3.4 Video Representation

Representation and Classification

How can we view an hour of video in just a few minutes and still have a correct perception of its content?

- Video content abstraction—the process to extract a presentation of visual information including salient features, typical style and all major subjects
  - Video icon construction—construction of a statical icon representing a video shot
  - Keyframe extraction—keyframe is the frame which represents the salient content of the shot

Video Icons

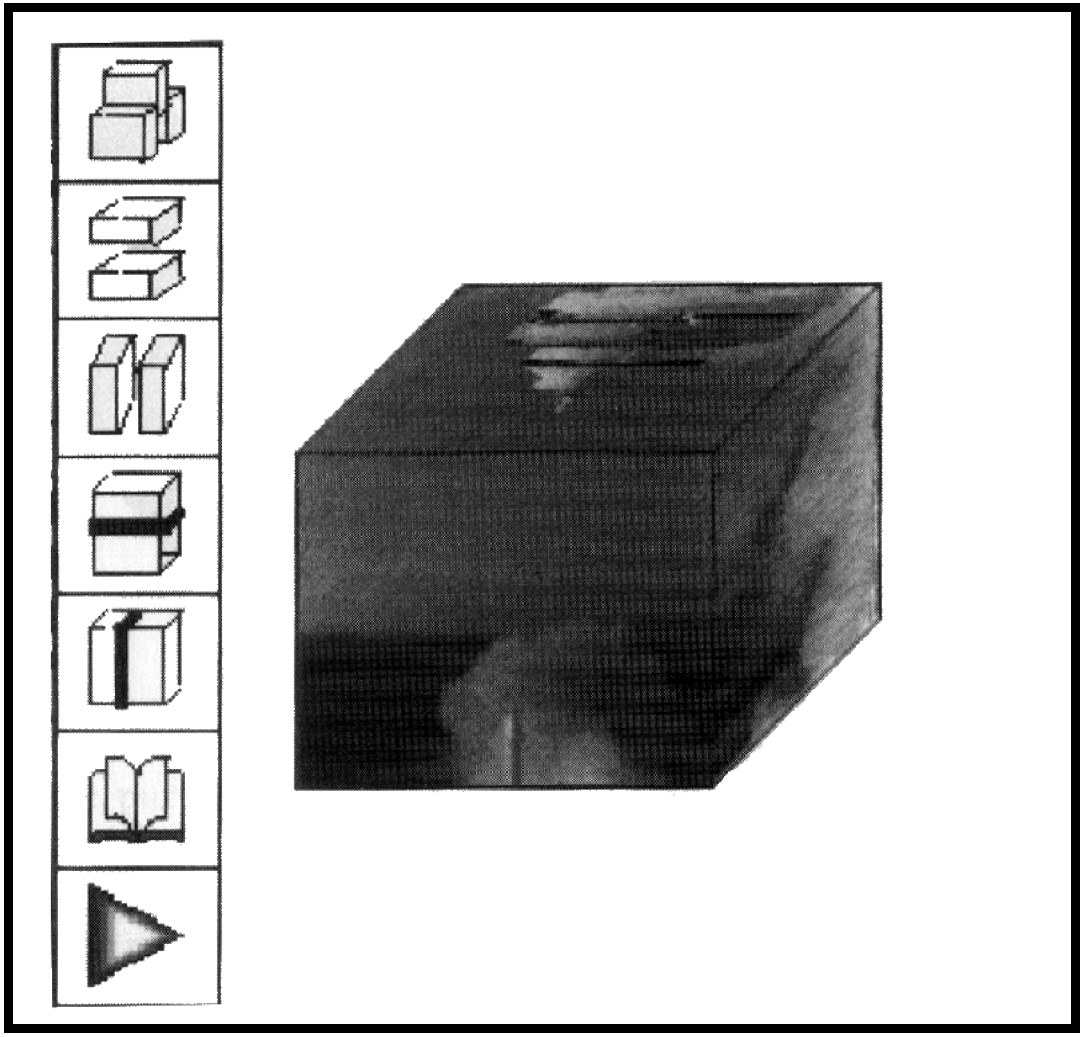
Several approaches—

- Icon based on a frame extracted from the shot, with pseudo-depth for representation of duration, arrows and signs for representation of object and camera motion
- Synthesis of an image representing the global visual contents of the shot—
  - salient stills
  - videospace icon
  - videomap

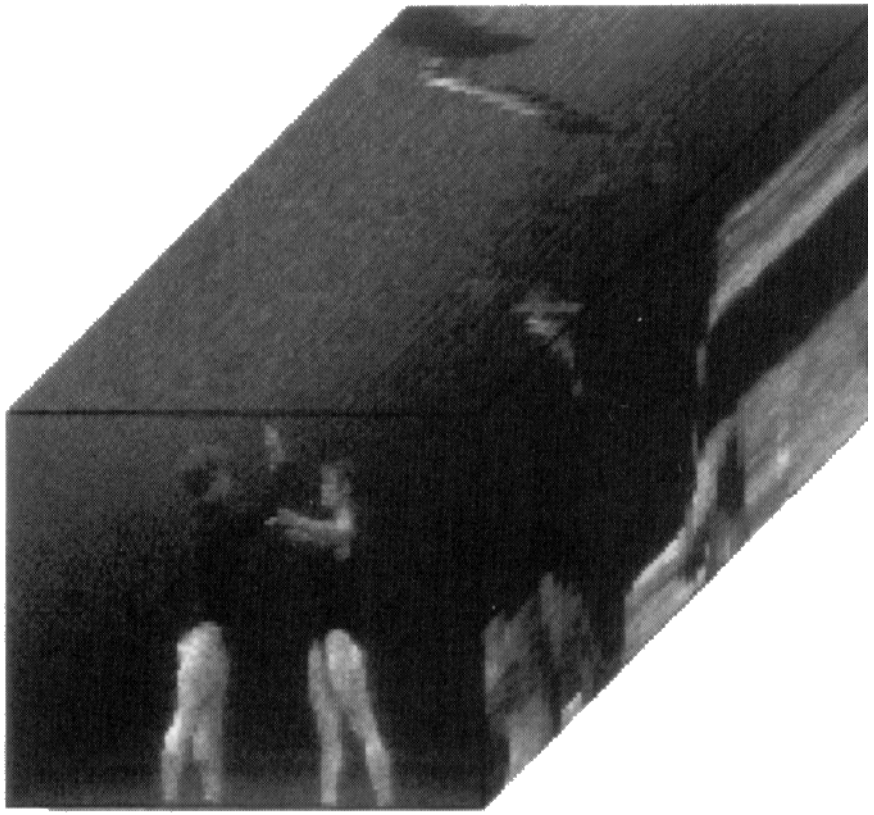
Interactive Tools

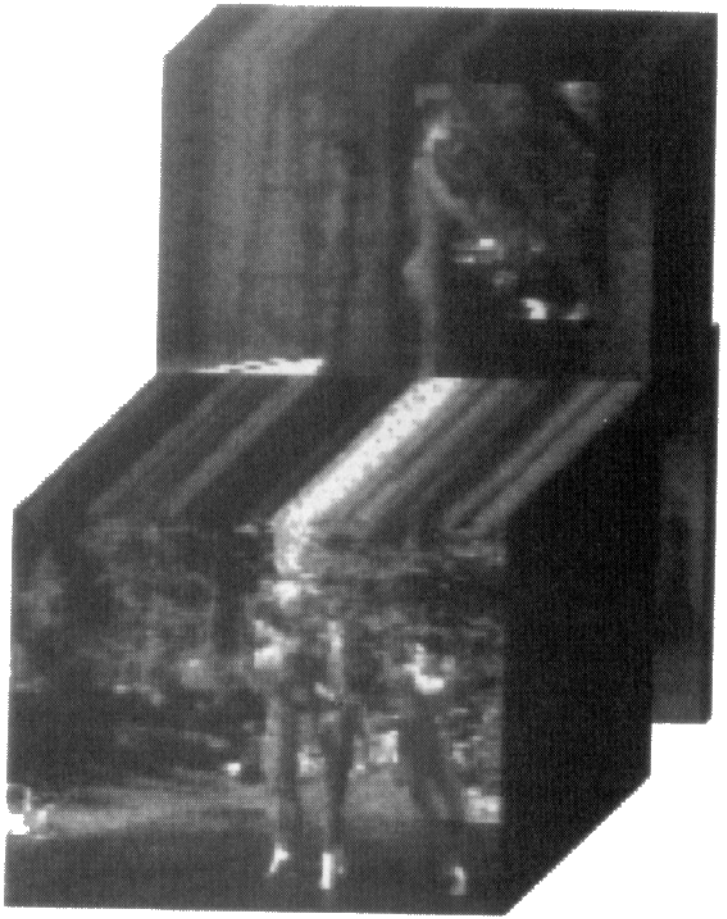
- *movie icon (micon)*—video is represented by a 3d volume
- *interactive micons*—examination and manipulation environment
- *paper-video*—”chart-based” video browser
- *video panorama*—indication of the video space captured
- *videoscope*—video content analyzer
- *sound browser*—detecting the presence of music

Interactive Video Icons

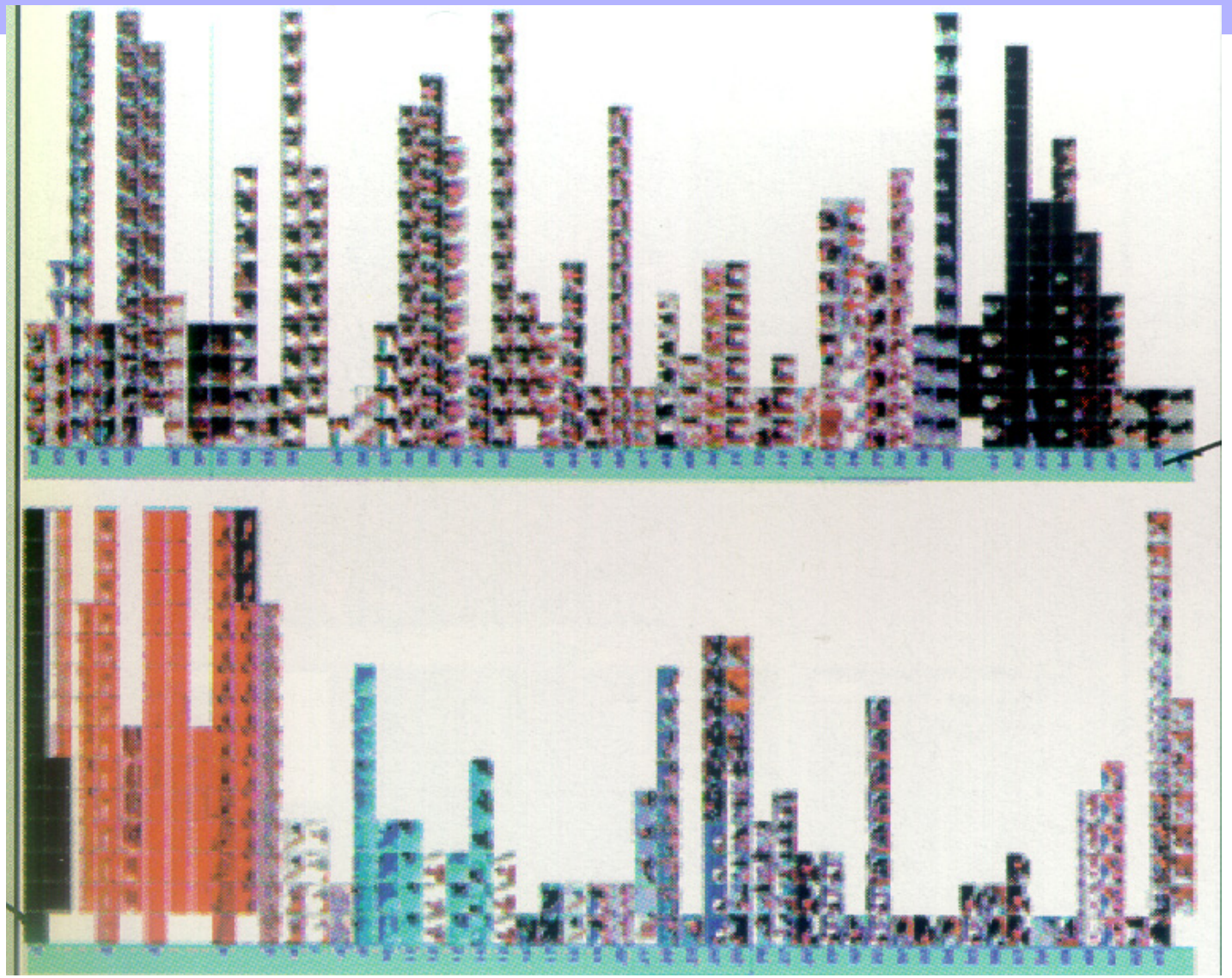


Interactive Video Icons



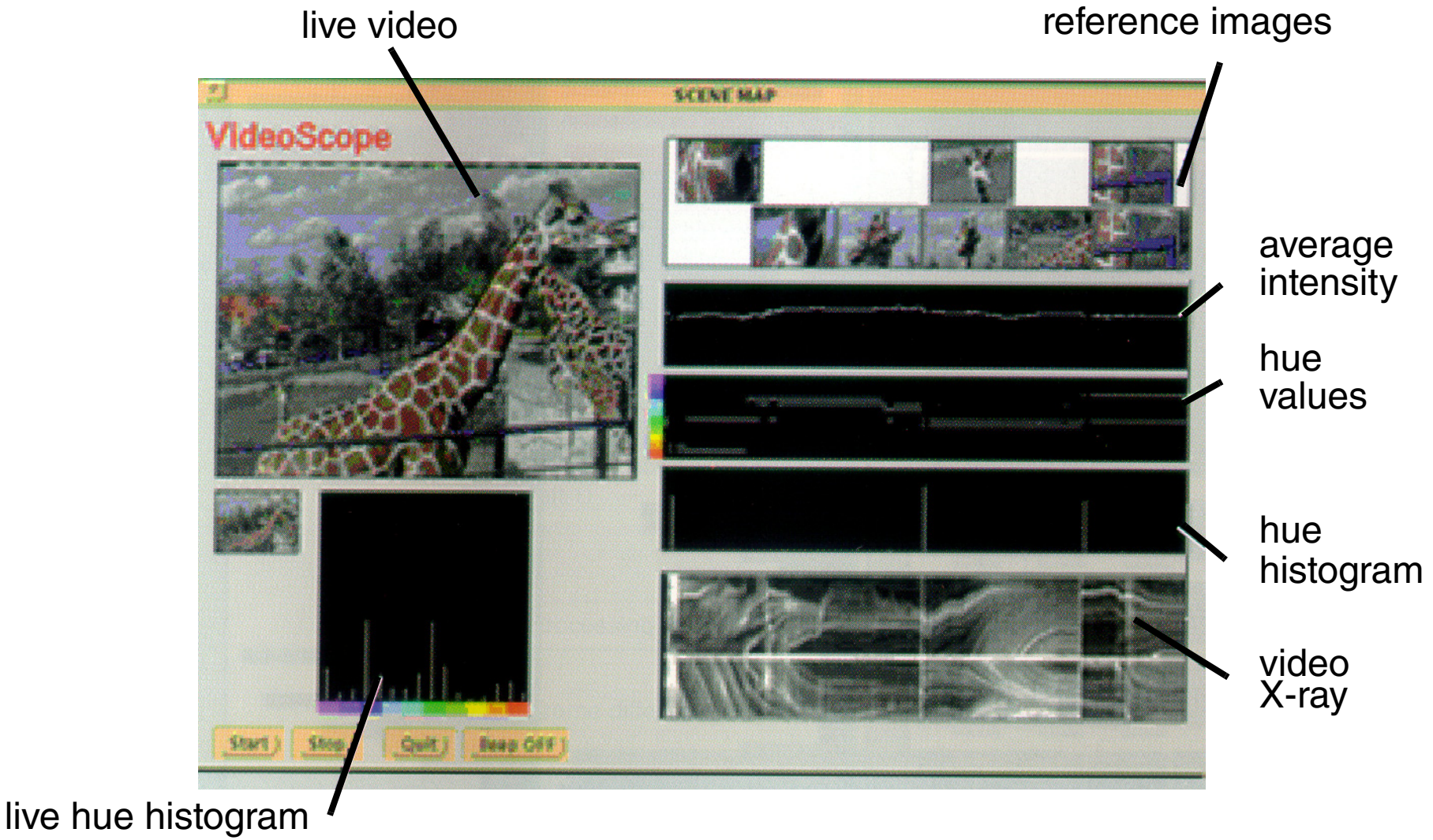


Paper Video



Video Panorama

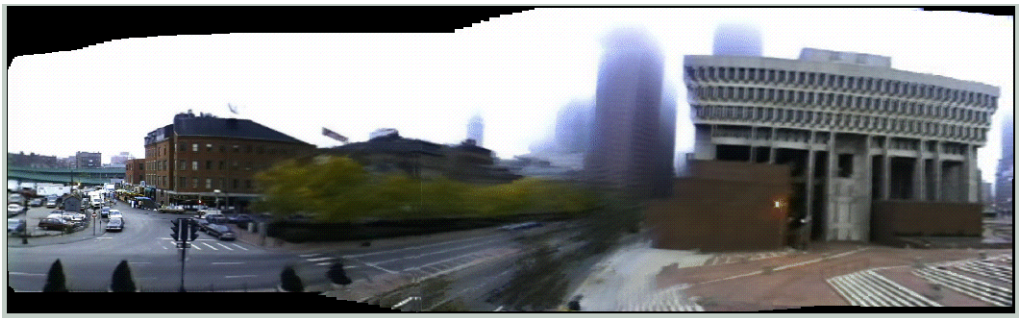
VideoScope


live video
reference images
VideoScope
average intensity
hue values
hue histogram
video X-ray
live hue histogram

Sound Browser

Salient Stills

salient still of a panning scene



start frame

end frame

salient still storyboard

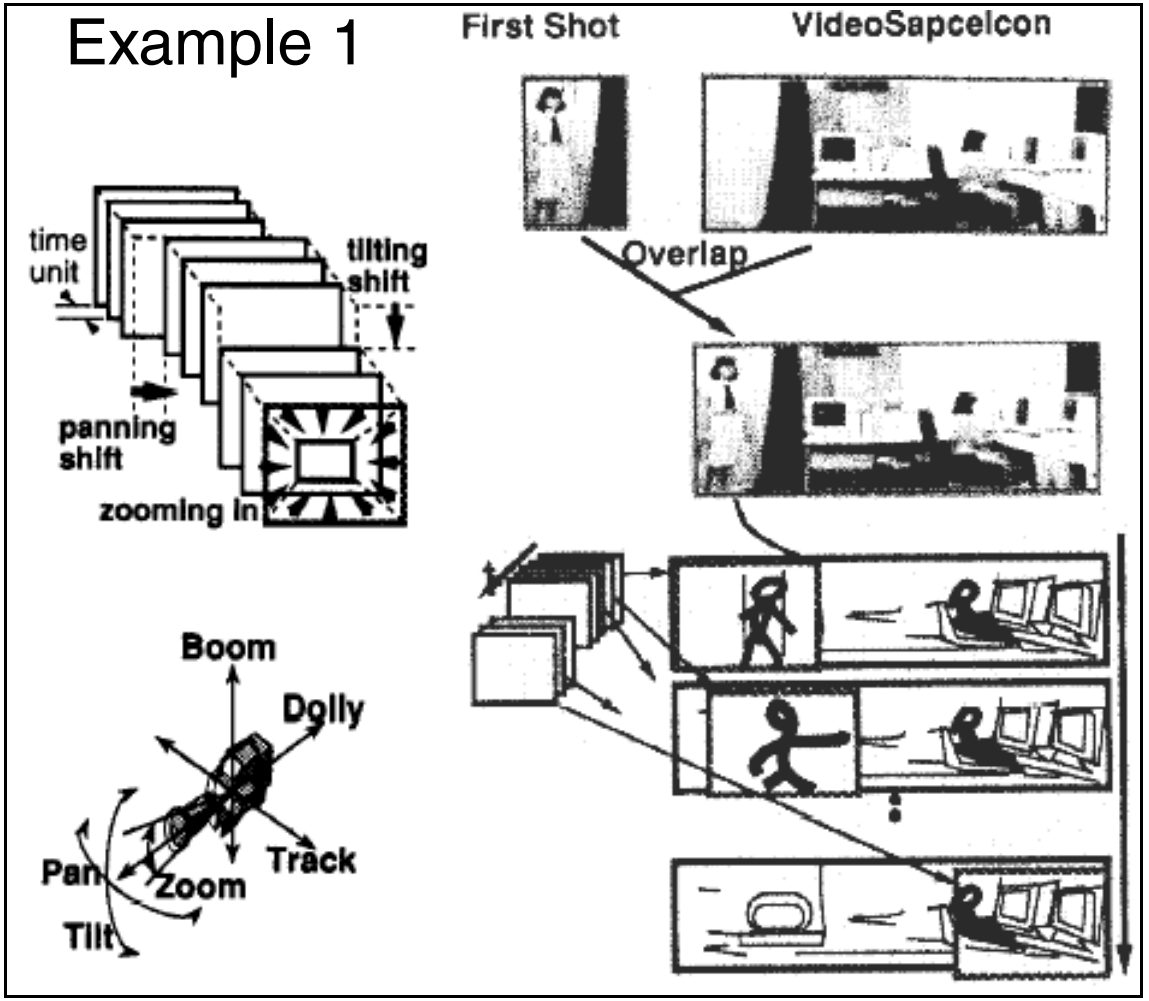
salient still

salient still "chrono-photo-graph"

