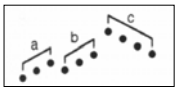
- Sound events are organised in groups
  - successive sounds form *melodies*
  - simultaneous sounds form *chords* or *harmonies*
- Music with one sound event at a time is called *monophonic*
- Music with more than one sound event at a time is called *polyphonic*
  - usually perceived as melody + chords
  - less frequent: 2 or more melodies

Gestalt principles

- Low level principles:
  - proximity
    - *rhythmic*
    - *pitch*
  - similarity
    - *duration*
    - *Timbre, articulation*
  - continuity
    - *melodic*
- These produce *closure* of wholes
- High-level principles
  - paralellism



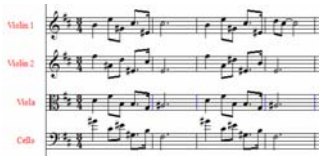




Melodic continuity

- In vision: a cross is interpreted as two straight lines

- In music, we tend not to hear crossings
  - instead, 'pitch proximity principle' dominates
  - overcome by timbral differentiation
- Example:
  Tchaikovsky 6th Symphony
  - first violin
  - second violin
  - together
  - whole orchestra



Outline

- Multimedia retrieval
- Perceptual issues
- Algorithmic issues
- Shape based music retrieval
- Indexing

Which algorithm?

Depends on

- which similarity measure, depends on
- which required properties, depends on
- which particular matching problem, depends on
- which application

Which problem?

- Computation problem: d(A,B)
- Decision problem: $d(A,B) \leq \varepsilon$?
- Decision problem: is there g: $d(g(A),B) \leq \varepsilon$?
- Optimization problem: find g: min d(g(A),B)
- Approximate optimization problem:
  find g: $d(g(A),B) < k\, d(g(A),B)$

Which properties?

- Metric properties
- Continuity
- Invariance

Metric Properties

- S set of patterns
- Metric: $d: S \times S \rightarrow R$ satisfying
  1. Self-identity: $\forall\, x \in S,\ d(x,x)=0$
  2. Positivity: $\forall\, x \neq y \in S,\ d(x,y)>0$
  3. Symmetry: $\forall\, x, y \in S,\ d(x,y)= d(y,x)$
  4. Triangle inequality: $\forall\, x, y, z \in S,\ d(x,z) \leq d(x,y)+d(y,z)$
- Semi-metric: 1, 2, 3
- Pseudo-metric: 1, 3, 4
- *S* with fixed metric *d* is called metric space

Symmetry

d(A,B) = d(B,A)

not always so for human perception

variant A:  prototype B: 

d(A,B) < d(B,A)

Triangle inequality

d(A,B)+d(B,C) ≥ d(A,C)

not always so for human perception,

in particular for partial matching:

A:  B:  C: 

Continuity

Robustness

Arbitrary small changes:

- deformation
- blurring
- cracks
- noise

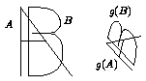
lead to arbitrary small change in similarity

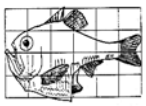
Invariance

d(g(A),g(B)) = d(A,B)

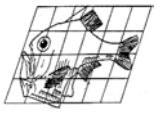
or d(g(A),B)=d(A,B)

for all g in transformation group G





Argyropelecus olfersi



Sternoptyx dialphana

(D'Arcy Thompson, 1911)

Invariance

Holbein: The ambassadors, 1533

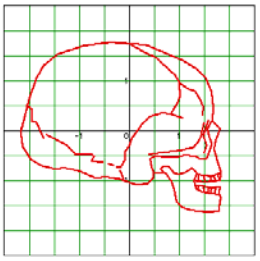
Invariance

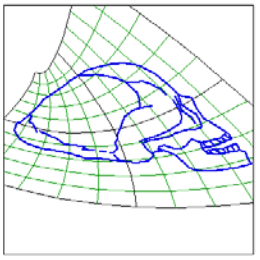
Holbein: The ambassadors, 1533

Projective transformation

Invariance?

for large enough transformation group ...

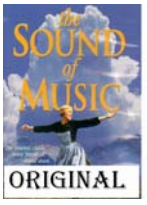


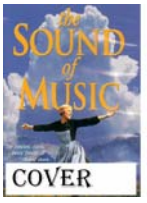


What is similarity?

What is similarity?

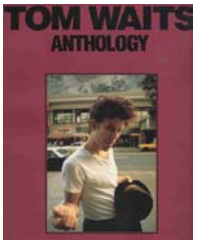
- "Cover": same melody but different performance



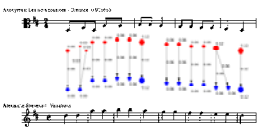


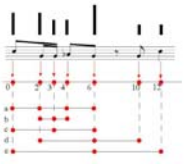
What is similarity?

