

# MULTIMEDIA VO

## Zusammenfassung

Version: 1.0 (Stoff vollständig abgedeckt)

Stand: 12.02.2016

Basierend auf den Folien vom WS 2015/2016

Vortragender: Prof. Breiteneder

---

Prof. Breiteneder war so freundlich, am Ende des Semesters eine Liste der nicht prüfungsrelevanten Folien bereitzustellen. Diese Folien wurden in der Zusammenfassung nicht berücksichtigt.

Kopie der Liste (zum Abgleich):

Prolog  
Folien 1-9

Kapitel I  
Folien 13-22, 44-49  
Folien 55-59, 65, 69, 72  
Folien 76-80, 83-86, 100-101, 104-107, 121-123,

Kapitel II  
Folien 129-132, 135-136, 181, 182,  
Folien 226-231, 248

Kapitel III  
Folien 299-322, 328,

Kapitel IV  
Folien 394-395, 409-415, 425  
Folien 431-436  
Folien 441-449,  
Folien 457-461  
Folien 520-522, 529-533  
Folien 545-555  
Folien 576-583  
Folien 609-611, 617-618  
Folien 623-627 (keine Details), 629, 643-645

Kapitel V  
Folien 663-664  
Folien 670-671, 676

Kapitel VIII  
ab 825

# Prolog

## Representation

- Zeitunabhängige (diskrete) Medien: Zeit spielt keine Rolle, z.B. Text, Graphiken
- Zeitabhängige (stetige) Medien: Zeitlich festgelegte Abfolge ist relevant; Verarbeitung ist zeitkritisch, z.B. Audio, Video

## I. Medientypen

### I.2. Bilder

#### Halbton-Approximation (Dithering)

Farb- und Helligkeitsabstufungen werden mittels Dot-Patterns angenähert. Der Mensch sieht das Pattern dann nur noch als einzelne Fläche. Pixel in einem Dot-Pattern sollten immer diagonal angeordnet sein um Artefakte zu verhindern (anstatt z.B. horizontal oder vertikal)

Cluster Dithering: Wahrgenommene Intensität wird durch die Größe des Dots bestimmt, dieser entspringt einem Mittelpunkt (siehe rechts). Verringert Kontureffekte und andere Artefakte.

Anwendung z.B. bei Druckern

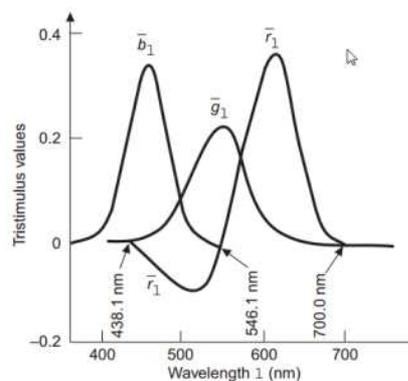
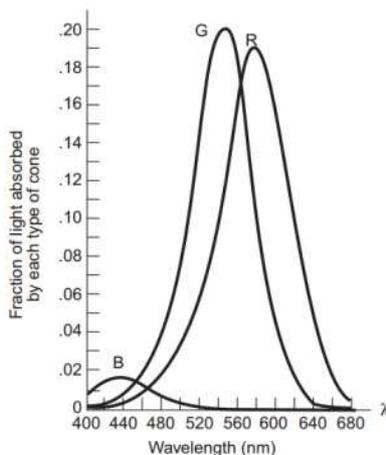


#### Farbe

- Verschiedene Modelle für verschiedene Ansprüche
- Ist immer subjektiv (abhängig von Person, Umgebung, etc.)
- Physikalisch: Farbe wird durch die Frequenz der elektromagnetischen Strahlen bestimmt, die auf die Retina treffen; Sichtbares Licht: 400-700nm Wellenlänge

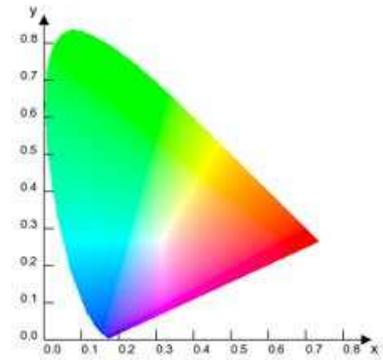
#### Tristimulus-Effekt

- Zapfen (Cones): 3 Arten, die für den blauen, grünen und roten Farbbereich zuständig sind
- Stäbchen (Rods): ermöglichen Sehen bei geringer Helligkeit (nur Graustufen)



## CIE-Farbraum

- Basiert auf dem Tristimulus-Effekt
- 3-dimensionaler Farbraum, kann 2-dimensional abgebildet werden.
- Am Außenrand befinden sich die reinen Spektralfarben, in der Mitte liegt Weiß. Komplementärfarben liegen genau gegenüber voneinander.
- Wird u.a. verwendet um andere Farbmodelle zu kalibrieren.