Video Space Icon

Scene-girl comes in, went to her computer; camera follows her

Example 1

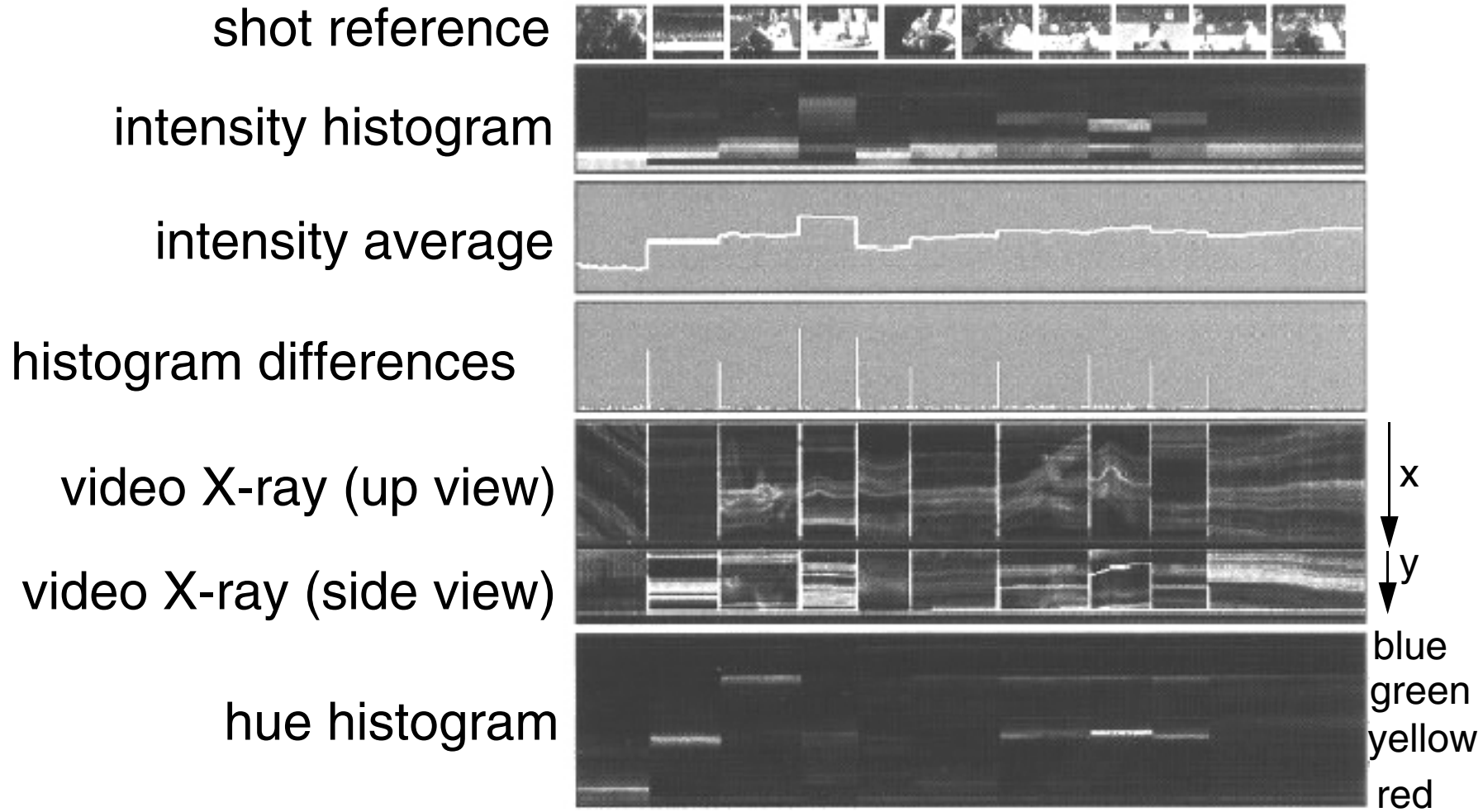


Example 2

front view

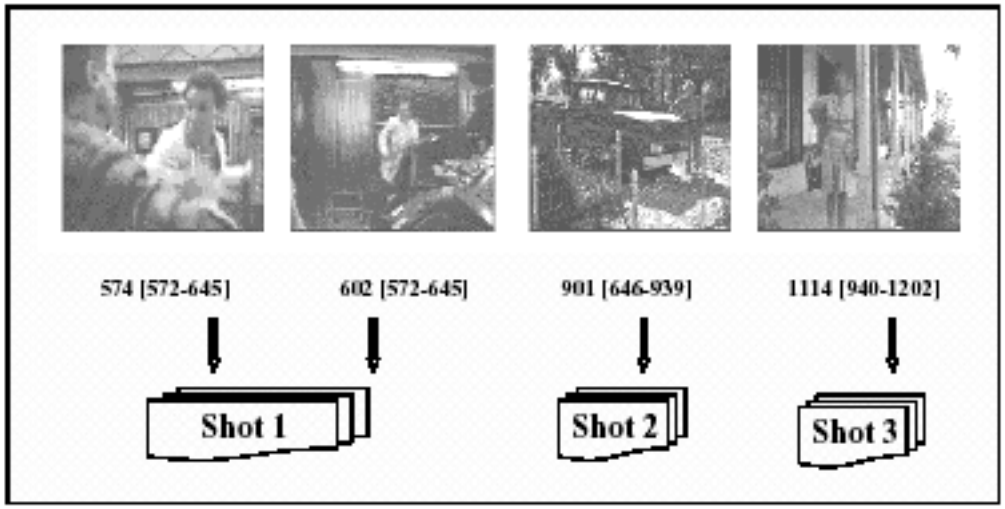
top-front view

Videomap


shot reference
intensity histogram
intensity average
histogram differences
video X-ray (up view)
video X-ray (side view)
hue histogram
x
y
blue
green
yellow
red

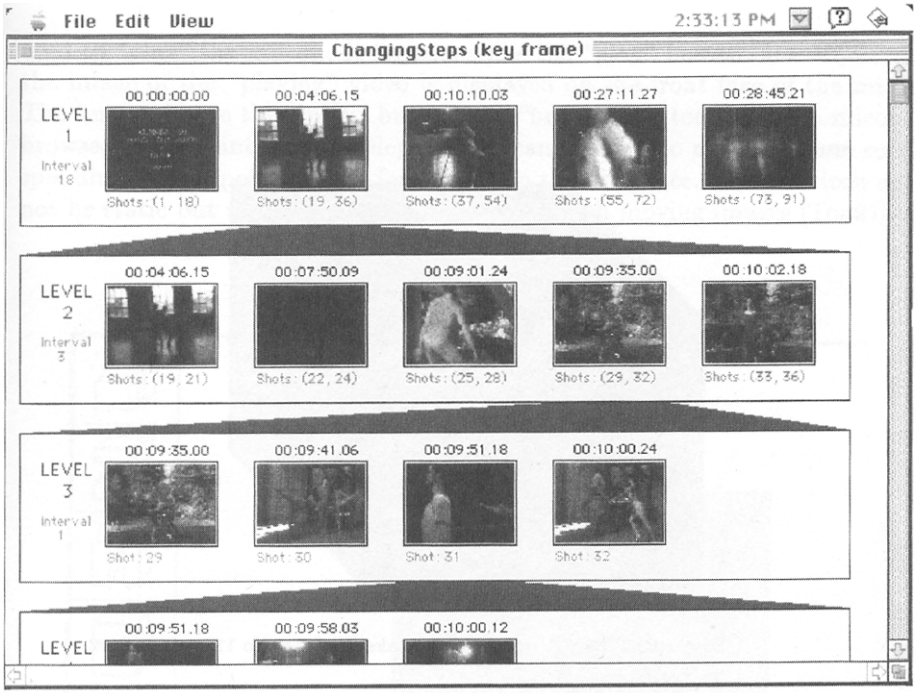
Key-Frame Extraction

- Advantage—simple calculation: when significant content change occurs between the current frame of the shot and the last key frame, the current frame is selected as a key frame - color, texture, motion
- Disadvantage—not very representative



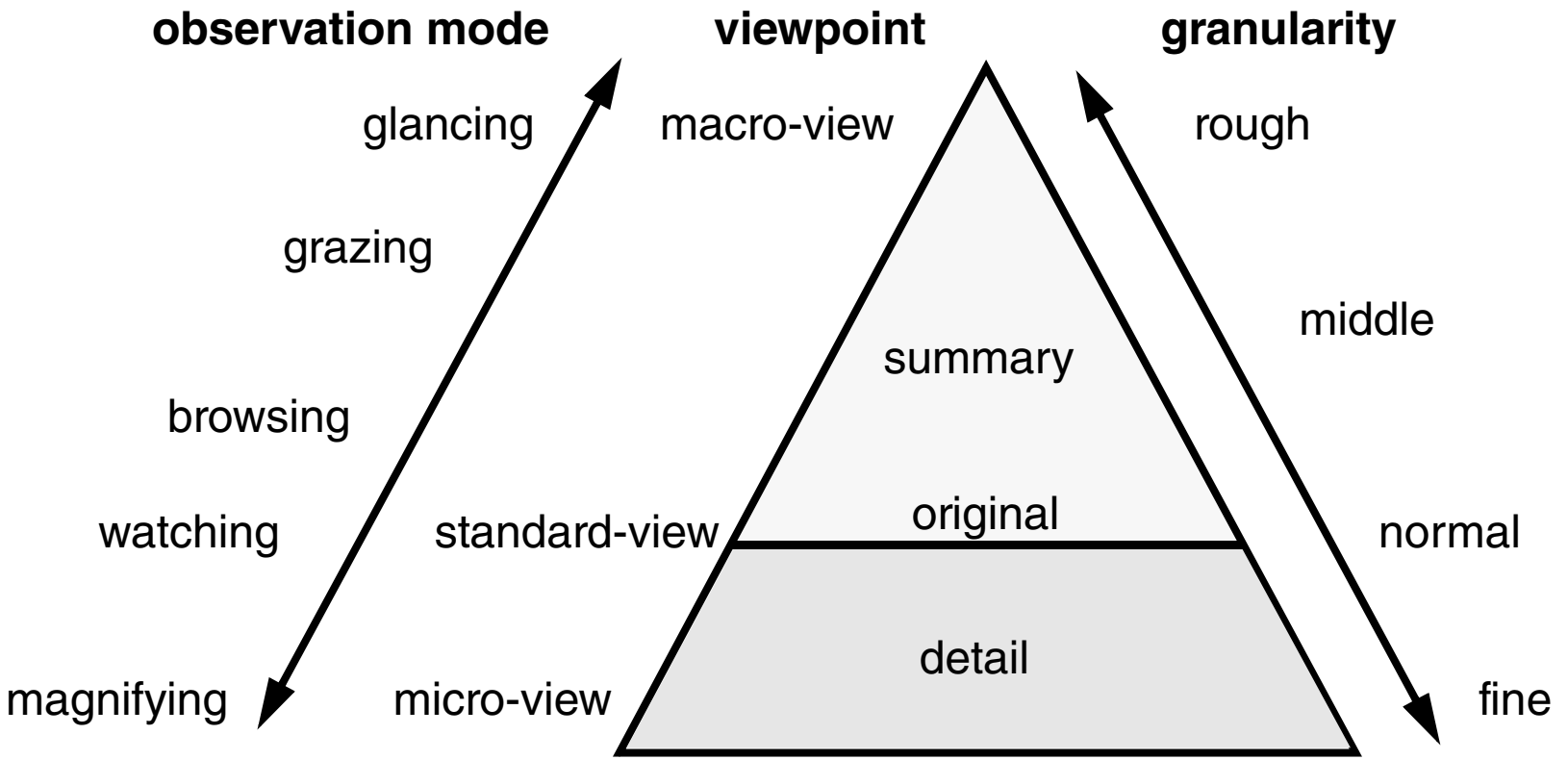
Video Browsing Tools

- Time line and strata browsers—shot-based image component line (basic); strata models provide further lines representing a particular angle of view on the document (shots, transitions, speech, music,...)
- hierarchical browsers—



Video Content Indication

- *Content indication* is the process of explicitly presenting some part of content for its better comprehension.



Video Content Indication

Interfaces for mm content indication should offer

- smooth stage transition between observation modes
- sense of overview—interface should cover a wide range of related contents
- sense of partitioning
- effective presentation—highly intuitive representations are needed; not easy for temporal material as e.g. video
- attractiveness—interface should reserve the original indication style as much as possible

3.5 Segmentation of Compressed Video

Segmentation of Compressed Video

Processing discribed so far is concerned with operations on pixel representations of individual image frames

NOW: Examine Videos in compressed form

- Hardware decompression—the procedures described before may be used without any computational overhead
- Software decompression—far less efficient
- If hardware decompression not available—use metrics performed directly on compressed representations
  - DCT Coefficients - both JPEG and MPEG representations
  - MPEG motion vectors

Algorithms Based on DCT Coefficients

- MPEG-Video —consists of I, P, B frames
  Only I-frames encoded with DCT coefficients
- GOP structure determines frequency of I-frames
- advantages —since only a small portion of all video frames are I-frames, computing time is reduced
  due to the large skip factor ∅ both gradual transitions and breaks will be dedected during one pass
- disadvantages —the loss of temporal resolution may introduce false positives

DCT Coefficients Correlation

DCT Correlation—DCT coefficients of consecutive frames of JPEG compressed video are compared.

- A subset of the DCT Coefficients of a subset of the blocks of the frame is extracted to construct a vector representation for that frame:
  $V_f = \{c_1, c_2, c_3, \ldots, c_k\}$
  the difference metric Ψ between two frames is defined

$$\Psi = 1 - \frac{|V_f \bullet V_{f+\varnothing}|}{|V_f|\,|V_{f+\varnothing}|}$$

normalized innerproduct

$\varnothing$....number of frames between the two frames being compared
$\Psi = 0$ if no change between the two frames

DCT Block Comparison

Pair-wise DCT Block Comparison

- Difference of a particular block in two frames—

$$Diff_l = \frac{1}{64} \sum_{k=1}^{64} \frac{|c_{l,k}(i) - c_{l,k}(i+\varnothing)|}{\max[c_{l,k}(i), c_{l,k}(i+\varnothing)]} \times 100\%$$

$Diff_l$.....content difference of block l

$c_{l,k}(i)$.....k^{th} DCT Coefficient of block l in frame i

- two thresholds—
  $Diff_l > t$ ....a particular block has changed across two frames
  $D(i,i+\varnothing) > T_b$ ...the percentage of blocks having changed: defines a camera break

Algorithm Based on Motion Vector

- Motion vectors in MPEG data stream—
  - P-frames—a single set of motion vectors
  - B-frames—two sets of motion vectors,for- and backward
- Field of motion vectors in a video—
  - within a single camera shot—relatively continuous changes
  - between differrent shots—continuity will be disrupted
- Definition M—
  P-frame: number of valid motion vectors
  B-frame: smaller number of valid mv (forward and backward)
- if $M < T_b$... camera break

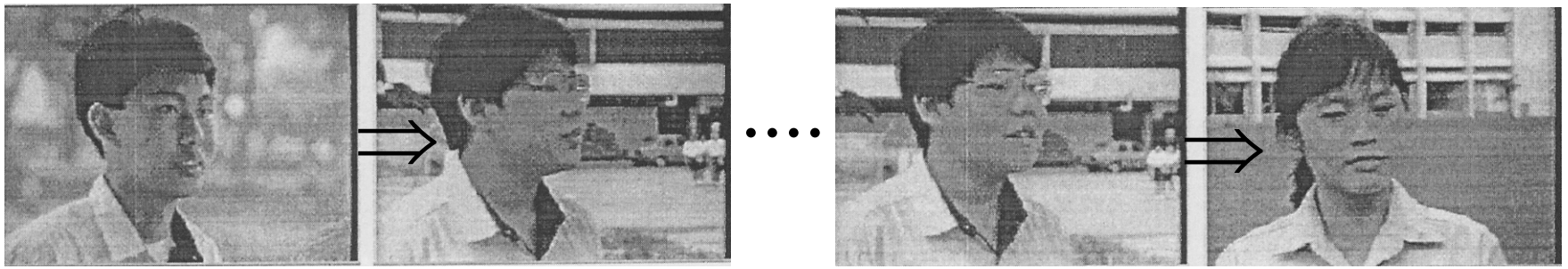
Hybrid Approach to Partitioning

Multiple passes and multiple comparisons

- first pass—DCT comparison (I-frames) with a large skip factor ∅ to dedect regions of potential breaks, gradual transitions, camera operations, object motions
- second pass—DCT comparison with smaller skip factor applied only to the neighborhood of potential breaks, transitions.... deletes false positives
- further passes—motion-based comparison either on the entire video or only on the sequences dedected before
  these passes verify DCT results and improve accuracy

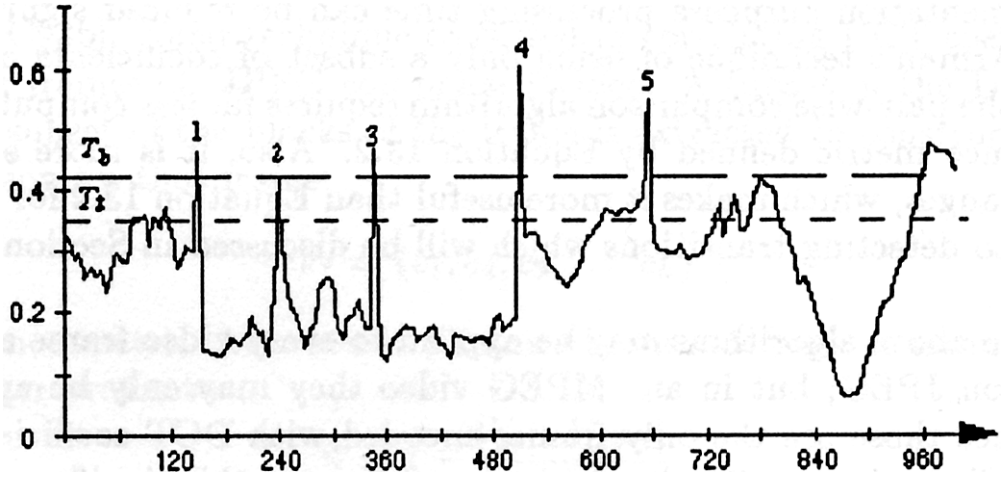
Evaluation of Algorithms

- Testcase MPEG compressed documentary video—78 seconds long (2340 frames), contains 25 shots separated by 17 breaks (labeled 1-17) and 7 gradual transitions
  - 4 shots involve camera panning
  - GOP— IBBPBB
  - specific images from the tested sequence—



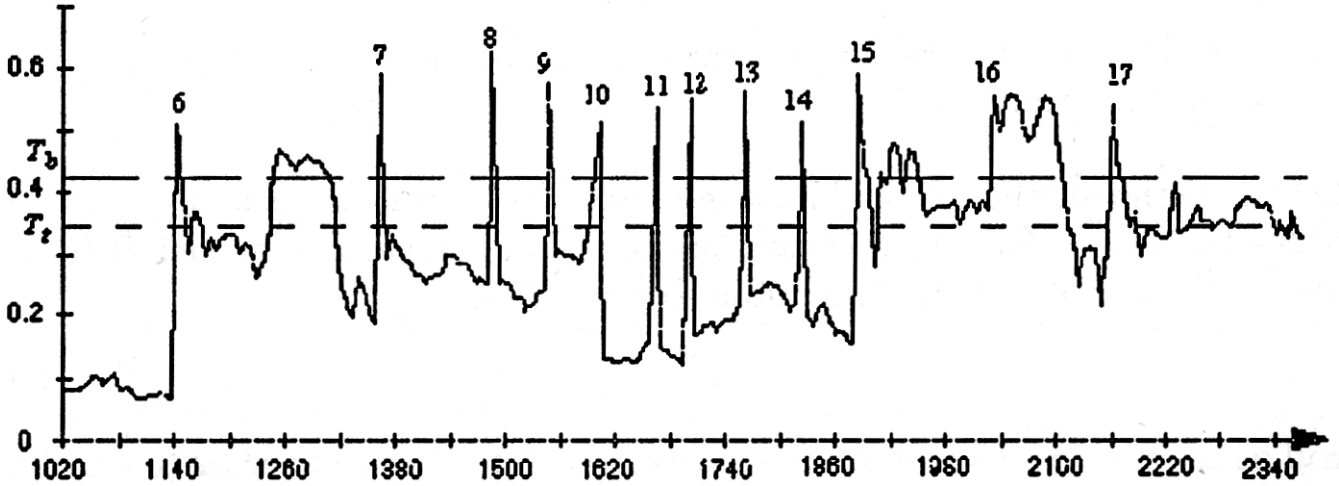
Break 2 Break 3

Evaluation DCT-Correlation



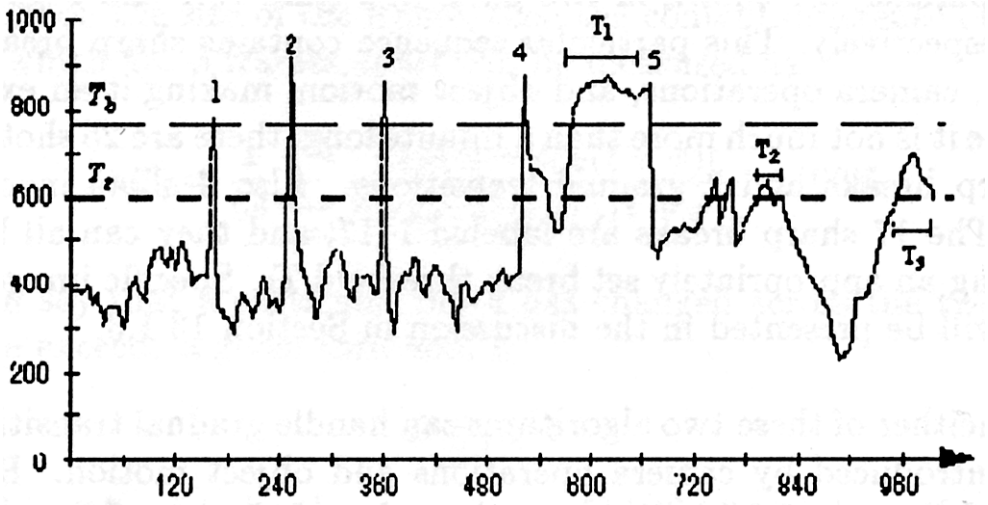
T_b....Threshold break
T_t.....for twin comparison