- Same timbre?
- Same atmosphere?



Shape

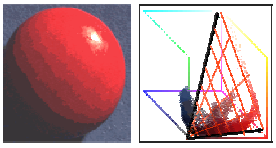




What makes shape?

Plato, "Meno", 380 BC:

"figure is the only existing thing that is found always following color"



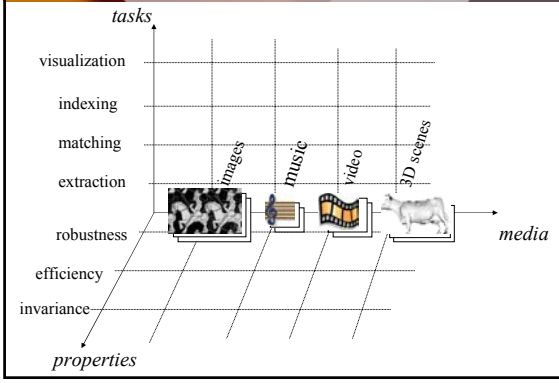
[Gevers]

What makes shape?

"terms employed in geometrical problems":

"figure is limit of solid"

Research space



Outline

- Multimedia retrieval
- Perceptual issues
- Algorithmic issues
- Shape based music retrieval
- Indexing

Which similarity?

Discrete metric:

$$d(A,B) = \begin{cases} 0 \text{ if A equals B} \\ 1 \text{ otherwise} \end{cases}$$

- Metric, invariant under all homeomorphisms!

Which similarity?

Discrete metric:

$$d(A,B) = \begin{cases} 0 \text{ if A equals B} \\ 1 \text{ otherwise} \end{cases}$$

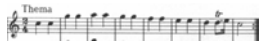
- Metric, invariant under all homeomorphisms!
- Exact congruence matching
- Lacks robustness properties

Problems of melody retrieval

- People remember high-level concepts, not notes
  - often confused with poor performance abilities
  - theme-intensive music (fugues) stimulate formation of such concepts
- Melodic variability and change (melodic confound): gradual shift of meaning
  - transposition
  - augmentation/diminution
  - ornamentation
  - variation
  - compositional processes: inversion, retrograde

One-dimensional melody retrieval

- Common assumption is (was?) pitch-only retrieval is sufficient
  - CCGGAAGGFFEEDDEC
  - wildcards



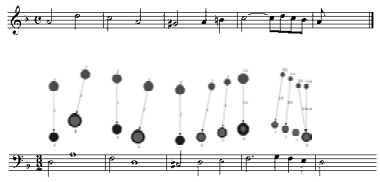
- Variants
  - interval (distance between 2 pitches)
  - pitch-contour
    - *repeat/up/down (Parson's Code)*
    - *RURURDRDRDRDRUD*
- String matching

Limitations

- Pitch contains only c. 50% of musical information
  - rhythm: 40% , timbre + loudness 10%
  - massive improvement expected from including rhythm
- No polyphonic retrieval
  - especially harmony (chord progressions) is important
  - state of the art polyphonic matching: OMRAS project
- No higher level musical concepts
  - e.g. melodic contour vs. ornamentation
  - input from music cognition and perception

Melody representation

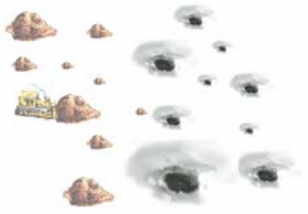
- Represent notes as weighted point sets in 2-dimensional space (pitch, time)
- Weight represents duration
  - other possibilities contour/metric position etc
- Interesting properties
  - tolerant against melodic confounds
  - suitable for polyphony
  - partial matching



after alignment, the weight is moved both along the temporal axis and along the pitch axis

Earth Mover's Distance

- The Earth Mover's Distance (EMD) calculates the *minimum flow* that would match two set of weighted points. One set emits weight, the other one receives weight
- Constraints:
  - no negative flow
  - no point emits or receives more than its weight
  - the lighter pointset is completely matched



Earth Mover's Distance

- $A=\{(x_i,w_i)\}$, $\sum w_i=W$, $B=\{(y_j,u_j)\}$, $\sum u_j=U$
- f_{ij} flow from x_i to y_j over d_{ij}
- $f_{ij} \geq 0$
- $\sum_j f_{ij} \leq w_i$
- $\sum_i f_{ij} \leq u_j$
- $\sum_i \sum_j f_{ij} = \min(W,U)$

$$EMD(A,B) = \frac{\min_F \sum_{ij} f_{ij} d_{ij}}{\min(W,U)}$$

Properties EMD

- Invariant under rigid motion
- Respects scaling
- Metric if d metric, and $W=U$
- If $W \neq U$:
  - No positivity, surplus not taken into account
  - No triangle inequality