## RGB-Modell

- Additiv: Mischung von Rot, Grün und Blau ergibt Weiß
- Für Lichtabstrahlende Oberflächen, z.B. Computerbildschirme

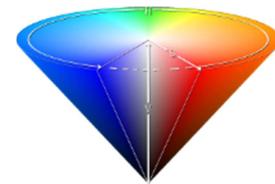
## CMY-Modell

- Subtraktiv: Mischung von Cyan, Magenta und Yellow ergibt Schwarz
- Für reflektierende Oberflächen, z.B. Print
- CMYK: Erweiterung durch vierte Farbe (Schwarz), v.a. bei Druckern

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} C \\ M \\ Y \end{bmatrix}$$

## HSV/HSB-Modell

- Hue (Farbe), Saturation (Sättigung), Value/Brightness (Helligkeit)
- Wird als Kegel dargestellt (siehe rechts)



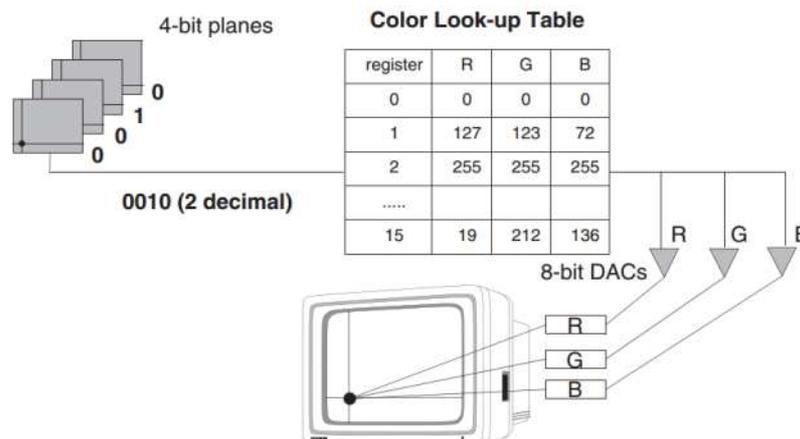
## YUV-Modell

- Luminance (Lichtstärke), U/V (Farbdifferenzwerte)
- Anwendung: TV

## Bildeigenschaften

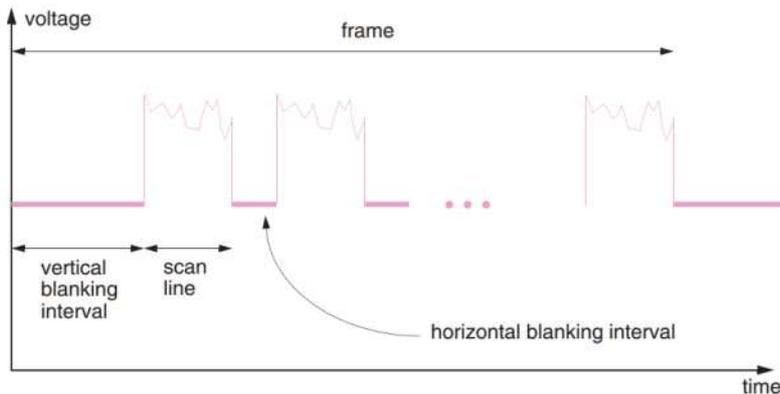
- Anzahl der Kanäle: Dimension des Farbmodells + Alphakanal (Transparenz)
- Channel depth: benötigte Bits pro Pixel
- Pixel Aspect Ratio: Pixelseitenverhältnis
- Interlacing: Bild wird in mehreren Schichten abgespeichert, schneller Aufbau eines Übersichtsbildes möglich
- Kompression: verlustlos bzw. -behaftet
- Indexing: Farben werden als Index in einer Color Map oder Color Lookup Table (CLUT) gespeichert (extern oder als Teil des Bildes)

## Color Mapping:



## I.2. Video

### Analoges Video-Signal:



### Interlaced Fields

- Zeilensprungverfahren
- Aufteilung in Bildzeilen mit gerader und ungerader Nummer („Halbbilder“)
- Bildwiederholrate kann verdoppelt werden, ohne zusätzliche Information zu übertragen → spart Bandbreite