...first half



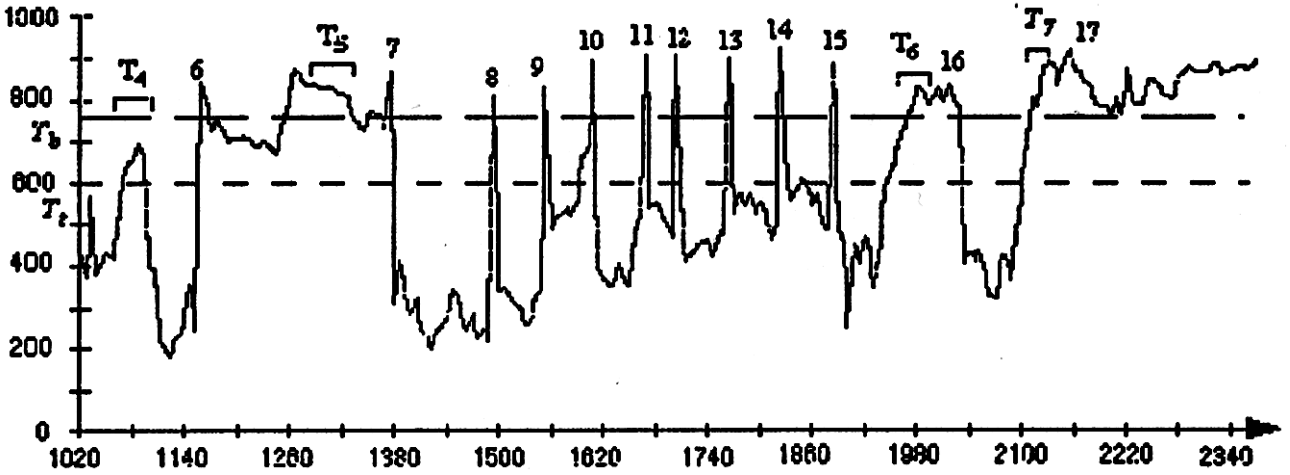
...second half

Evaluation Pair-wise Block Comparison



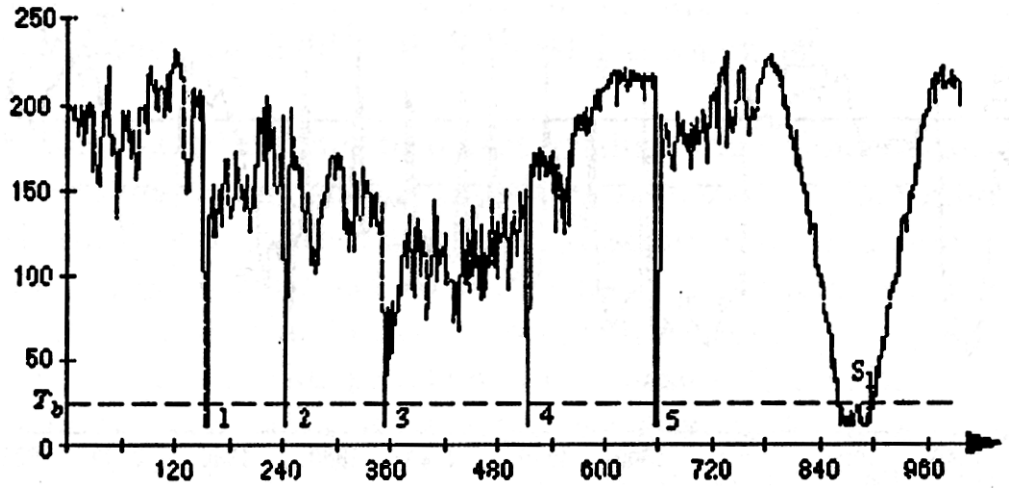
T_1-T_7..gradual Transitions

...first half



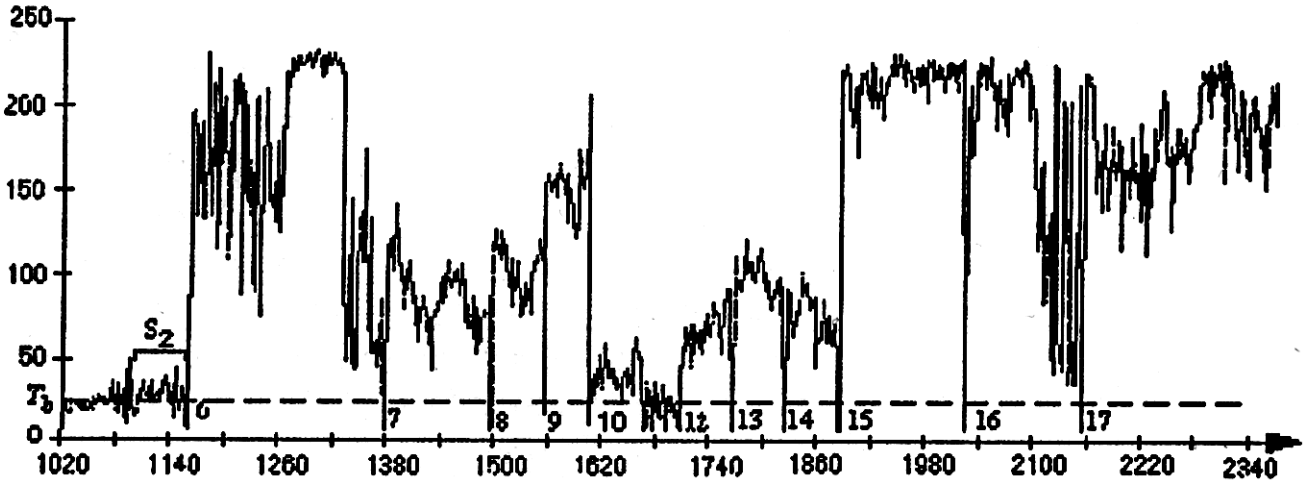
...second half

Evaluation Motion Vectors



S_1, S_2.....false positives

...first half



...second half

Evaluation Summary of Results

BREAKS	dedected	undedected	falsely dedected
DCT 1	16	1	4
DCT 2	17	0	4
motion vector	17	0	0
hybrid	17	0	0

GRADUAL TRANSITIONS	dedected	undedected	falsely dedected
DCT 1	3	5	1
DCT 2	4	4	1
motion vector	0	7	0
hybrid	7	0	0

Evaluation Summary of Results

- Discussion—
  - breaks—motion vector and hybrid dedected all breaks DCT less effective; false dedection mainly resulted from the fact that only I frames can be used
  - gradual transitions—motion vector failed completley
  - conclusion—hybrid approach provides highest accuracy in dedecting both breaks and gradual transitions.
- Camera operation algorithm successfully dedected all four of the camera pans in the test data

3.6 Case Studies

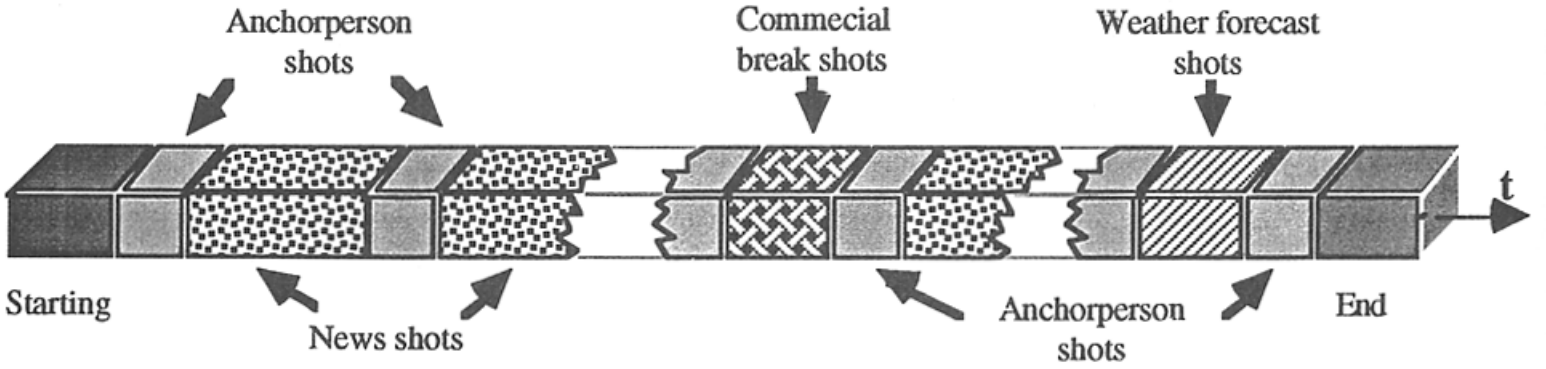
3.6.1 Television News

Case Study —Television News

Automatic extraction of semantic information of a video only possible when its structure is based on domain knowledge.

- Example — TV News
  - Spatial structure—anchorperson shots

  - Temporal structure—in the order of shots



News Video Parsing Algorithms

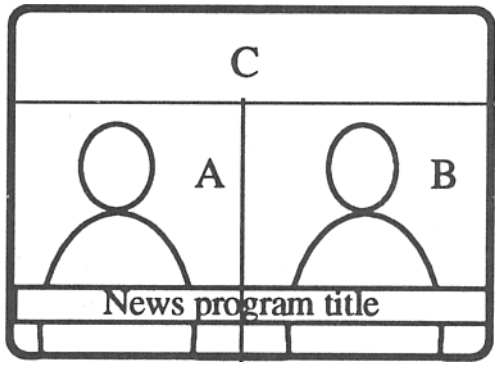
Video parsing process consists of three steps

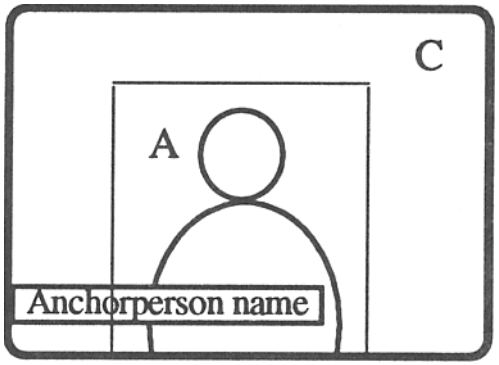
- temporal segmentation— see *Video Segmentation*
- classification of the shots—
  - define a model of an A shot (anchorperson shot)
  - develop similarity measures to be used in matching these models
  - temporal structure model of an entire news program—sequence and episode identification
- visual abstraction—k-frames for each shot to represent its visual content

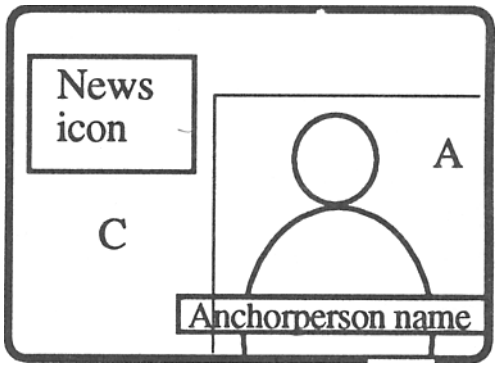
Anchorperson Shot Model 1

First task in shot classification is data-modeling. Three models are proposed—

- Frame models— a frame of an A shot is strictly composed according to one of the following spatial structures







Identifying an A-frame requires first classification of its regions structure model

Anchorperson Shot Model 2

- Region models—
  - region A and/or B—*anchorperson(s)*
  - region *news icon*
  - region *news program title bar*
  - region *name bar*
  - region C—*background*
- Shot models—
  background tends to be relatively fixed, talking-head(s) tend(s) to be static → frame-to-frame changes (both histogram and pixel comparison) are very small

Matching A-shots 1

- first step—using shot model; mean μ and variance σ of the difference values over the entire shot

  if $\sigma^2 < t_1,\ \mu < t_2$

  shot is declared a potential A shot

- second step—frames divided into regions *A*, *B*, *C* and/or *A*,*C*

  if $\mu_A, \mu_B > \mu_t$ *and* $\sigma_A^2, \sigma_B^2 > \sigma_t^2$

  caused by movements of head, face, hands....

  and $\mu_C \approx 0, \sigma_C \approx 0$

  shot is declared a A shot. This step also establishes which of the spatial structure models is used

Matching A-shots 2

- creating a model image—

$$p_A(i,j) = \frac{1}{N}\sum_{n=1} p_n(i,j)$$

N...number of frames in the shot
$p_n(i,j)$...value of a pixel in frame n
$p_A(i,j)$...average pixel value over a shot

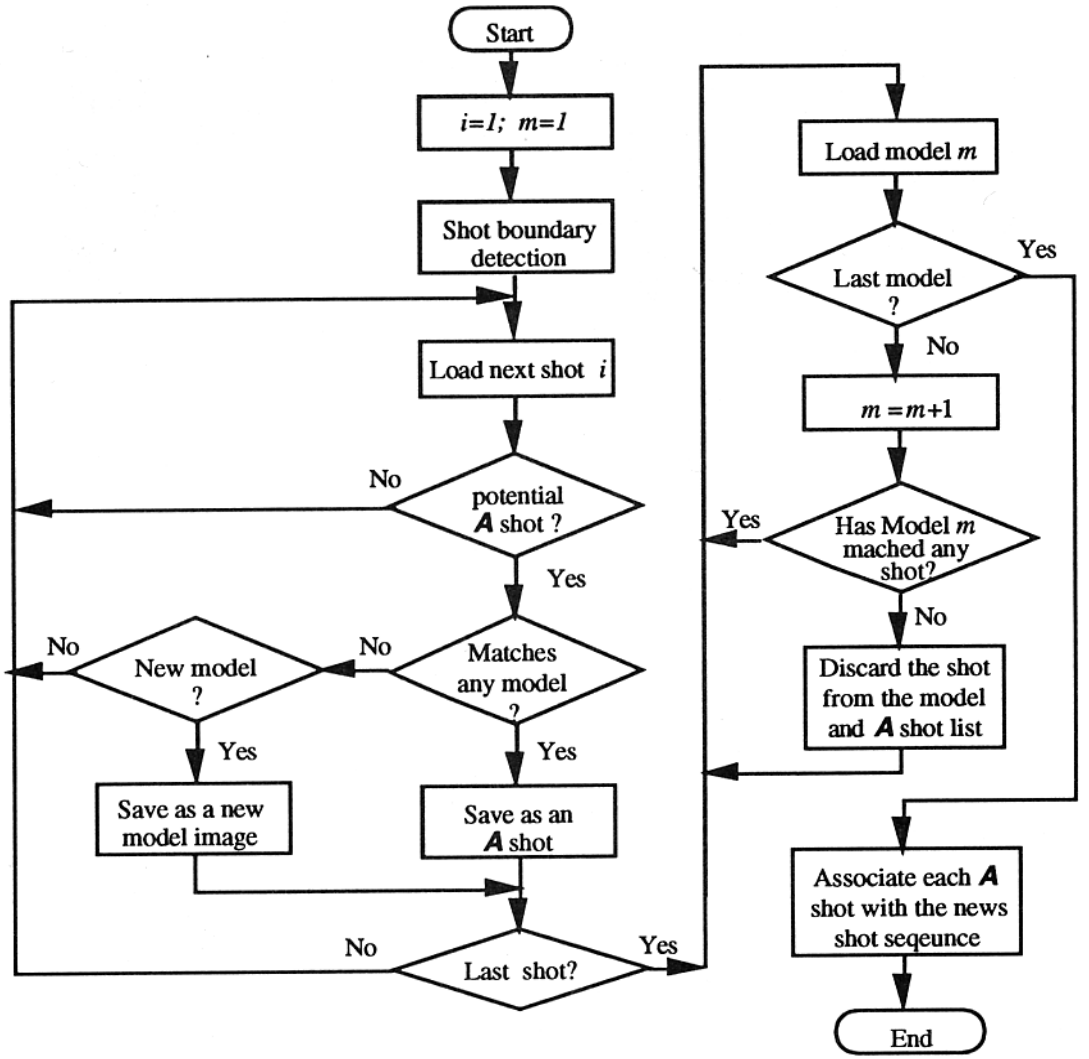
Image $\{p_A(i,j)\}$ and its histogram stored as *model image*— will be used to identify the same type of A shots among the remaining candidates. Comparison (histogram, pixel) only computed in the region *A* and *B* →

- reduction of computation time
- avoids problem of variation in the rest of the frame (e.g. induced by changing news icons)

Matching A-shots 3

- Model images are constructed sequentially from the beginning of the broadcast.
- Once a candidate A shot is verified as a given type, region image and histogram are stored as the model for that type.
- If a new candidate shot is of a new type model image processing is repeated for that type.
- Models are constructed incrementally until all types have been modeled.
- Models with no match in the subsequent processes will be discarded - will eliminate mistaking an interview news shot for an A shot.

Matching A-shots 4



Matching other Shot Types

- Starting and ending sequences—used as a signature of a production, employed very consistently, has a fixed and predefined temporal and spatial structure, may be easily identified
- Commercials—mostly pre-defined starting sequence when the program returns from a commercial break
- Weather Forecast Shots—considerable variation, in general the frame which is displayed tends to be identical across the entire shot, often a pre-defined starting sequence
- News shots—too much variety to support any structure model may be identified as those that do not conform to any model

IV.4 Audio Retrieval

Einleitung / Motivation

- Weltweit beherbergen Archive (Filmarchive, Fernseh-, Radiostationen, Musikarchive, etc.) eine gigantische Anzahl an Video- und Tondokumenten.
- exponentielles Wachstum von Musikangeboten
- nur automatische Indizierung kann solche Archive auf Dauer nutzbar machen

Audio vs. Visuelles Retrieval

- Heute wird oft mehr Augenmerk auf visuellen Aspekt (Bild, Video) gelegt.
- akustischer Teil oft (noch) außer Acht gelassen
- In vielen Fällen Ton aussagekräftiger als Bild
  - Beispiel Videoszene mit Dialog
    Gesprochenes ist aufschlussreicher als Bild

Audio-Indexierung und -Retrieval

einfachste Methode: über Titel und Dateiname ...