Proportional Transportation Dist

- $A=\{(x_i,w_i)\}$, $\sum w_i=W$, $B=\{(y_j,u_j)\}$, $\sum u_j=U$
- f_{ij} flow from x_i to y_j over d_{ij}
- $f_{ij} \geq 0$
- $\sum_j f_{ij} = w_i$
- $\sum_i f_{ij} \leq u_j W/U$
- $\sum_i \sum_j f_{ij} = W$

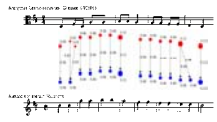
$$PTD(A,B) = \frac{\min_F \sum_{ij} f_{ij} d_{ij}}{W}$$

Properties PTD

- Invariant under rigid motion
- Respects scaling
- PTD is pseudo-metric:
  - Triangle inequality holds
  - No positivity
    - *but only when same relative weights*
    - *surplus taken into account*

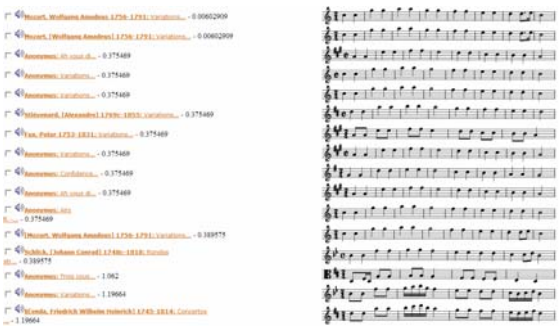
Application

Musical Search Engine



Example use

Ah, vous dirai-je maman/Twinkle twinkle little star/Altijd is Kortjakje ziek

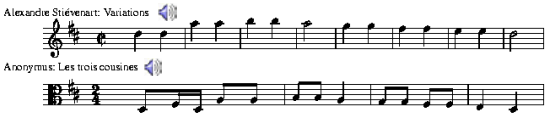


Outline

- Multimedia retrieval
- Perceptual issues
- Algorithmic issues
- Shape based music retrieval
- Indexing

Imagine

- 480,000 musical theme notations in RISM
- 80,000 labeled anonymous

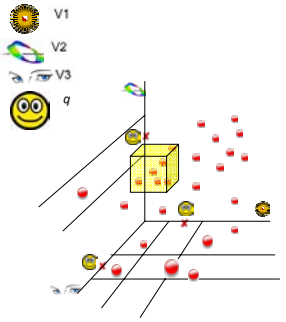


Ah! Vous dirai-je Maman/Altijd is Kortjakje ziek/Twinkle twinkle little star

- De-anonymization: 32,000,000,000 comparisons
- 1 ms per comparison: 370 days
- We identified 17,895 anonymous pieces, (*k* expensive + *O(logN)* cheap, i.s.o. *O(N)* expensive comparisons)

Vantage Indexing

- [Vleugels, Veltkamp; Pattern Recognition 2002]
- Select *k* vantage objects from the dataset
- Embed database in vantage space
  - compute distances from all objects to *k* vantage objects
- Query with *q*
  - position *q* in vantage space
- Take range ε around *q*
- Candidate matches: return intersection: *p candidate matches*
- Compute exact distance between *q* and the *p* candidate matches: get rid of false positives (optional)



Vantage Indexing

- Properties
  - Online only *k (+p)* distance calculations
    - *So complex distance measures possible*
  - No false negatives
    - *If triangle inequality holds for distance measure*

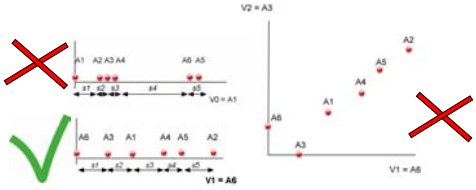
Selecting vantage objects:
Vantage object quality

- No false negatives possible
- But still false positives:
  - $d_{o\text{-}space}(A_q, A_{fp}) \geq \varepsilon$ *however*
  - $d_{v\text{-}space}(A_q, A_{fp}) \leq \varepsilon$

  - But we don't know what A_q nor ε will be…
    - *Obtain good performance averaged over all possible queries (A) over all possible ε*
  - How?
    - *Strive for small return sets (averaged over A) given fixed ε*