- sehr verbreitet
- Namen allerdings unvollständig und subjektiv –schwierig zu finden
- außerdem keine Möglichkeit, Audio-Aufnahmen zu finden, die so klingen wie etwas, was gerade zu hören ist
- www.pandora.com

Audio-Indexierung und -Retrieval

Inhalt verwenden

- Vergleich Messwert für Messwert
  - wenig erfolgversprechend, da Unterschiede in Abtastrate und Auflösung nicht berücksichtigt
- daher Merkmale (Features) extrahieren und nutzen
  - mittlere Amplitude
  - Frequenz-Verteilung

Allgemeiner Ansatz

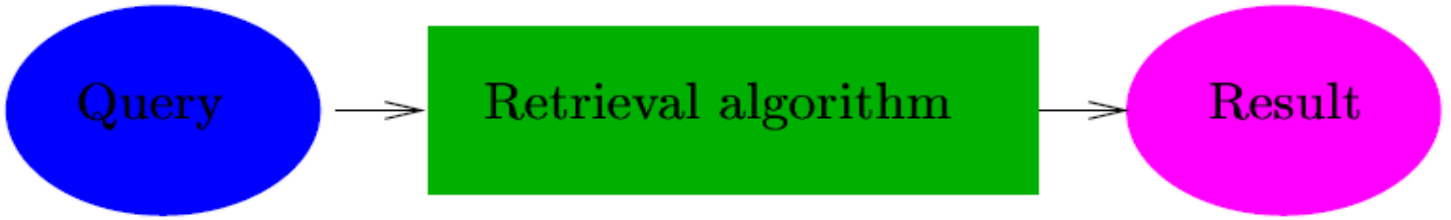
- Klassifikation
  in verbreitete Typen wie Sprache, Musik, Geräusch
- differenzierte Behandlung jeder Klasse
  z. B. Sprache: Spracherkennung und Indexierung des Textes
- Anfragen
  ebenso klassifiziert, verarbeitet und indexiert
- Retrieval
  beruht auf der Ähnlichkeit der Anfrage-Merkmale mit den Merkmalen der gespeicherten Tondokumente

Klassifikation

- verschiedene Typen verlangen unterschiedliche Verarbeitung und unterschiedliche Indexierungstechniken
- verschiedene Typen haben unterschiedliche Bedeutung für eine Anwendung
- Sprache ist der wichtigste Typ, und es gibt heute recht erfolgreiche Spracherkennungs-Techniken und –Systeme
- die Typinformation selbst ist in einigen Anwendungen sehr nützlich
- der Suchraum reduziert sich auf eine Klasse

II.4.1 Grundlagen

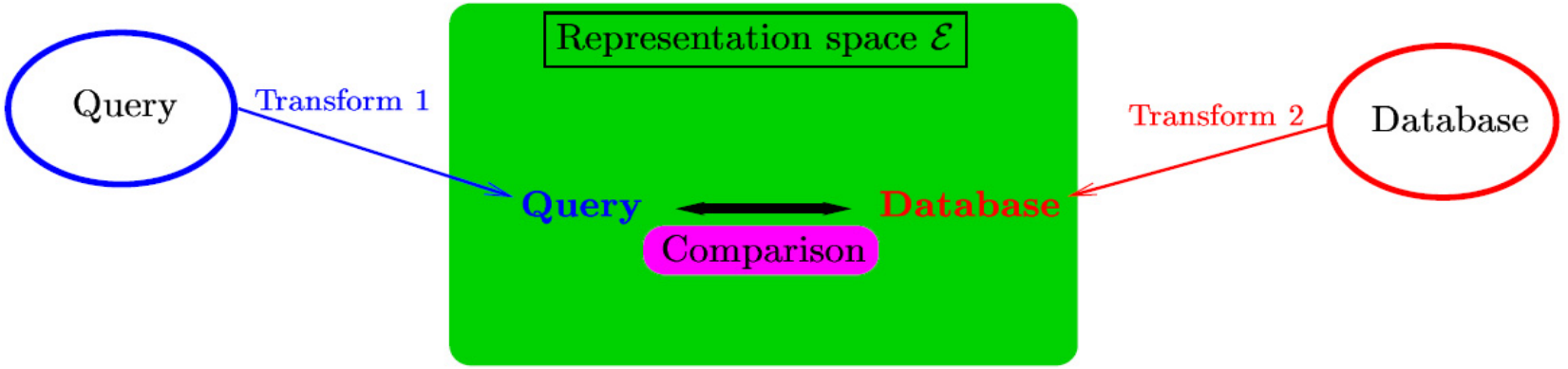
Retrieval Prozess



Beispiele für Anfragen

- Query by Humming
- Finde ein bestimmtes Wort in einem Nachrichtenarchiv
- Finde einen bestimmten Tierlaut in einem Tierlaute-Archiv

Repräsentationsraum



- Abfrage-Audiosignal und Audiosignale der Datenbank werden in Repräsentationsraum (e) transformiert
- Vergleich erfolgt im Repräsentationsraum

Repräsentationsraum

- In Abhängigkeit von der Aufgabe: viele Repräsenta-tionsräume und entsprechend viele Ähnlichkeitsmaße
- 2 Gründe für e:
  - Abfrage nicht notwendigerweise in Waveform
  - e muß für die konkrete Aufgabe ***diskriminieren***:
    e ist extrem abhängig vom konkreten Problem
- e ist ein Raum von Audio Features

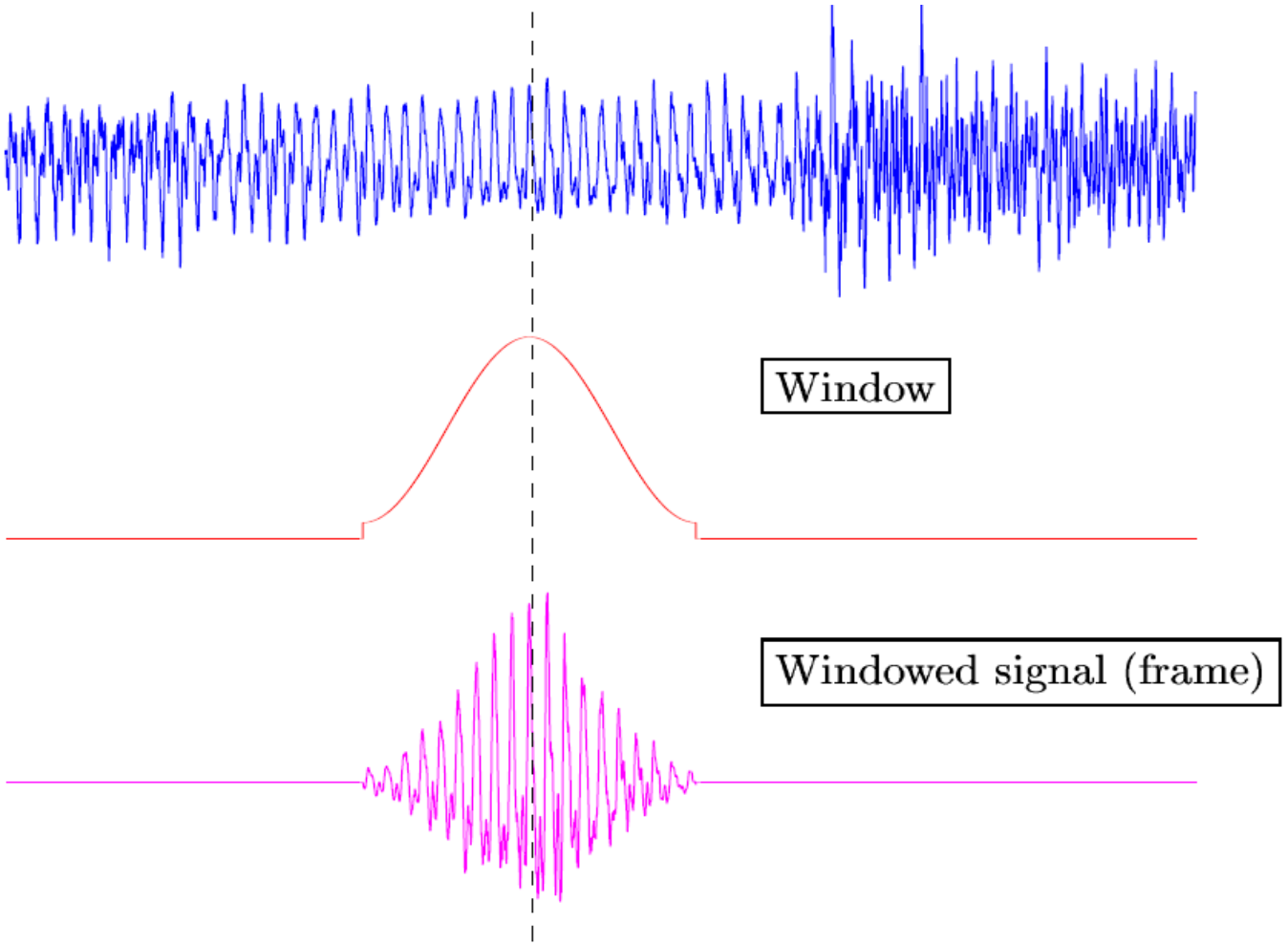
Segmentierung des Signals

Frame

- jeder Frame ist an einer bestimmten Zeit t positioniert
- charakterisiert das Signal zur Zeit t
- Multiplikation des Drucksignals mit einem Fenster(-signal) (positioniert zur Zeit t) ergibt einen Frame
- t korrespondiert mit dem Zentrum oder einem der Endpunkte
- Windows können verschiedene Formen haben (z. B. Hamming, Bartlett, Dreieck, Rechteck)

Frames

g

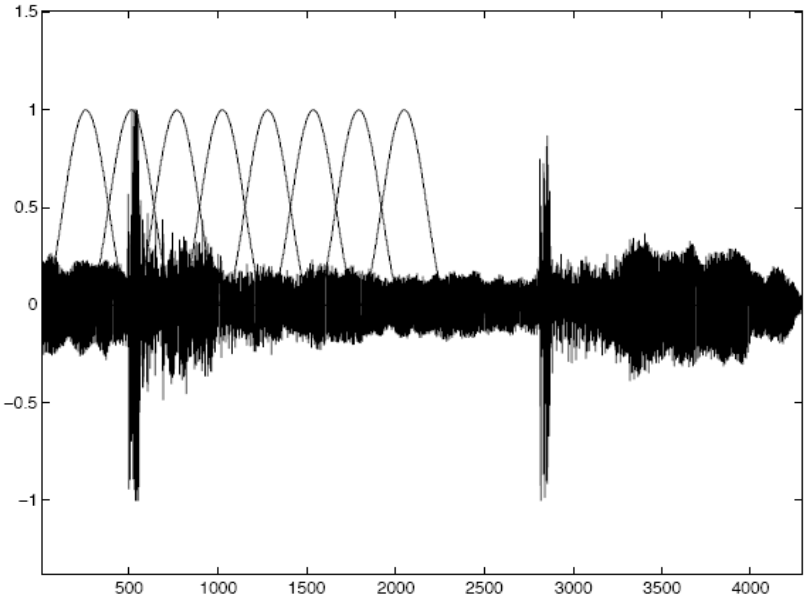


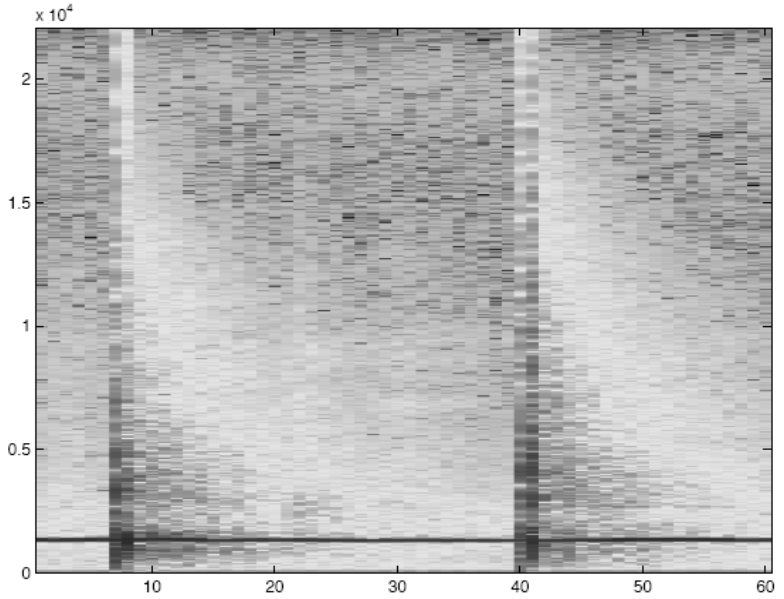
Spektrogramm

- einfache Darstellungen haben Grenzen:
  - Zeit-Domäne zeigt Frequenz-Anteile eines Signals nicht
  - Frequenz-Domäne zeigt nicht, wann Frequenzen auftreten
- kombinierte Darstellung - Spektrogramm
  - x-Achse: Zeit, y-Achse: Frequenzanteile
  - Schwärzung (Farbe) eines Punkts: Energie der Frequenz zu dieser Zeit
- Analysen

  z.B. Regelmäßigkeit des Auftretens von Frequenzen, Musik vs. Geräusch

Zeit-Frequenzdarstellung





Vorgangsweise

- Eingangssignal wird blockweise verarbeitet
- überlappende Segmente des Signals werden verwendet
- sinusoide Fensterfunktionen in Abb. deuten Signalausschnitte an, auf die sich die Analyse in einem Schritt „konzentriert“
- punktweise Multiplikation von Signalblock mit Fensterfunktion
- resultierendes Signal wird Fouriertransformiert
- Aneinanderreihung der Spektralvektoren liefert Zeit-Frequenzdarstellung des Signals

Short Time FT (STFT)

- Fenster (Frame, Windowed Signal):

$$x'_t(\tau) = x(\tau)w(\tau - t)$$

- Fenster w, Zeit t, t Fensterposition
- Fourier Transformation von x'_t ist STFT von x zur Zeit t mit Fenster w

Spektrogramm

- Spektrogramm von *x* ist das Quadrat der STFT:

$$\begin{aligned} \mathrm{Spect}_x(t,f) &= \left| \int x(\tau)\, w(\tau - t)\, e^{-j2\pi f\tau}\, d\tau \right|^2 \\ &= |\mathrm{STFT}_x(t,f)|^2 \end{aligned}$$

- häufig statistische Interpretation als nicht normalisierte Dichtefunktion (pdf) über die Frequenz
- erlaubt Berechnung statistischer Parameter, z.B. Lagemaße, Streuung

Audio Klassifikation

- Sprache
  - männliche oder weibliche Sprache
- Musik
  - Arten von Musik
- Umgebungsgeräusche (z.B. Tierlaute)

Sprache

- Bandbreite vergleichsweise gering, 100 bis 7000 Hz
- Zentroid deshalb niedriger als bei Musik
- häufige Pausen (zwischen Worten und Sätzen)
- höherer Anteil der Stille
- charakteristische Struktur: Folgen von Silben, die aus kurzen Perioden von Friktionen (Konsonanten) bestehen, auf die längere Perioden von Vokalen folgen
- während der Friktionen hohe Nulldurchlaufrate, ZC variiert stärker

Musik

- hohe Bandbreite, 16 bis 20.000 Hz
- Zentroid deshalb höher
- niedriger Anteil der Stille
- Ausnahmen: Soloinstrument, A-Capella-Gesang
- Nulldurchlauf variiert nicht so stark
- regular beat

Vorgang Klassifikation

Schritt für Schritt

- ein Merkmal nach dem anderen
- z. B. erst Zentroid

  wenn hoch: Musik
- dann Anteil der Stille

  wenn niedrig: Musik
- dann ZC-Variabilität

  wenn niedrig: Solo-Musik

  sonst: Sprache

Vorgang Klassifikation

- Reihenfolge wichtig
  algorithmische Komplexität und Differenzierungsvermögen
  einfach zu berechnen und hohe Differenzierung zuerst
- ein Merkmal allein auch schon nutzbar:
  nur ZC: bis zu 90 % korrekt klassifiziert
  nur Anteil der Stille: bis zu 82 %

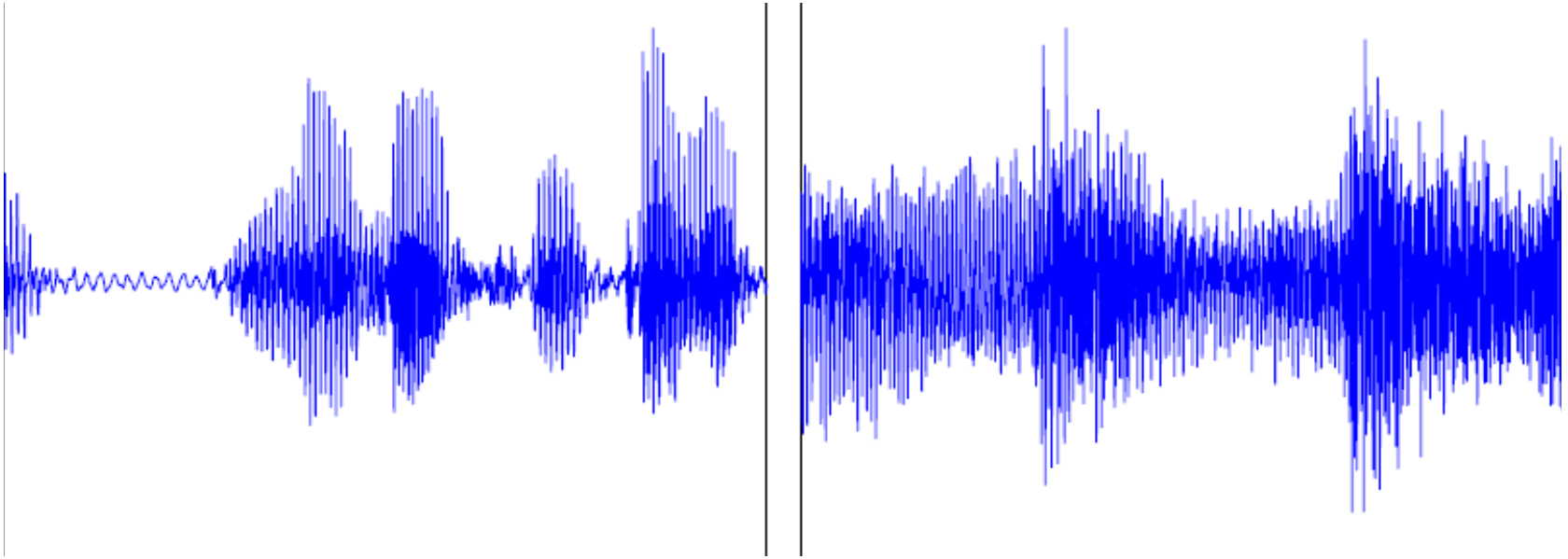
Musik-Indexierung

- noch in den Anfängen
- www.pandora.com
- 3 Arten
  - strukturierte Musik
  - synthetische Musik (MIDI)
  - aufgezeichnete Musik

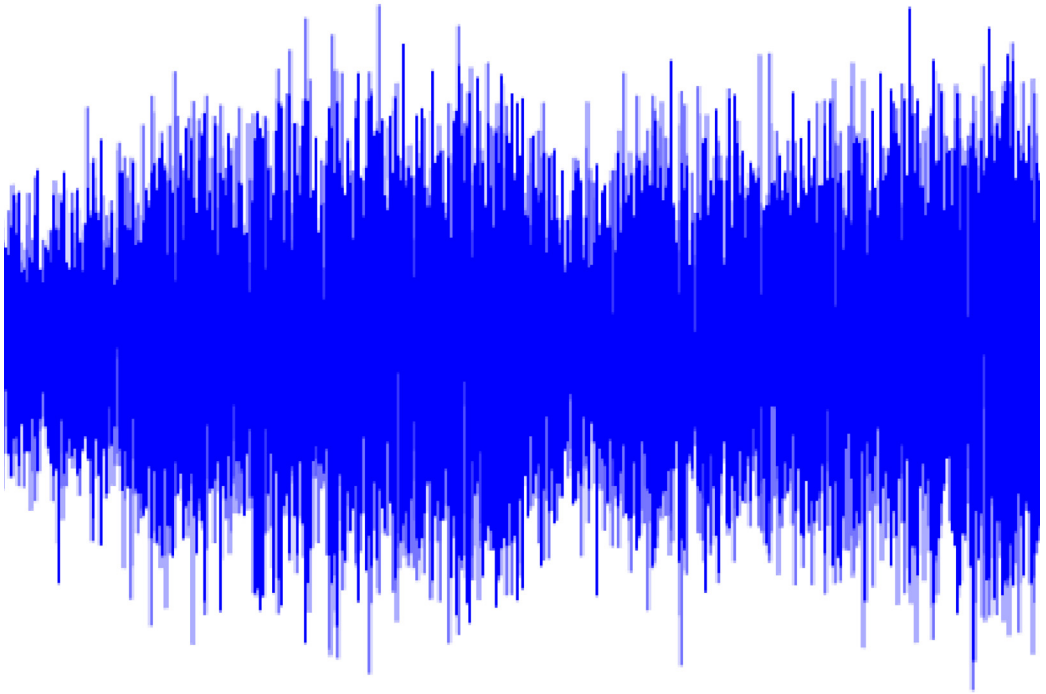
Indexierung strukturierter Musik

- keine Extraktion von Merkmalen erforderlich
- sogar exakte Übereinstimmung als Suchmethode denkbar
- allerdings könnten Instrumente nicht übereinstimmen
- Ähnlichkeit schwierig zu definieren
- eine Möglichkeit: nur den Pitch-Wechsel berücksichtigen
  Up, Down, Repeat – U, D, R
- Retrieval durch Zeichenkettenvergleich
- www.melodyhound.com: (Uni Karlsruhe)

Beispiele: Sprache / Musik

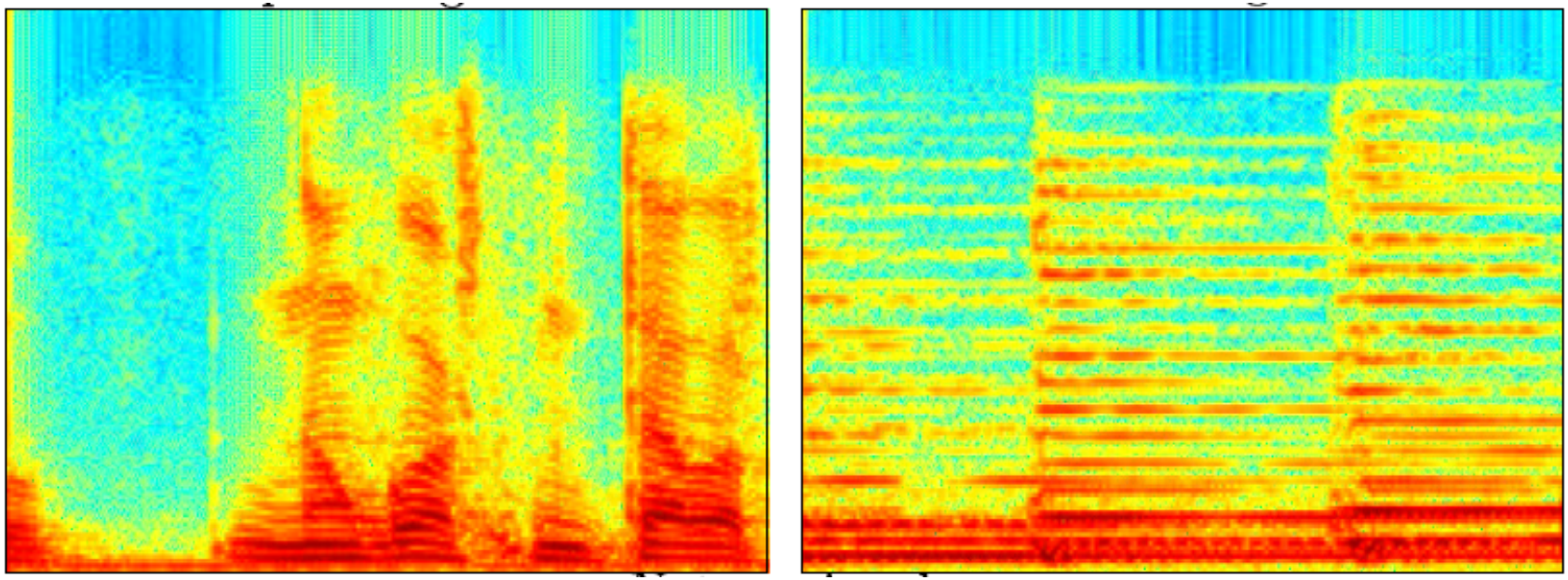


Beispiel Tierlaut



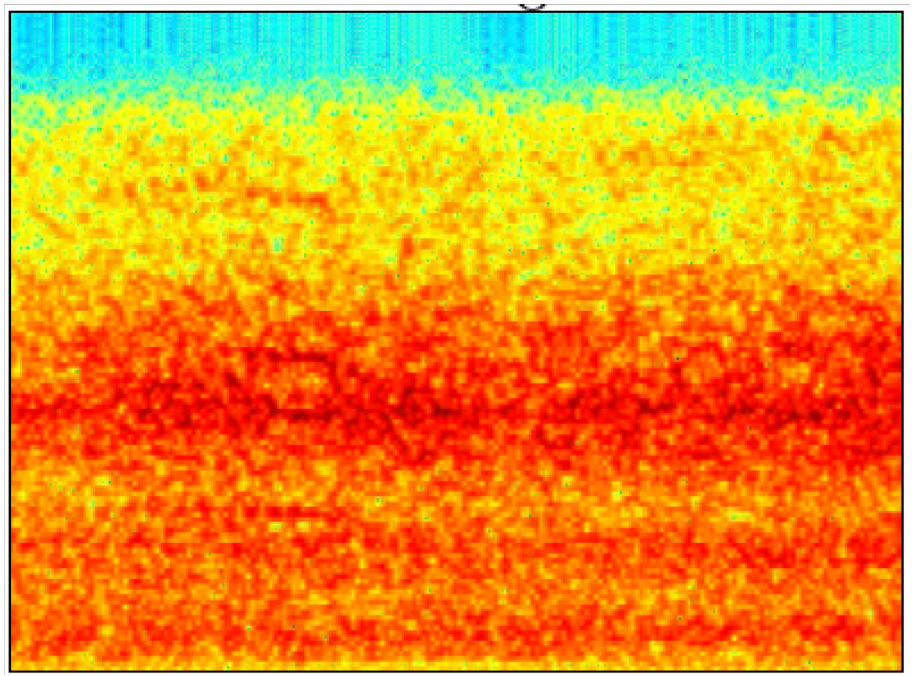
Beispiel-Spektrogramme

- Sprache / Musik (Zeit 0-1.14 s, Frequenz 0-5512 Hz)



Beispiel-Spektrogramme

- Tiergeräusch



II.4.2 Audio Merkmale

(im Spektrogramm Bereich)

Merkmale Einteilung

unterschiedliche Einteilungen in der Literatur

- Sprache, Musik, Geräusch
- Frames oder Clips
- Subjektivität / Interpretation (z.B. Timbre)
- nach Repäsentationsräumen (Zeit, Frequenz, oder spezieller)

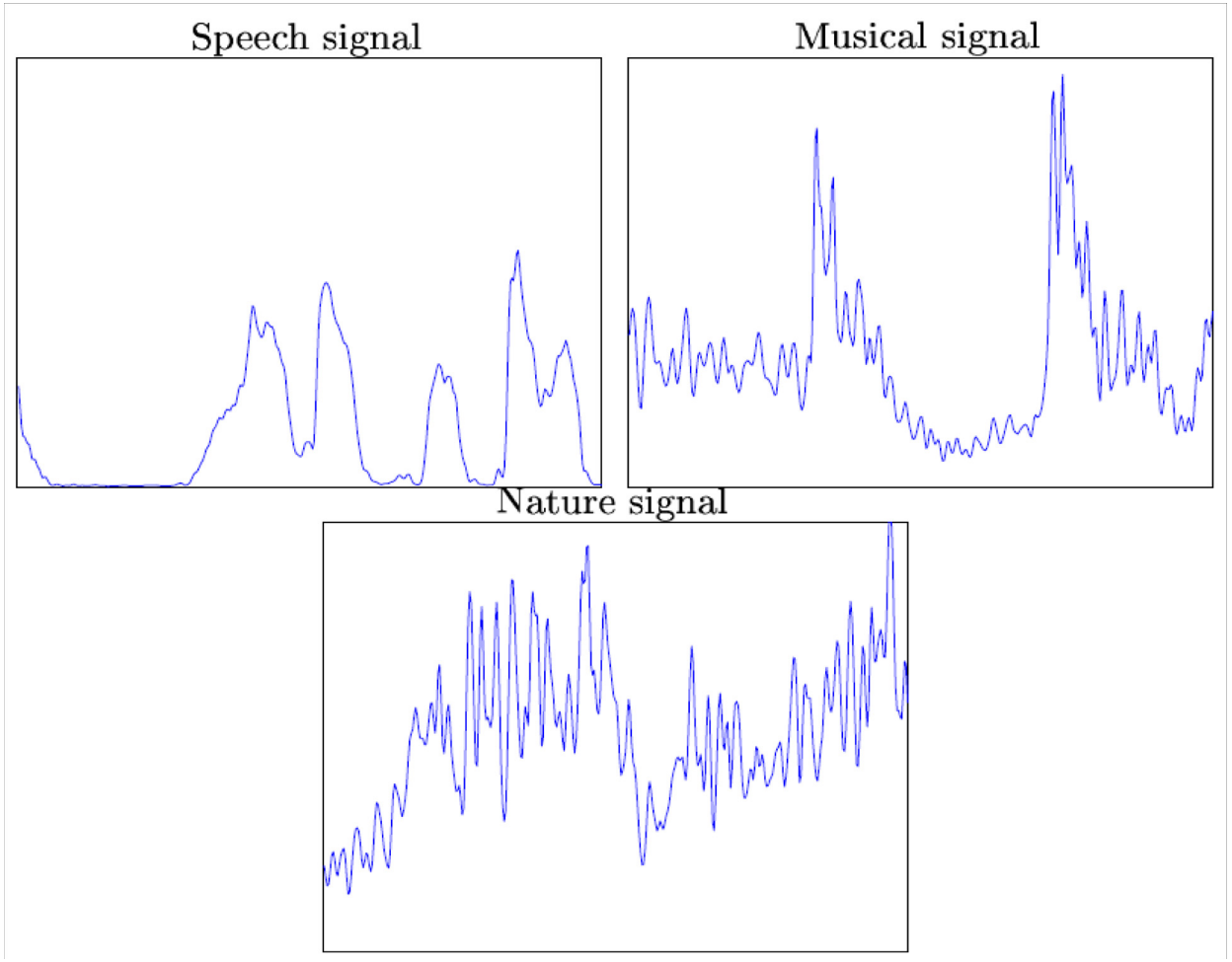
Volume

- auch Short Time Energy, Lautstärke
- viele Definitionen in der Literatur

$$\mathrm{Vol}(t) = \frac{\int \mathrm{Spect}_x(t, f)df}{\int w(\tau)d\tau}$$

$$\mathrm{Vol}(t) = \frac{\int |x(\tau)w(\tau - t)|^2 d\tau}{\int w(\tau)d\tau}$$

Volume Beispiele



Bandenergie

- Energie innerhalb eines Frequenzbereichs (Band) f_0 - f_1:

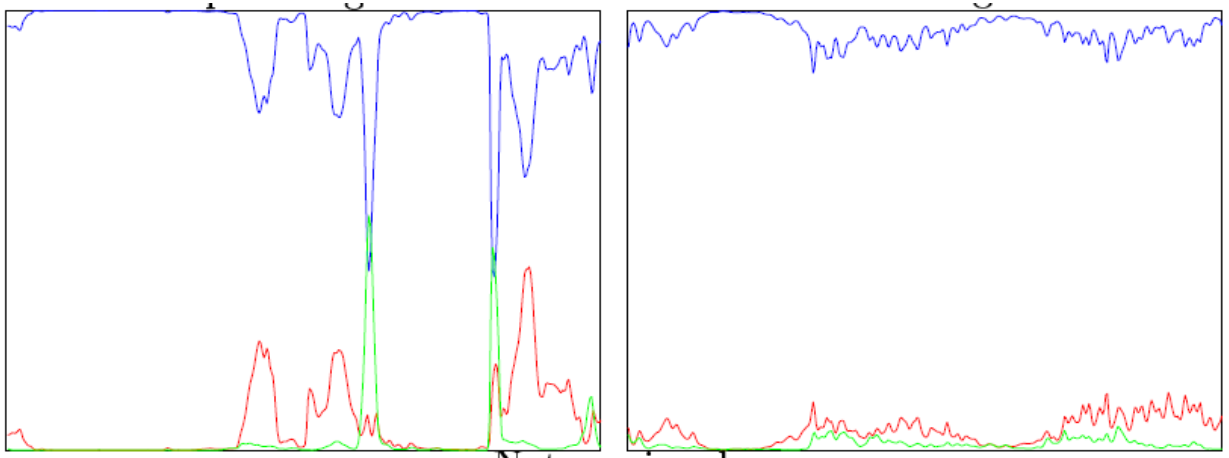
$$\mathrm{BE}_{[f_0,f_1]}(t) = \frac{\int_{f_0}^{f_1} \mathrm{Spect}_x(t,f)df}{\int w(\tau)d\tau}$$

- Band Energy Ratio:

$$\mathrm{BER}_{[f_0,f_1]}(t) = \frac{\mathrm{BE}_{[f_0,f_1]}(t)}{\mathrm{Vol}(t)}$$

- Histogramm Approximation von pdf $\mathrm{Spect}_x(t,f)$

Bandenergie Beispiele



blau 0-630 Hz

rot 630-1720 Hz

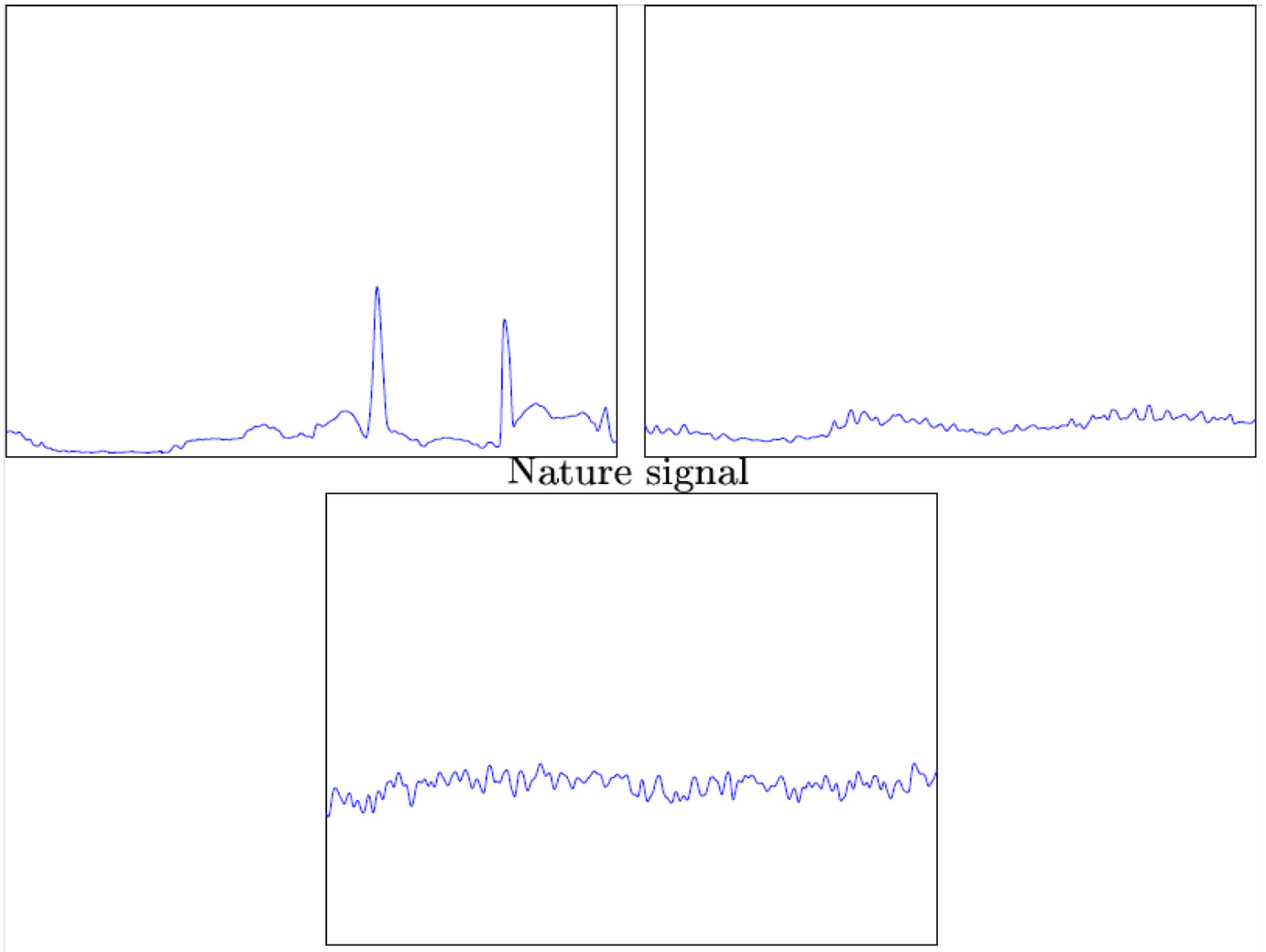
grün 1720-5512 Hz

Median Frequenz

- Median des Spektrums eines Frames, auch Centroid Frequenz oder Brightness genannt
- Lagemaß

$$\mathrm{MedF}(t) = \frac{\int f \, \mathrm{Spect}_x(t,f)df}{\int \mathrm{Spect}_x(t,f)df}$$

Median Frequenz Beispiele

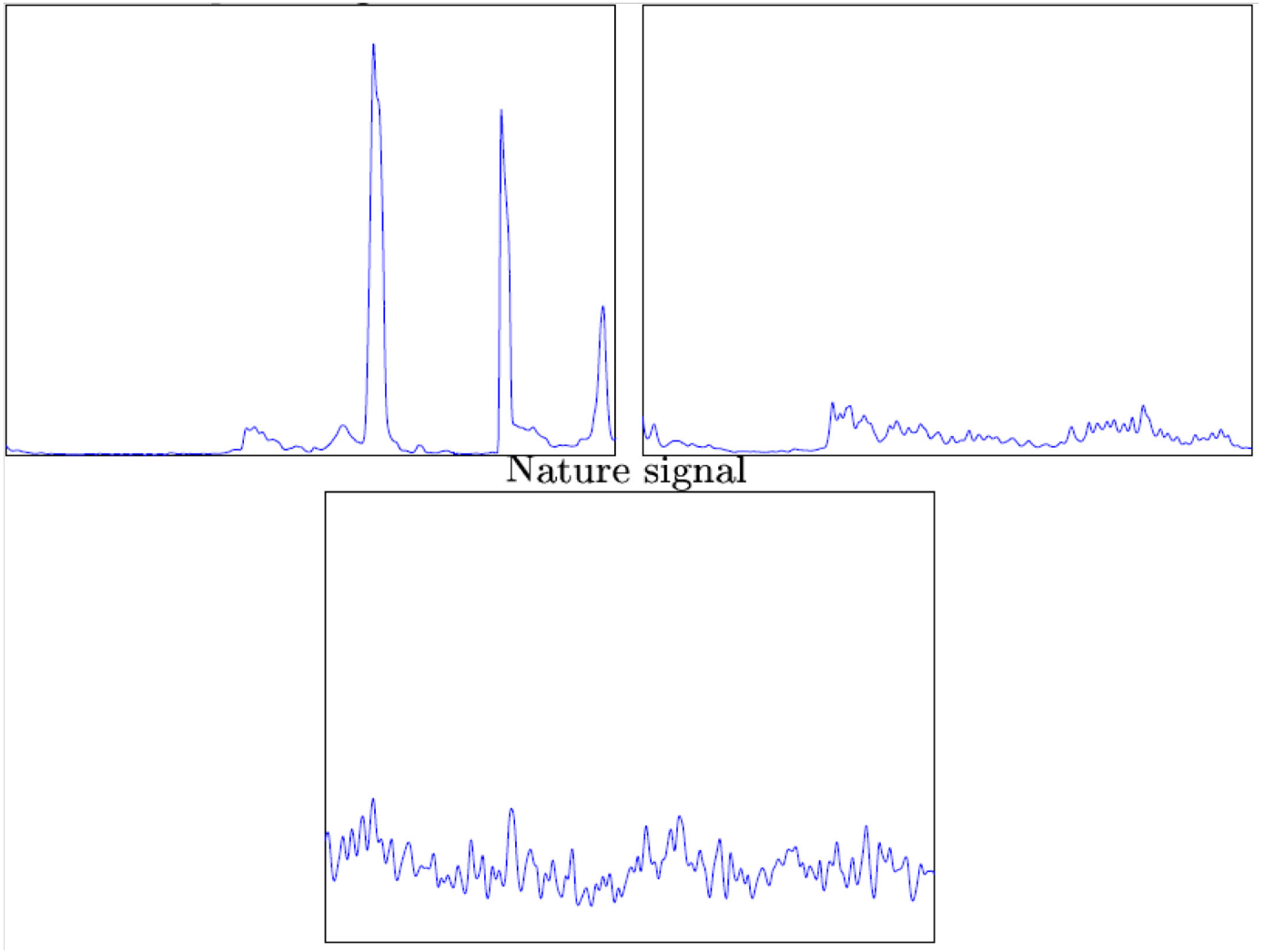


Bandwidth

- Streuungsmaß

$$\mathrm{BndW}(t) = \frac{\int [f - \mathrm{MedF}(t)]^2 \ \mathrm{Spect}_x(t, f) df}{\int \mathrm{Spect}_x(t, f) df}$$

Bandwidth Beispiele



Anteil der Stille

- silence ratio
- Anteil der Messwerte an der Gesamtzahl, die einer Periode (!) der Stille angehören
- zwei Schwellenwerte:
  - Amplitudenwert, unterhalb dessen Stille angenommen wird
  - Anzahl unmittelbar aufeinanderfolgender Messwerte, die mindestens still sein müssen, um eine Stilleperiode zu bilden

Harmonie

- spektrale Komponenten oft Vielfache der niedrigsten und lautesten Frequenz ("fundamental frequency")
- Musik in der Regel harmonischer als andere Geräusche
- Prüfung, ob eine Tonaufnahme harmonisch ist: dominante Komponenten Vielfache der fundamentalen Frequenz?
- Beispiel: Flöte spielt Note G4; Spitzen bei den Frequenzen 400 Hz, 800 Hz, 1200 Hz, 1600 Hz usw.
- f, 2f, 3f, 4f usw. Harmonische der Note

Tonhöhe

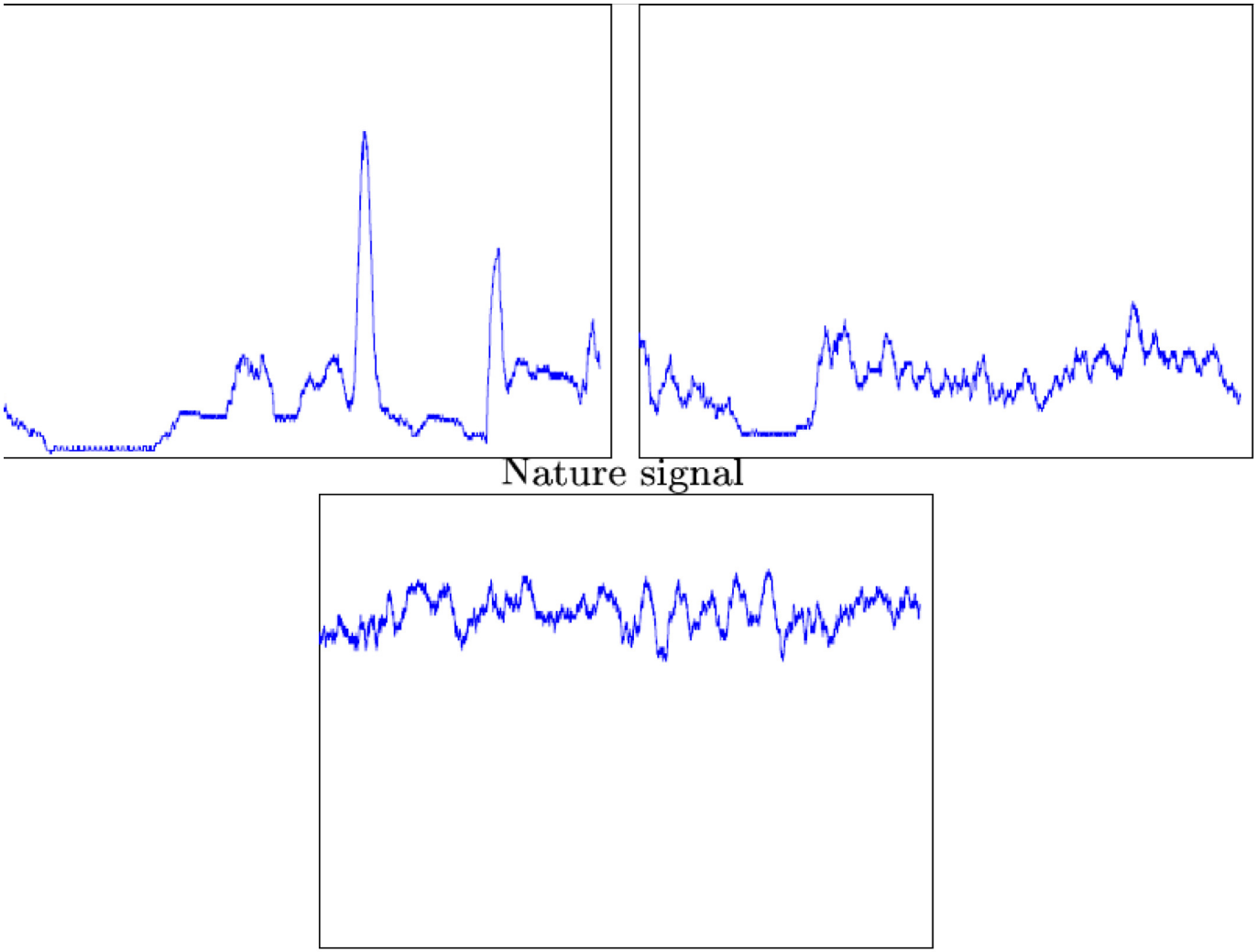
- nur periodische Klänge (Instrumente, Stimme)
- Perkussion dagegen nicht
- subjektiv; verwandt, aber nicht gleichbedeutend mit der fundamentalen Frequenz, die (oft als Näherung verwendet wird)
- viele unterschiedliche Berechnungsverfahren