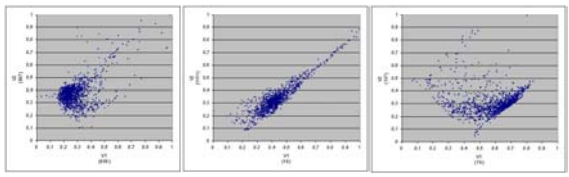
Selecting vantage objects:
Two criteria

- Given: Database A = {A1…A6} and d: A x A → R
- Goal: spread out A over the v-space as much as possible
  - Individual vantage object: distances as uniformly distributed as possible
    - *Interpretation: Low Variance of Spacing*
  - Combined vantage objects: as different vantage objects as possible
    - *Interpretation: Weak correlation coefficient*



Selecting vantage objects:
Some real examples

- Dataset : MPEG-7 CE-Shape 1 Part B
- Distance Measure: Curvature Scale Space [Mokhtarian et al.]



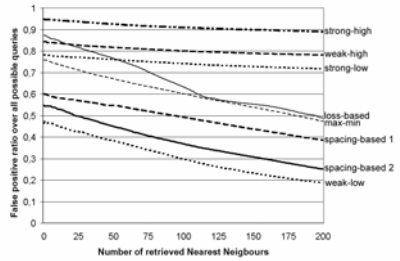
Selecting Vantage Objects:
Now we have the criteria, how to select them?

- Random incremental construction:
  - Create index by adding database objects one by one in random order
  - While doing so keep an eye on the index' spacing and correlation properties
  - Fix index where necessary

Algorithm 1 Spacing-based Selection
Input: Database A with objects $A_1, ..., A_n$, $d(A, A) \rightarrow \mathbb{R}$, thresholds ϵ_{corr} and ϵ_{spac}
Output: Vantage Index with Vantage objects $V_1, V_2, ..., V_m$

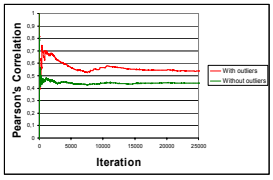
```
1: select initial V_1, V_2, ..., V_m randomly
2: for All objects A_i do in random order
3:     for All objects V_j do
4:         compute d(A_i, V_j)
5:         add A_i to index
6:         if var(Spacing(V_j)) > ε_spac then
7:             remove V_j
8:             select new vantage object randomly
9:     if for any pair p(V_k, V_l), Corr(V_k, V_l) > ε_corr then
10:        remove p's worst spaced object
11:        select new vantage object randomly
```

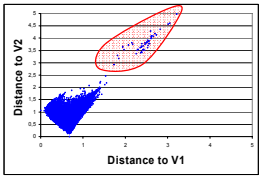
Selecting vantage objects:
Results: false positive ratios



Outliers

- Pearson's correlation is very sensitive to outliers
- Risk: throw away a possible good pair
- Solutions:
  - Ignore outliers
    E.g. detect sharp increase of correlation
  - Use other coefficients
    E.g. rank correlations like Spearman's or Kendall's



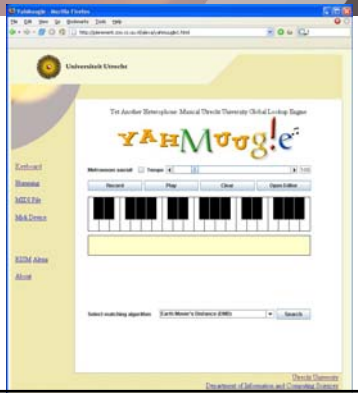


Ongoing

- Optimal vantage space dimensionality
  - More vantage objects: less false positives but longer querying times
- Vantage Indexing with partial matching
  - Partial matching: no triangle inequality guaranteed, thus false negatives possible
  - Possible solution: weak triangle inequality

$$d(A,B)+d(B,C) \geq k\, d(A,C)$$

Putting it all together



Where are we now?

- Computer is still stupid in seeing and listening
- The language/picture barrier: semantic annotation
- Invariant features and similarity must be designed
- Perceptually relevant features and similarity must be designed
- Interaction needs are high
- Compute power and algorithmic efficiency are necessary
- Multimedia integrated approach

Combining image and music retrieval

Holbein, The Ambassadors, 1533

Johann Walther: *Geistlich Gesangbuchli* (1525)

(Musical) content adds to meaning

Scientific Future

1. Scalability
2. Multimodal (text, picture, speech etc.)
3. Invariance and perception
4. Feedback and learning
5. Benchmarking

Acknowledgment

- Martijn Bosma, Panos Giannopoulos, Herman Haverkort, Reinier van Leuken, Mirela Tanase, Rainer Typke, Frans Wiering
- FP6 IST 511572 PROFI

PROFI

- FP6 IST 506766 AIM@SHAPE

AIM@SHAPE
FP6 IST NoE 506766
www.aim-at-shape.net