Nullkreuzungsrate

- Cero Crossing Rate
- Anzahl der Vorzeichenwechsel in einem Frame
- diskreter Fall:

$$\mathrm{ZCR}[t] = \frac{1}{L}\sum_{\tau=1}^{L}\left|\mathrm{sign}(x'_t[\tau+1]) - \mathrm{sign}(x'_t[\tau])\right|$$

Nullkreuzungsrate Beispiele

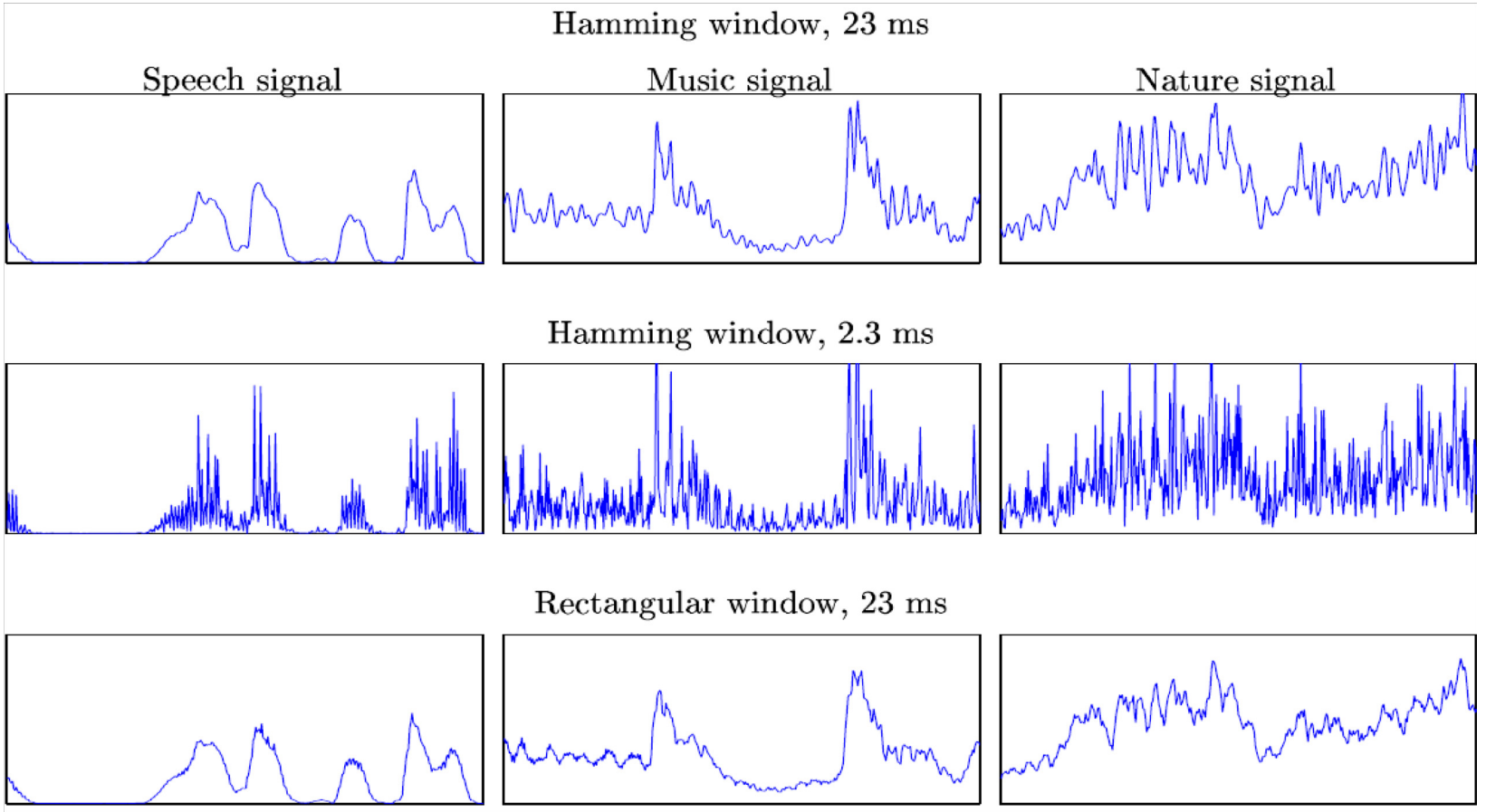


Probleme

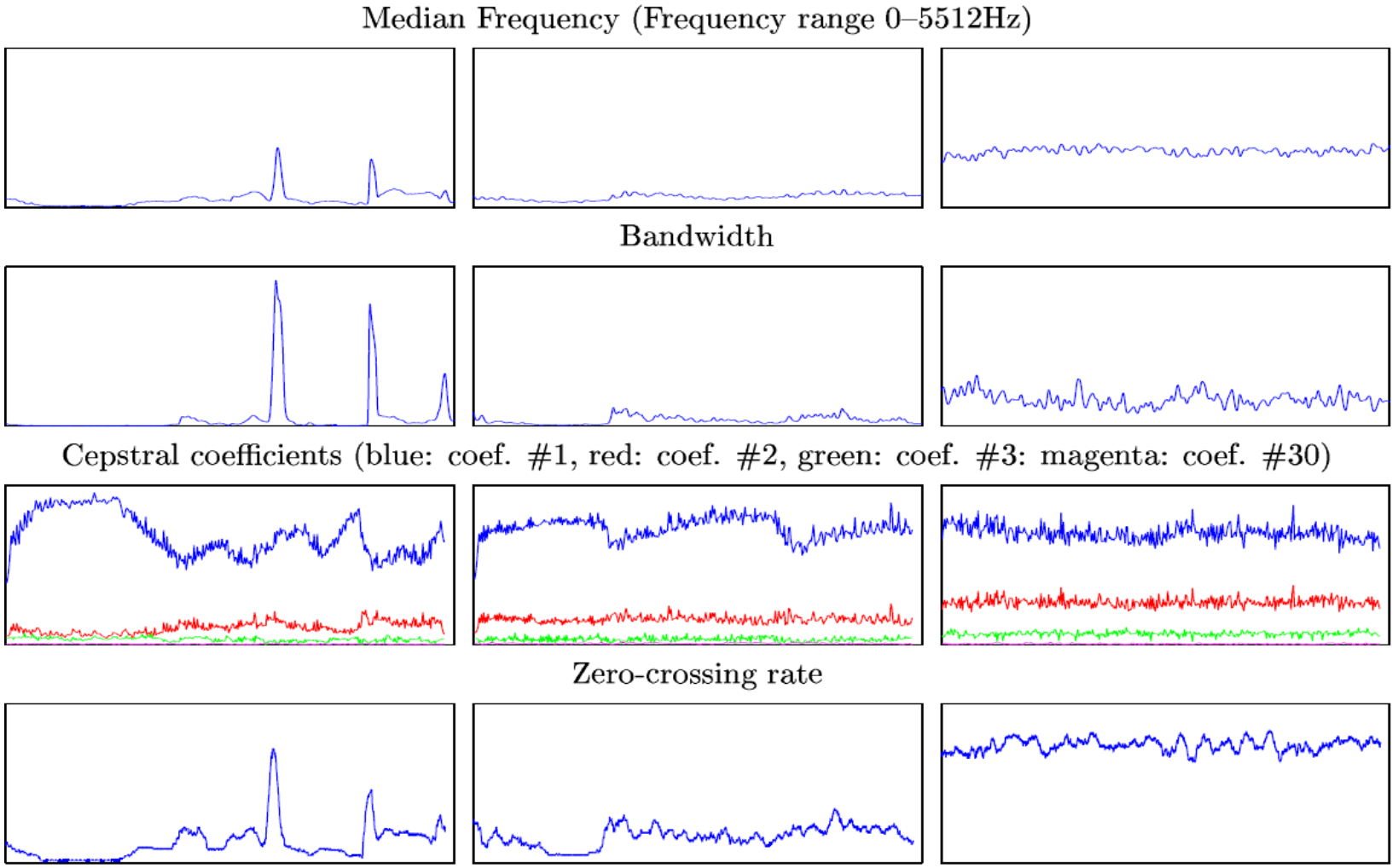
außer Vorzeichenwechselrate alle diskutierten Features verwandt mit Spektrogramm

- Wahl von Fenster und Fensterlänge hat **entscheidenden Einfluß** auf Merkmale (siehe nächste Folie)
- redundante Features (siehe übernächste Folie)

Beispiel Berechnung Volume



Vergleich von Merkmalen



IV.5
Distanz und Ähnlichkeit

II.5.1 Distanzfunktionen

Einleitung

- Distanzfunktionen vergleichen die Merkmalswerte zweier Medienobjekte
- Invarianz—drückt aus, welche Merkmale zum Vergleich nicht herangezogen werden sollen
- Bsp.: $d(g(o_1),g(o_2)) = d(o_1,o_2)$, d invariant gegenüber g

Definitionen

eine binäre Funktion

$$d : O \times O \longrightarrow \mathbb{R}_0^+$$

mit den Eigenschaften

- *Selbstidentität (Si):* $\forall o \in O : d(o, o) = 0$
- *Positivität (Pos):* $\forall o_1 \neq o_2 \in O : d(o_1, o_2) > 0$
- *Symmetrie (Sym):* $\forall o_1, o_2 \in O : d(o_1, o_2) = d(o_2, o_1)$
- *Dreiecksungleichung (Dreieck):*

$$\forall o_1, o_2, o_3 \in O : d(o_1, o_3) \leq d(o_1, o_2) + d(o_2, o_3)$$

nennen wir **Distanzfunktion**,

die Kombination mit Trägermenge *O* **Metrik**.

Definition

Klasse	Si	Pos	Sym	Dreieck
Distanzfunktion	✓	✓	✓	✓
Pseudo-Distanzfunktion	✓	–	✓	✓
Semi-Distanzfunktion	✓	✓	✓	–
Semi-Pseudo-Distanzfunktion	✓	–	✓	–

Distanzfunktionen

- einfache Distanzfunktion

$$d_{abs} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_0^+, d_{abs}(r_1, r_2) \mapsto |r_1 - r_2|$$

- euklidische Distanzfunktion

$$d_{L_2} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_2}(p_1, p_2) \mapsto \sqrt{\sum_{i=1}^{n} (p_1[i] - p_2[i])^2}$$

Minkowski-Distanzfunktion

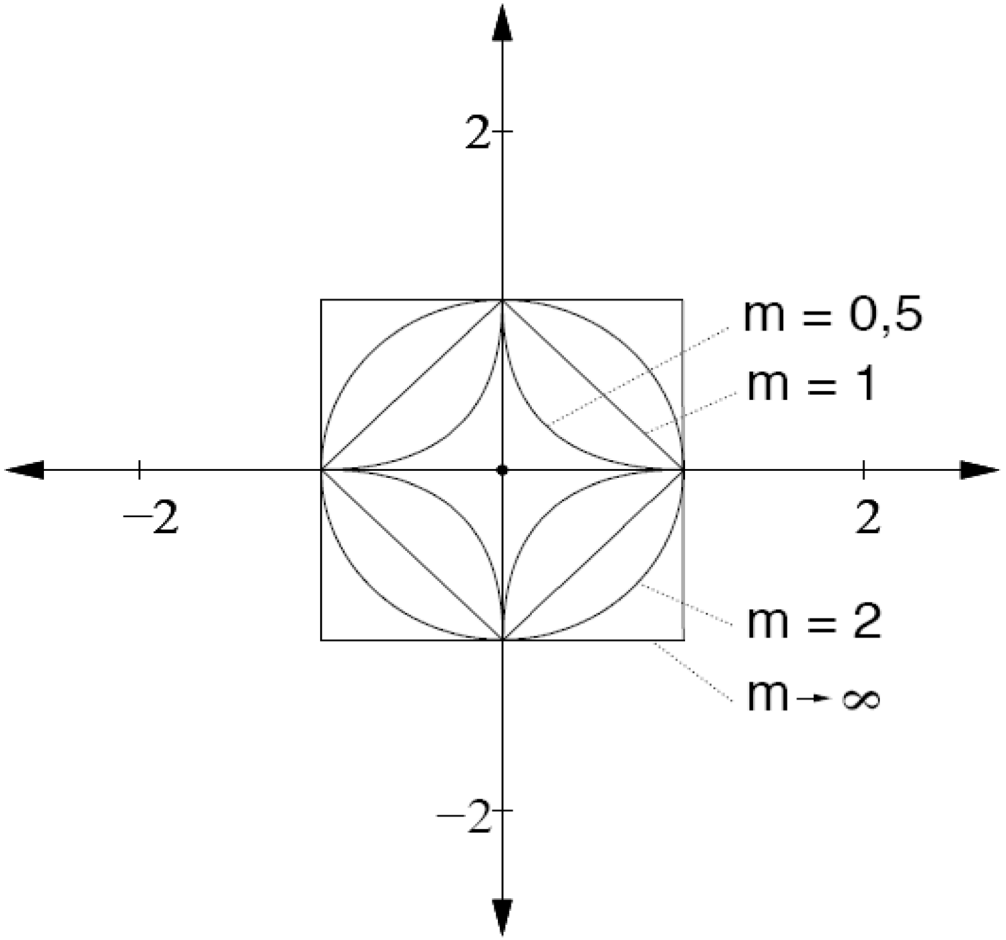
- Minkowski-Distanzfunktion (Punkte)

$$d_{L_m} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^{n} |p_1[i] - p_2[i]|^m \right)^{1/m}$$

- Werte für *m*
  - *m* = 1: Manhattan- oder Blockdistanz
  - *m* = 2: euklidische Distanz
  - *m* = unendlich: Max- oder Tschebyscheff-Distanz
- translationsinavariant, aber nicht rotations- (Ausnahme: *m* = 2) und skalierungsinvariant

„*m*-Einheitskreise“

Distanz zwischen 2 Punkten umso größer, je kleiner *m*



Gewichtete Minkowski-Distanz

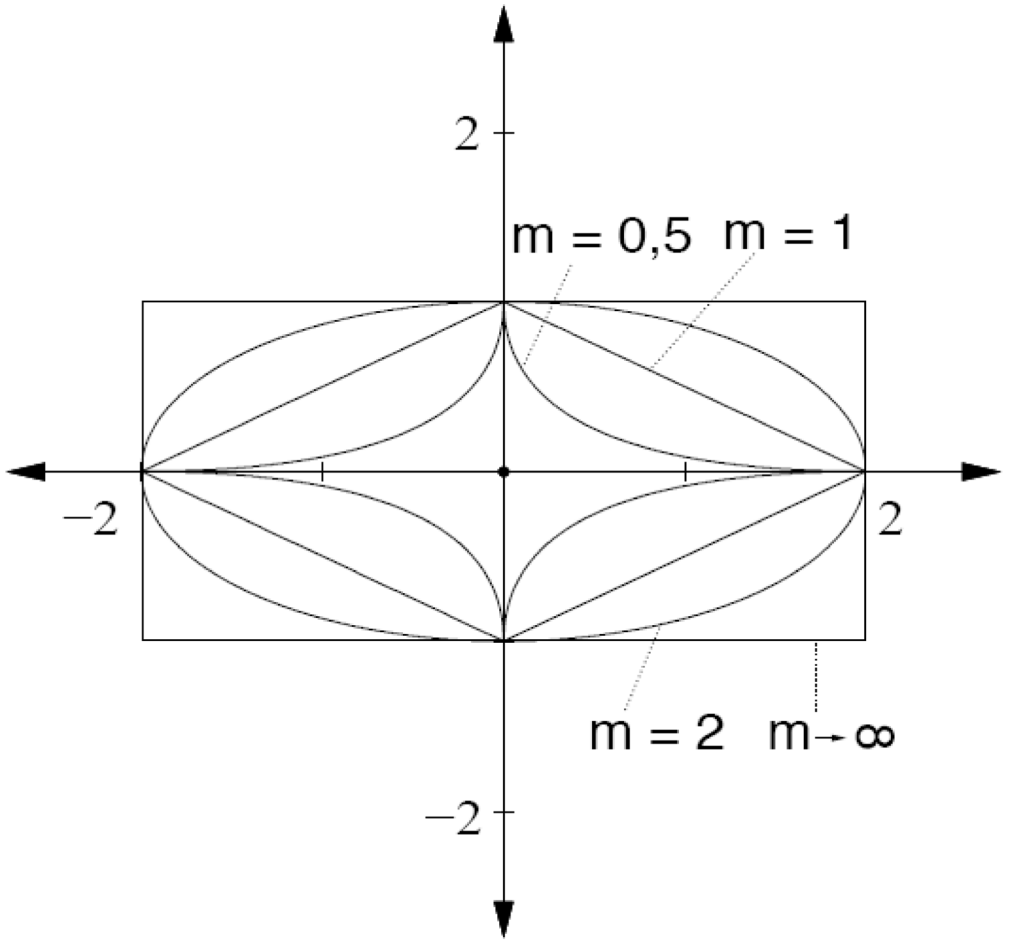
- häufig verwendete Variante

$$d^w_{L_m} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^+_0, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^{n} w_i * |p_1[i] - p_2[i]|^m \right)^{1/m}$$

- oft zusätzlich

$$\sum_{i=1}^{n} w_i = 1$$

Einheitskreise



Quadratische Distanz

Erweiterung der gewichteten euklidischen Distanz durch Drehung

$$d_q(p_1, p_2) = (p_1 - p_2)^T * A * (p_1 - p_2)$$

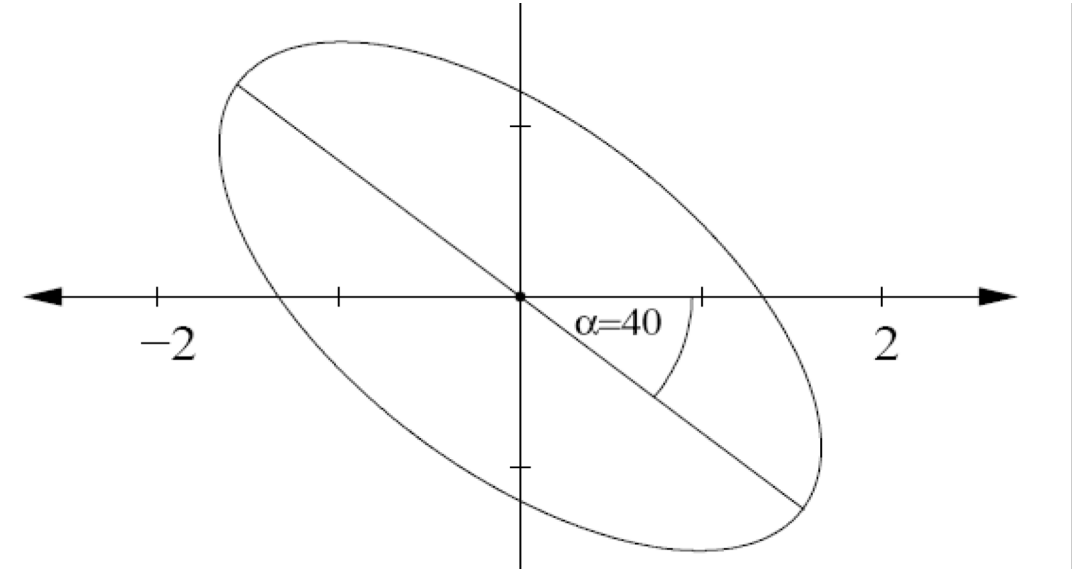
parametrisierbar über die Wahl von *A*:

- **Einheitsmatrix**: quadrierte euklidische Distanz
- **Diagonalmatrix**: quadrierte gewichtete euklidische Distanz
- **orthonormale** M.: Rotation bez. quadrierte euklidische Dist.
- **symmetrische** M.: identisch mit quadrierten euklidischen Distanz nach geeigneter Transformation

Beispiel

$$A = \begin{pmatrix} 0,5599 & 0,3693 \\ 0,3693 & 0,6901 \end{pmatrix}$$

$$= \begin{pmatrix} \cos 40 & \sin 40 \\ -\sin 40 & \cos 40 \end{pmatrix} * \begin{pmatrix} 0,25 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}$$



Mahalanobis Distanzfunktion

- quadratische Distanzfunktion basiert auf Kovarianzmatrix
- Kovarianzmatrix *C* hat folgenden Einfluß auf die Berechnung:

$$d_M(p_1, p_2) = |det\ C|^{1/d}(p_1 - p_2)^T * C^{-1} * (p_1 - p_2)$$

Quadratische Pseudodistanz

- Abstand 0 auch für nichtidentische Punkte
- unsymmetrische Translationsinvarianz, bei der Punkte um Vektoren *t* eines Unterraums *T* verschoben werden können:

$$pd_q(p_1, p_2 + t) = pd_q(p_1, p_2)$$

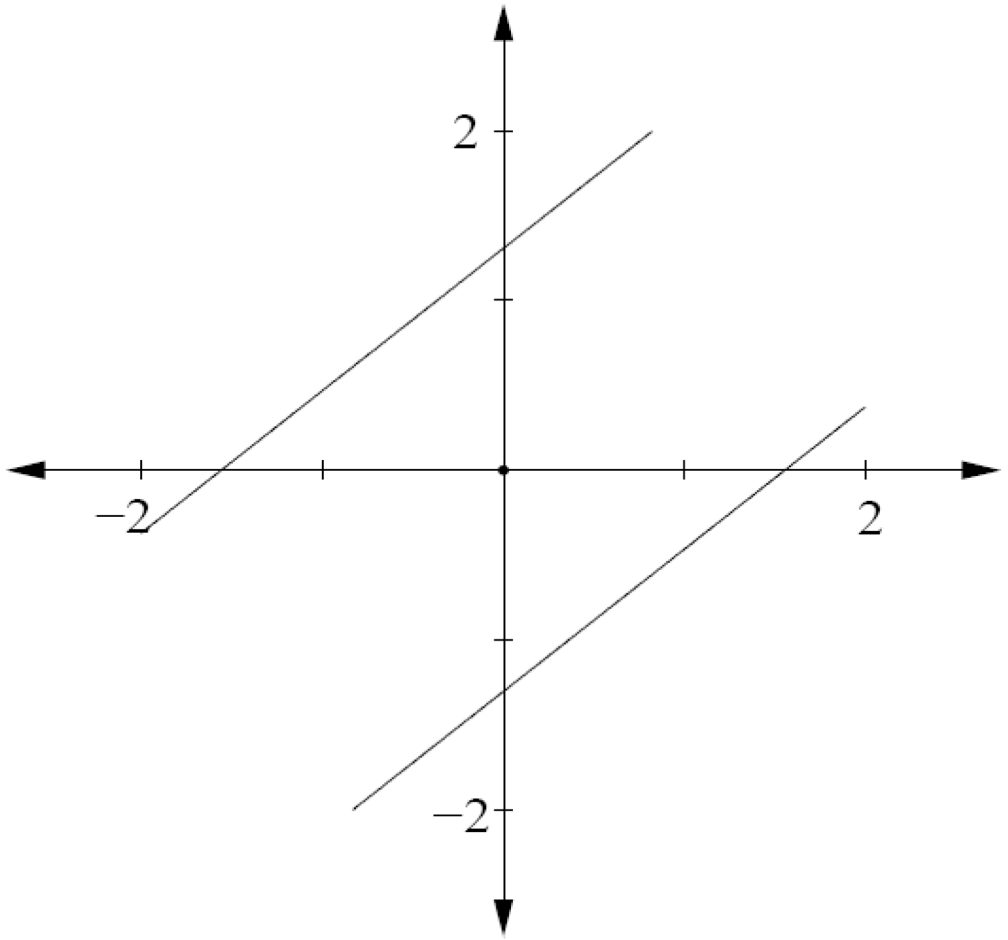
- Beispiel

$$U = \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$$U * L * U^T = \begin{pmatrix} 0,4132 & -0.4924 \\ -0,4924 & 0,5868 \end{pmatrix}$$

„Einheitskreis“

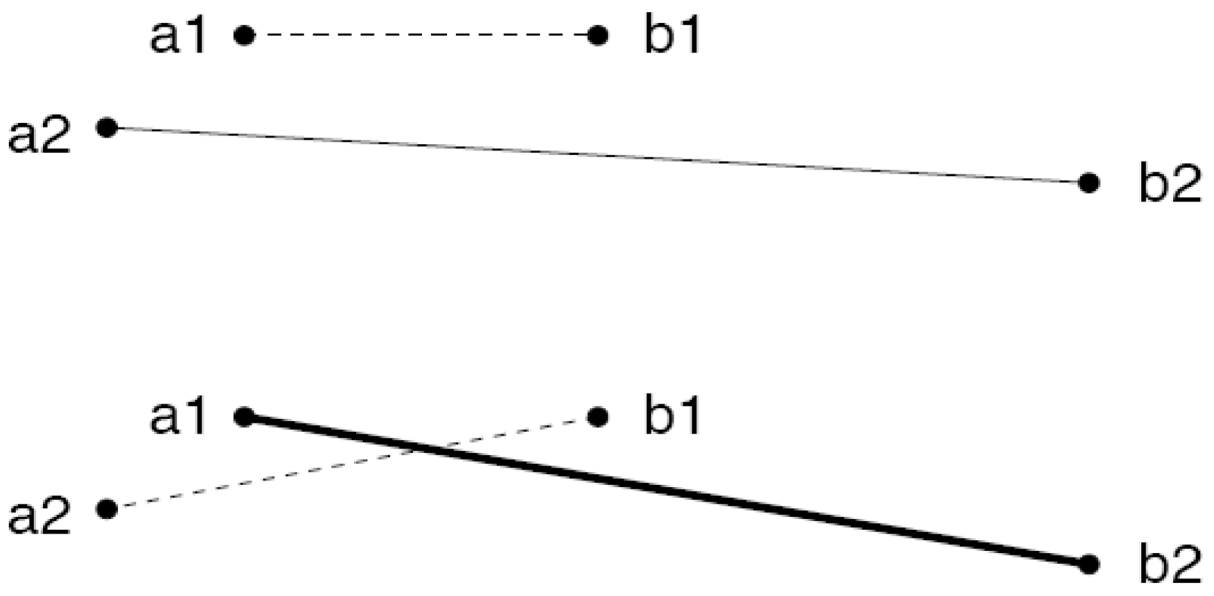


Bottleneck-Distanz

- Distanzfunktion auf Mengen
- Kardinalität der Mengen muß gleich sein
- sucht das Minimum der maximalen Elementepaardistanzen über alle möglichen Bijektionen *f* aus *F*(*A*, *B*):

$$d_B(A,B) = \min_{f \in F(A,B)} \max_{a \in A} d_X(a, f(a))$$

Beispiel



Übersicht

Distanzfunktion	Objekte	Si	Pos	Sym	Dreieck
Minkowski	Punkte	✓	✓	✓	✓
quadratisch (pos. def.)	Punkte	✓	✓	✓	✓
quadratisch	Punkte	✓	–	✓	✓
Dynamical-Partial	Punkte	✓	–	✓	–
Chi-Quadrat	Histogramm	✓	–	✓	–
Kullback-Leibler	Wahrsch.-Verteil.	✓	✓	–	–
Bhatacharayya	Wahrsch.-Verteil.	✓	✓	✓	–
Minkowski (binär)	Binärdaten	✓	✓	✓	✓
Hamming	Binärdaten	✓	✓	✓	✓
Größendifferenz	Binärdaten	✓	–	✓	–
Musterdifferenz	Binärdaten	✓	–	✓	–
Varianz	Binärdaten	✓	✓	✓	✓
Form	Binärdaten	✓	–	✓	–
Lance & Williams	Binärdaten	✓	✓	✓	–

Übersicht Invarianzen

Distanzfunktion	Objekte	Skal.	Rot.	Transl.
Minkowski	Punkte	–	–[13]	✓
quadratisch (pos. def.)	Punkte	–	–[14]	✓
quadratisch	Punkte	–	–	✓
DFT-L_2	Sequenzen	✓[15]	–	✓[16]

IV.5.2 Ähnlichkeitsmaße

Einführung

- Objekte werden als ähnlich wahrgenommen, wenn sie bei Menschen zu ähnlichen Reizen (Stimuli) führen
- keine allgemein akzeptierte Definition von Ähnlichkeit
- Ähnlichkeitsmodelle in Mathematik, Statistik, Bildverarbeitung und Mustererkennung
- **Ähnlichkeitsmaß**: Funktion, die einem Paar von Objekten eine relle Zahl aus [0,1] zuordnet
- Wert 1 korrespondiert mit maximaler Ähnlichkeit

Distanz vs. Ähnlichkeit

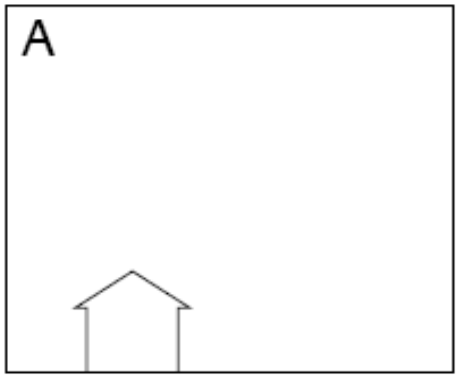
- viele Ansätze verwenden Distanzfunktion auf Featurewerten
- Distanzwerte werden auf [0,1] abgebildet
- Distanzeigenschaften für Ähnlichkeitsempfinden zu restriktiv (Untersuchungen in der Psychologie)
- bedeutet nicht automatisch, dass Distanzfunktionen für Ähnlichkeitsmaße ungeeignet sind
- nur nicht grundsätzlich für alle Anwendungen geeignet

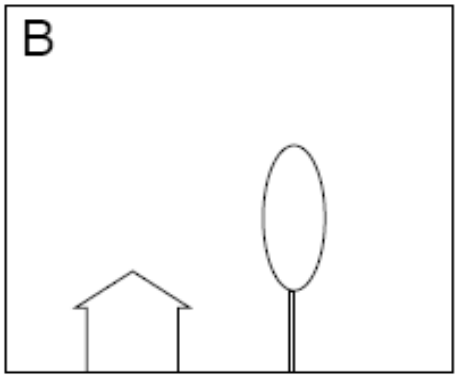
Probleme

- **Selbstidentität**: gilt nicht grundsätzlich [Krumhansl]
- **Positivität**: von Tversky als allgemeine Bedingung für menschliches Ähnlichkeitsempfinden widerlegt
- **Symmetrie**: Rollentausch macht Unterschied
- **Dreiecksungleichung**: Unterschiede zwischen 2 Objekten werden zu hoch bewertet, wenn kein drittes für den Vergleich vorliegt

Symmetrieprobleme

- Abfrage mit A als Suchbild vs. B als Suchbild

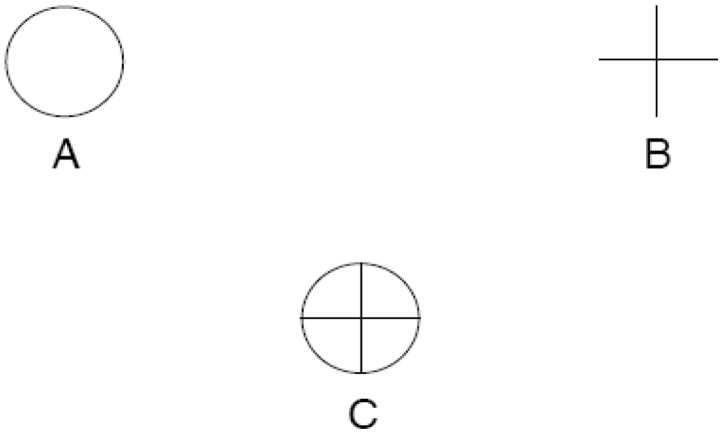




- **salient features**

Dreiecksungleichung

- Unähnlichkeit zwischen A und B wird i. a. stärker eingeschätzt als Summe der Unähnlichkeiten zu C



$$(d(A,B) > d(A,C) + d(B,C))$$

Ähnlichkeitsabstand

- Ähnlichkeitsabstand ist ein Unähnlichkeitsmaß
- Mindesteigenschaften [Tversky und Gati]:
  - **Dominanz**

$$d\left(\begin{pmatrix}x_1\\y_1\end{pmatrix},\begin{pmatrix}x_2\\y_2\end{pmatrix}\right) \geq \max\left\{d\left(\begin{pmatrix}x_1\\y_1\end{pmatrix},\begin{pmatrix}x_1\\y_2\end{pmatrix}\right), d\left(\begin{pmatrix}x_1\\y_1\end{pmatrix},\begin{pmatrix}x_2\\y_1\end{pmatrix}\right)\right\}$$

  - **Konsistenz**
  - **Transitivität**

Ähnlichkeitsabstand

Konsistenz

$$d\left(\begin{pmatrix}x_1\\y_1\end{pmatrix},\begin{pmatrix}x_2\\y_1\end{pmatrix}\right) > d\left(\begin{pmatrix}x_3\\y_1\end{pmatrix},\begin{pmatrix}x_4\\y_1\end{pmatrix}\right)$$

$$\Longleftrightarrow$$

$$d\left(\begin{pmatrix}x_1\\y_2\end{pmatrix},\begin{pmatrix}x_2\\y_2\end{pmatrix}\right) > d\left(\begin{pmatrix}x_3\\y_2\end{pmatrix},\begin{pmatrix}x_4\\y_2\end{pmatrix}\right)$$

Transitivität

$$d\left(\begin{pmatrix}x_1\\y\end{pmatrix},\begin{pmatrix}x_3\\y\end{pmatrix}\right) > \max\left\{d\left(\begin{pmatrix}x_1\\y\end{pmatrix},\begin{pmatrix}x_2\\y\end{pmatrix}\right), d\left(\begin{pmatrix}x_2\\y\end{pmatrix},\begin{pmatrix}x_3\\y\end{pmatrix}\right)\right\}$$

Ähnlichkeitsabstand

- Eigenschaften des Ähnlichkeitsabstandes sind allgemeiner als die Distanzeigenschaften (z. B. Symmetrie nicht gefordert)
- bei Anwendung einer monoton wachsenden Funktion auf Werte eines Abstandsmaßes bleiben Eigenschaften erhalten

Grenzen v. Ähnlichkeitsmaßen

- **Weltwissen** spielt bei menschlicher Ähnlichkeitsempfindung bewusst oder unbewusst eine Rolle
- Ebenen der Inhaltsverarbeitung
  - syntaktische Ebene (ohne Bedeutung d. Objekte)
  - semantische Ebene (Ähnlichkeitsvergleich)
  - pragmatische Ebene (Interpretation, thematische Kategorien)

Pre-Attentive vs. Attentive

- bei menschlicher Wahrnehmung von Reizen wird die **pre-attentive** von der **attentiven** Wahrnehmung unterschieden
- pre-attentative Wahrnehmung in den ersten 250 ms; ohne Interpretation (Weltwissen)
- Features der pre-attentiven Phase
  - Linienorientierung
  - Länge, Breite, Größe von Objekten
  - Krümmung
  - Anzahl von Objekten
  - Farbe und Intensität von Objekten

Ähnlichkeitsmaße

- viele verschiedene Funktionen und Maße vorgeschlagen
- viele Kombinationen
- leider keine allgemein anerkannte Kombination
- viele Alternativen zur Auswahl

Feature-Kontrast Modell

- Für ein Ähnlichkeitsmaß s(a,b) zwischen Objekten a, b auf Grundlage der korrespondierenden Eigenschaftsmengen A und B gelten [Tversky]:
  - **Matching**

$$s(o_1, o_2) = f(A \cap B, A \setminus B, B \setminus A)$$

  - **Monotonie**

$$s(a,b) \geq s(a,c) \text{ gdw.}$$
$$A \cap C \subseteq A \cap B, \quad A \setminus B \subseteq A \setminus C, \quad B \setminus A \subseteq C \setminus A$$

Feature-Kontrast Modell

- **Unabhängigkeit**
  - **Übereinstimmung** — *f*(*X*,*Y*,*Z*) sei ein Ähnlichkeitsmaß mit $X = A \cap B$, $Y = A - B$ und $Z = B - A$. Wir schreiben *V* ~ *W*, wenn *X*, *Y* und *Z* existieren, für die eine oder mehrere der Bedingungen gelten:

$$\begin{aligned} f(V,Y,Z) &= f(W,Y,Z) \\ f(X,V,Z) &= f(X,W,Z) \\ f(X,Y,V) &= f(X,Y,W) \end{aligned}$$

Feature-Kontrast Modell

- Übereinstimmung (Forts.) — Zwei Objektpaare (*a*,*b*) und (*c*,*d*) stimmen in einer (2 oder 3) Komponenten überein, wenn gelten:

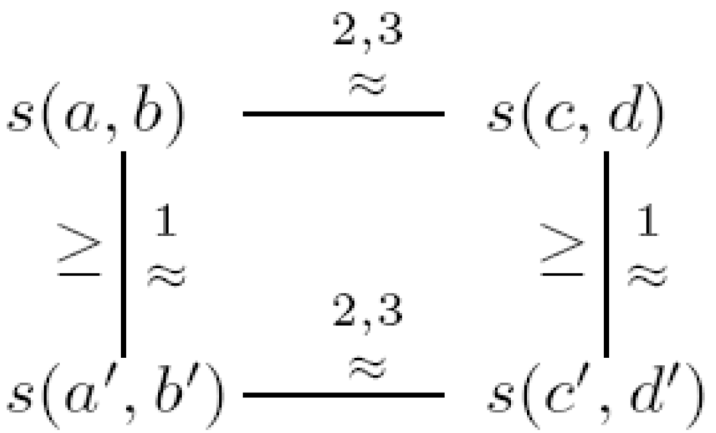
$$
\begin{aligned}
(A \cap B) &\approx (C \cap D) \\
(A \setminus B) &\approx (C \setminus D) \\
(B \setminus A) &\approx (D \setminus C)
\end{aligned}
$$

Unabhängigkeit

- Angenommen die Paare (*a*,*b*) und (*c*,*d*) sowie die Paare (*a'*,*b'*) und (*c'*,*d'*) stimmen in den selben 2 Komponenten überein,
- während die Paare (*a*,*b*) und (*a'*,*b'*) sowie die Paare (*c*,*d*) und (*c'*,*d'*) in der übrigbleibenden Komponente übereinstimmen.
- für **Unabhängigkeit** muss dann gelten:

$$s(a,b) \geq s(a',b') \iff s(c,d) \geq s(c',d')$$

Unabhängigkeit



Repräsentationssatz

- *s* sei ein Ähnlichkeitsmaß, für das Matching, Monotonie und Unabhängigkeit erfüllt sind.
- Es existiert dann eine Ähnlichkeitsfunktion *S*, eine nichtnegative Funktion *f* sowie 2 Konstanten a,b >= 0, sodass für alle Objekte *a*,*b*,*c*,*d* gelten:

$$S(a,b) \geq S(c,d) \iff s(a,b) \geq s(c,d)$$

$$\text{und}$$

$$S(a,b) = f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A)$$

Nachteile FKM

- Abhängigkeit von der Eigenschaftsanzahl
  Korrektur durch Normierung
- Skalierung
- binäre Eigenschaftswerte
  Verbesserung durch Santini und Jain

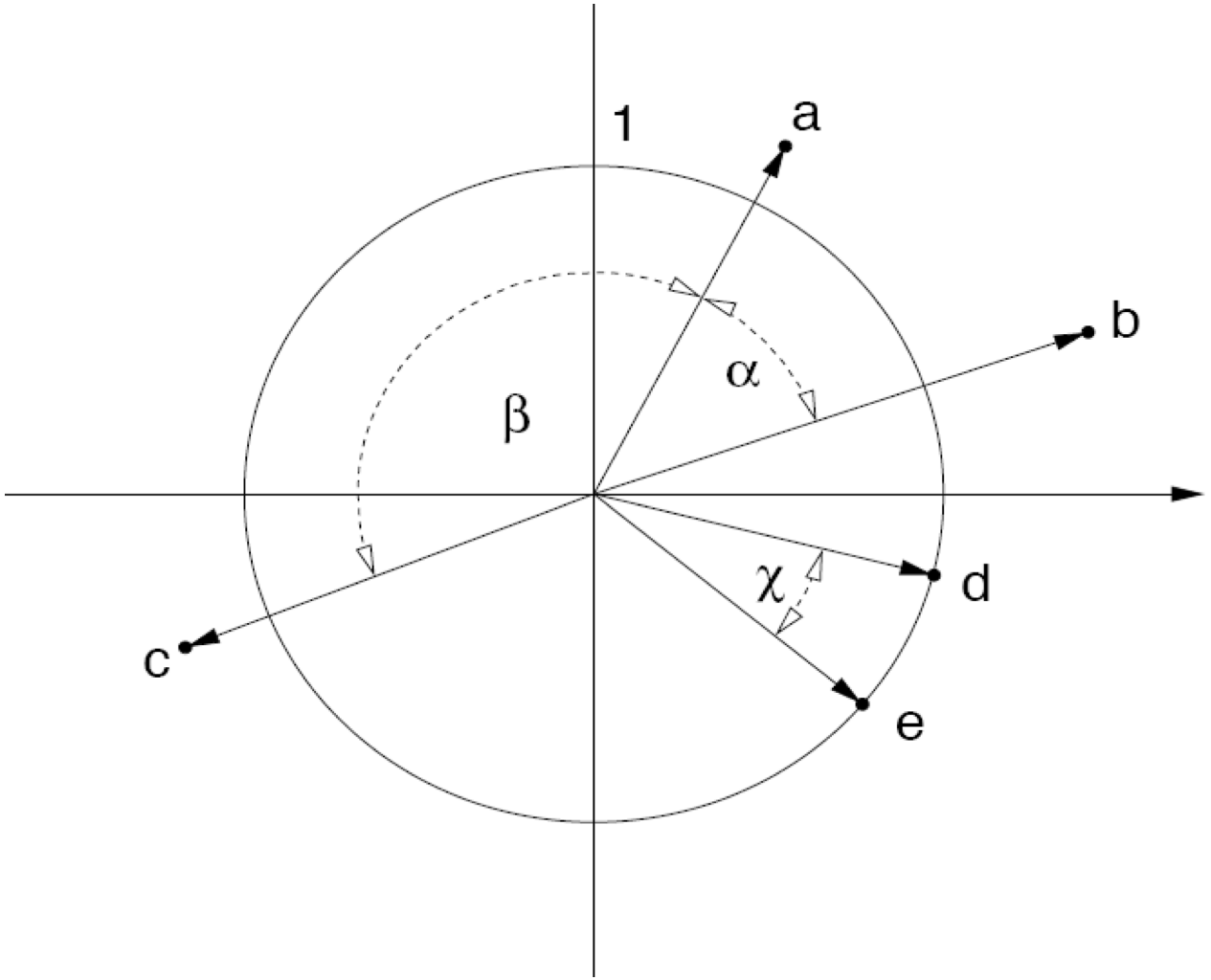
Kosinusmaß

- weit verbreitetes Maß
- Vektorraummodell
- Skalarprodukt von Vektoren

$$S_{cos}(a,b) = \frac{\langle a,b \rangle}{||a|| * ||b||}$$

- Intervallanpassung durch Halbierung und Addition von 0.5
- nichtnegative Werte — Intervallanpassung kann entfallen

Beispiel



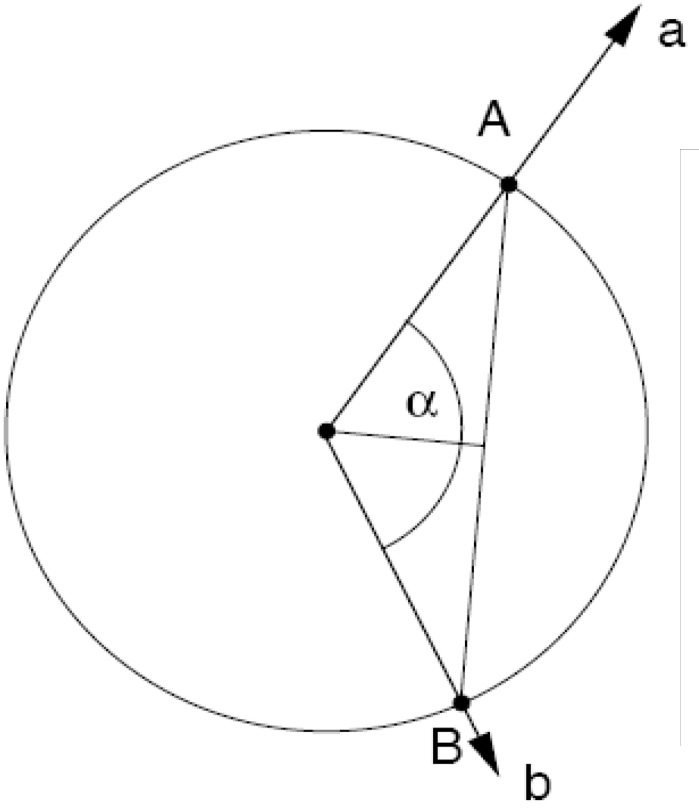
Abstandsfunktion

- um aus dem Ähnlichkeitswert einen Ähnlichkeitsabstand zu erzeugen:

$$d_{cos}(a,b) = 1 - S_{cos}(a,b) = 1 - \frac{\langle a,b \rangle}{||a|| * ||b||}$$

- Semi-Pseudo-Distanzfunktion:
  - Selbstidentität und Symmetrie
  - Positivität bahängig von Längenormierung
  - Dreiecksungleichung nicht erfüllt

Kosinusmaß — Alternative



$$
\begin{aligned}
d_{cos2}(a,b) &= ||A-B|| \\
&= 2 * \sin \alpha/2 \\
&= \sqrt{2*(1-\cos\alpha)} \\
&= \sqrt{2*(1-S_{cos}(a,b)} \\
&= \sqrt{2*\left(1-\frac{\langle a,b\rangle}{||a||*||b||}\right)}
\end{aligned}
$$

Aggregation v. Werten

- Anforderungen an eine Aggregatfunktion *agg*:
  - **Ähnlichkeitswerte** $agg : [0,1]^n \longrightarrow [0,1]$
  - **Monotonie**

$$x_1 \leq y_1 \wedge \ldots \wedge x_n \leq y_n \implies agg(x_1, \ldots, x_n) \leq agg(y_1, \ldots, y_n)$$

  - **strikte Monotonie**

$$x_1 < y_1 \wedge \ldots \wedge x_n < y_n \implies agg(x_1, \ldots, x_n) < agg(y_1, \ldots, y_n)$$

  - **Stetigkeit**
  - **Idempotenz**
  - **Unabhängigkeit von der Reihenfolge**

Generalisiertes Mittel

$$agg^{\alpha}_{gm}(x_1, \ldots, x_n) = \left(\frac{x_1^{\alpha} + \ldots + x_n^{\alpha}}{n} \right)^{1/\alpha}$$

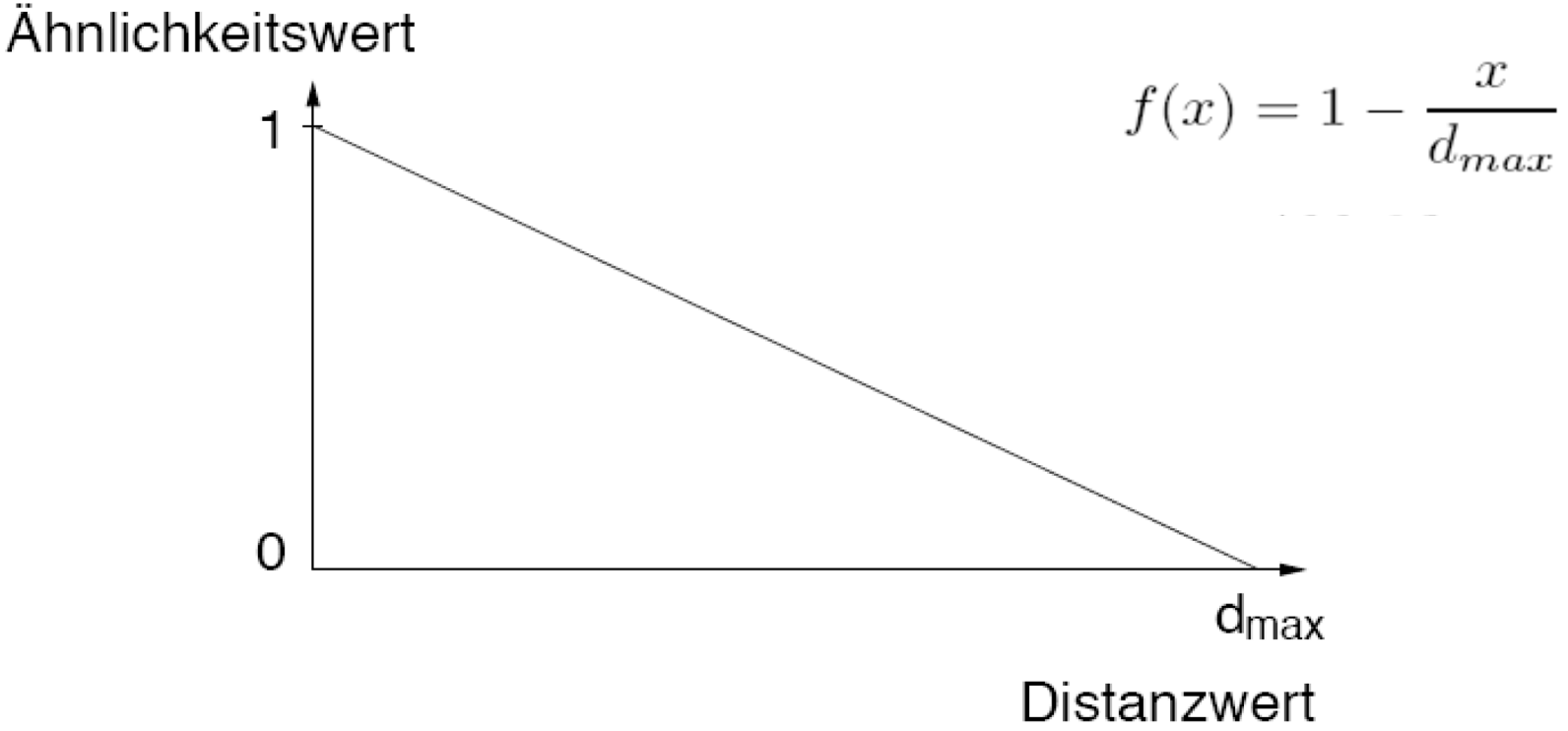
- a ungleich 0
- Spezialfälle
  - a = 1: arithmetisches Mittel
  - a unendlich: Maximum
  - a negativ unendlich: Minimum

Umwandlungsfunktionen

- Distanzen —> Ähnlichkeitswerte
- eine **Umwandlungsfunktion** ist eine Funktion *f*, die nichtnegative, reelle Werte auf [0,1] abbildet.
- Eigenschaften:
  - **Grenzbedingung max. Ähnlichkeit**: $f(0) = 1$
  - **Grenzbedingung min. Ähnlichkeit**: $f(d_{max}) = 0$
  - **streng monoton fallend**: $x_1 > x_2 \Rightarrow f(x_1) < f(x_2)$
  - **Stetigkeit**

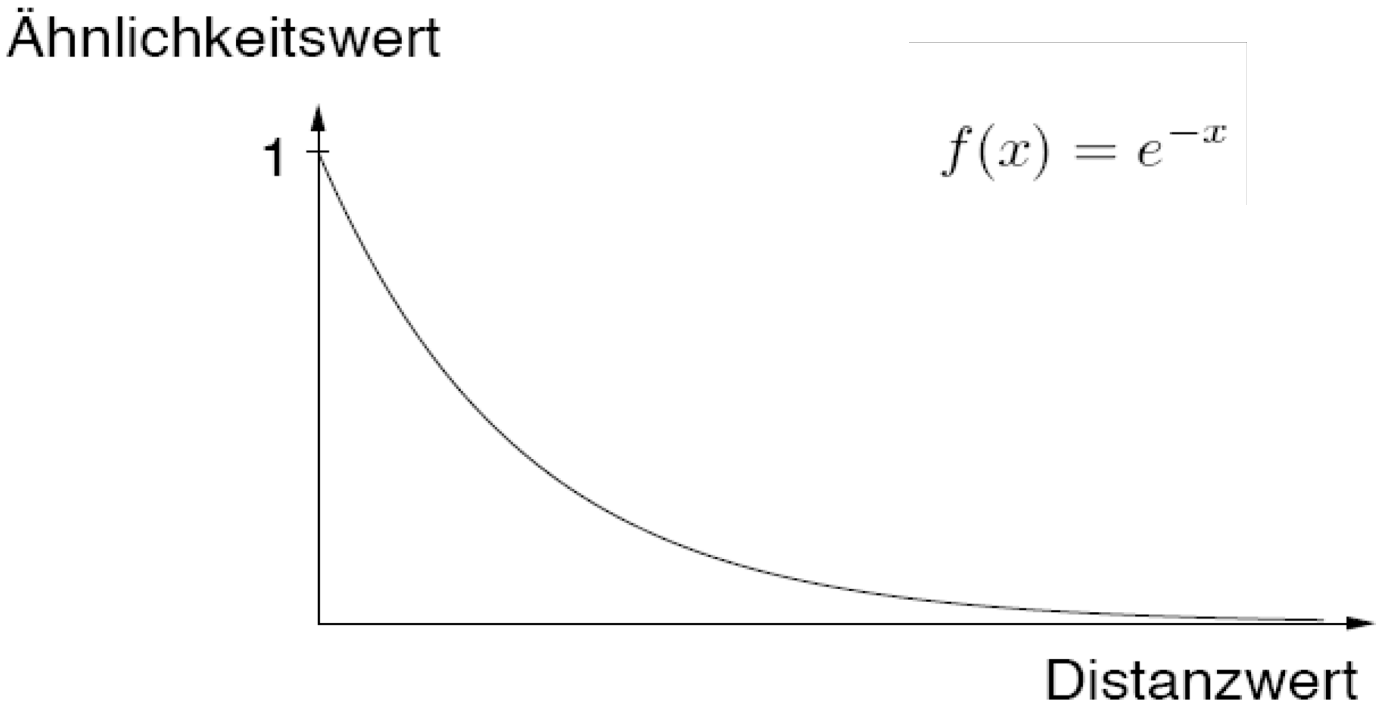
Umwandlungsfunktionen

- Linearkombination der Grenzbedingungen



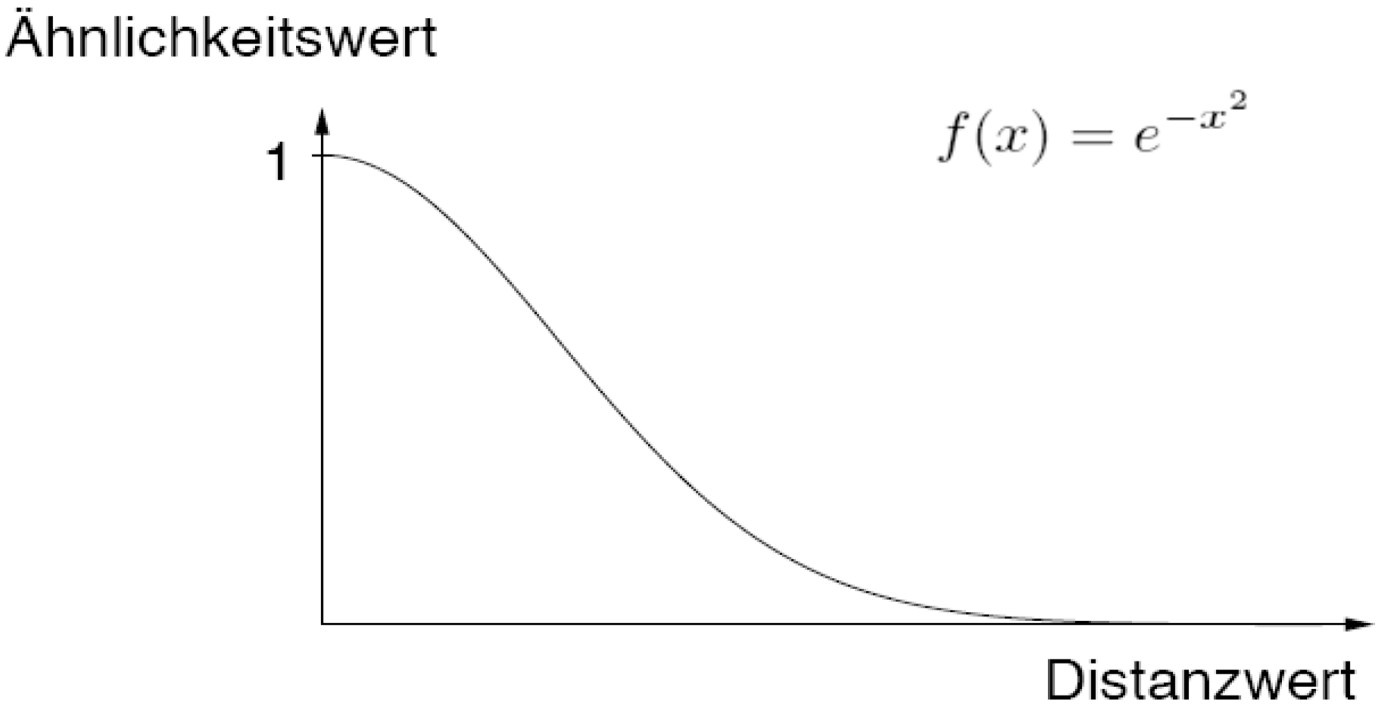
„Dynamische Sensibilität“

- hohe Sensibilität bei geringen Distanzen



Modifikation

- Abschwächung bei sehr geringen Distanzen

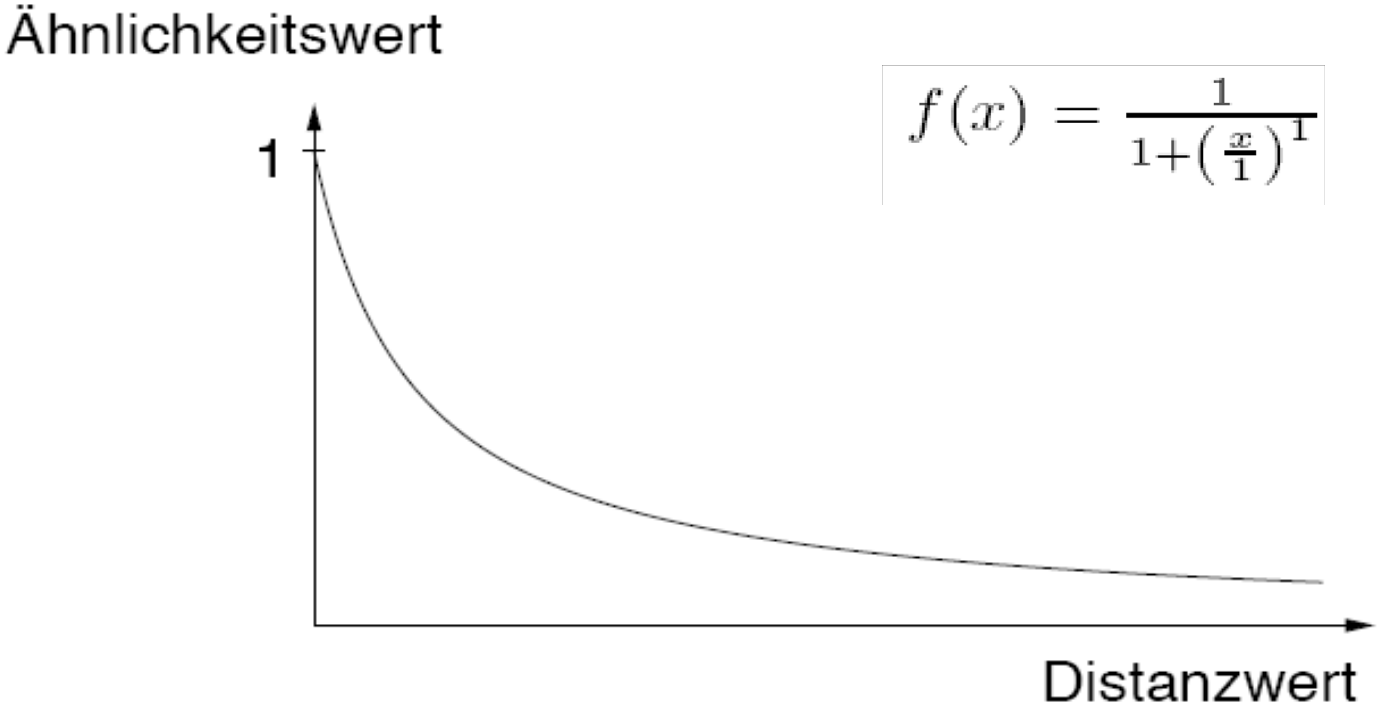


Parametrisierbare Funktion

- Beeinflussung der Sensibilität
  - *t* für Distanzwerte bei 0
  - *s* globlae Auswirkung

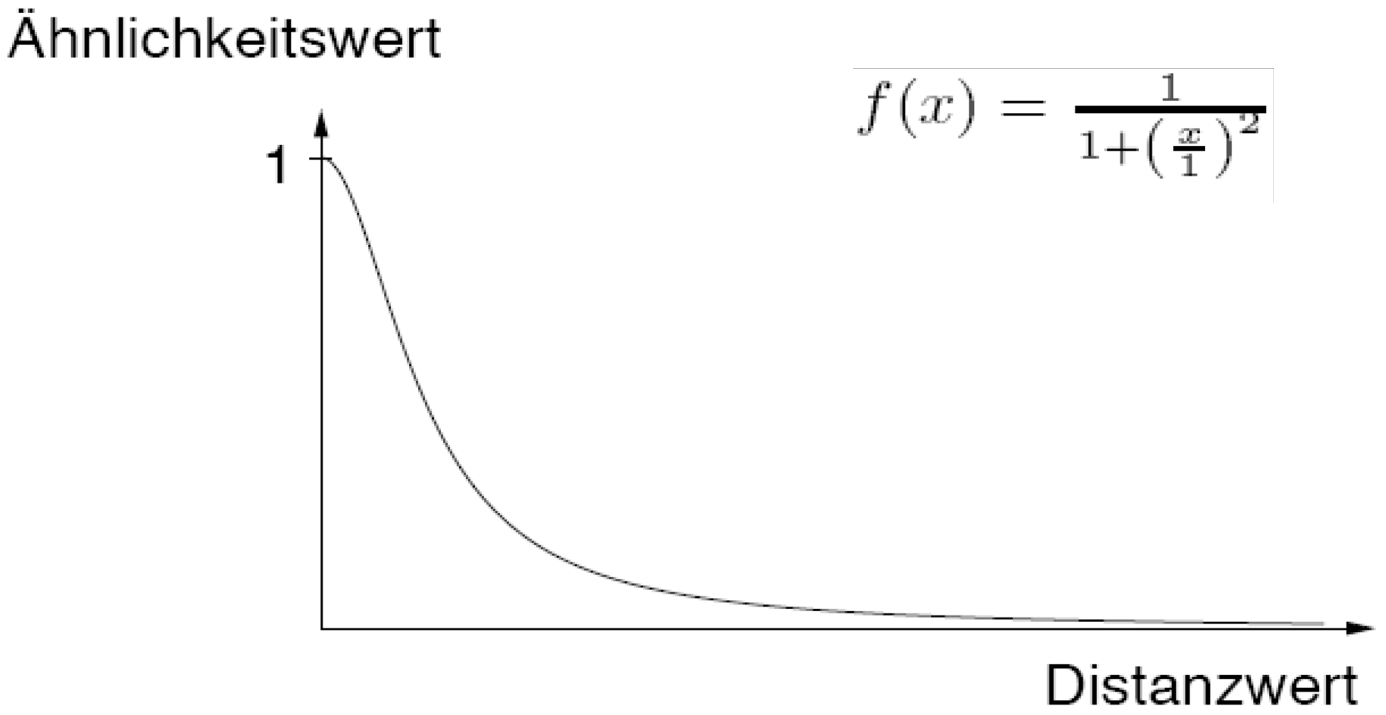
$$f(x) = \frac{1}{1 + \left(\frac{x}{s}\right)^t}$$

Beispiele



$$f(x) = \frac{1}{1+\left(\frac{x}{1}\right)^1}$$

Beispiele



$$f(x) = \frac{1}{1+\left(\frac{x}{1}\right)^2}$$

Beispiele