### Analogvideo – Eigenschaften

- Frame Rate: z.B. 25fps, 30fps
- Scan Lines: Anzahl der Interlacing-Zeilen
- Aspect Ratio: z.B. 4:3, 16:9
- Interlacing: z.B. 2:1, non-interlaced = progressive
- Signal Quality: consumer, professional, broadcast
- Signalübertragung: composite, component (siehe unten)
- Stability: „Editing Degradation“

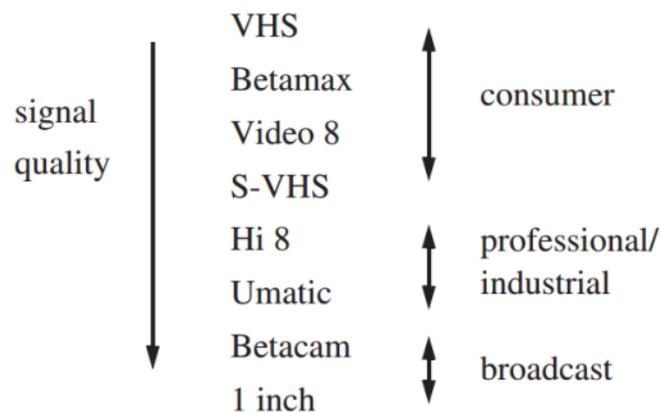
### Composite Video

- RGB/YUV-Informationen in einem einzigen Signal zusammengefasst
- Besser für Broadcasting, einfache Verkabelung, dafür schlechte Qualität

### Component Video

- Jede RGB/YUV-Information wird in eigenem Signal gesendet
- Bessere Qualität, dafür weniger geeignet für Broadcasting (Synchronisation nötig), kompliziertere Verkabelung

### Analoge Video-Tape-Formate:

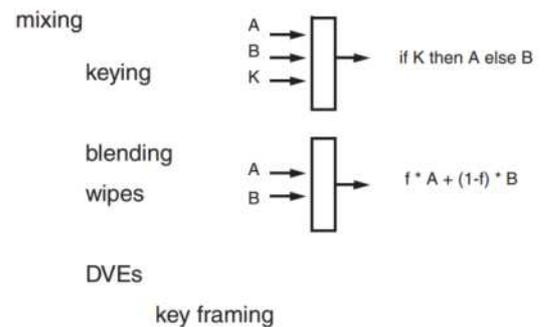


## Video-Equipment

- Routing Switcher (Matrix Switch): vereinfacht die Verkabelung zwischen Inputs und Outputs
- Distribution Amplifier: teilt 1 Input-Signal auf mehrere Outputs auf (ohne Qualitätsverlust)
- Timebase Corrector (TBC): rekonstruiert Signale, um Timing-Fehler zu vermeiden
- Sync Generator: Master-Clock-Signal synchronisiert verschiedene Geräte
- Frame (delay) buffer: synchronisiert externe Quellen
- Video production switcher: siehe unten

## Video Production Switcher

- Vereint mehrere Input-Streams zu einem einzigen Output-Stream
- Special Effekte möglich, z.B. Schriften, Übergänge, Bild-in-Bild

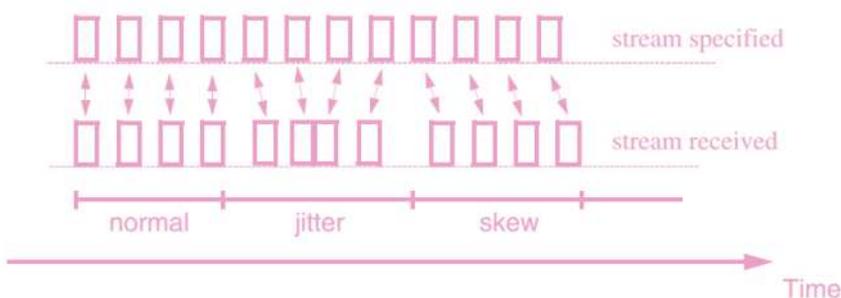


## Video Time Codes

- SMPTE (HH:MM:SS:FF)
  - Non-drop frame: FF-Wert läuft von 0 bis 29 durch
  - Drop frame: FF-Wert läuft von 2 bis 29, außer in jeder Zehntelminute 0 bis 29
- Time Code recorded on tape
  - LTC (longitudinal time code): gespeichert im Audio-Track
  - VITC (vertical interval time code): gespeichert in vertikalen Stanzen
  - RCTC (rewritable consumer time code): Sony Video-8

## Video-Synchronisation

Intra-flow-Synchronisation (skew, jitter):



## Digitales Video

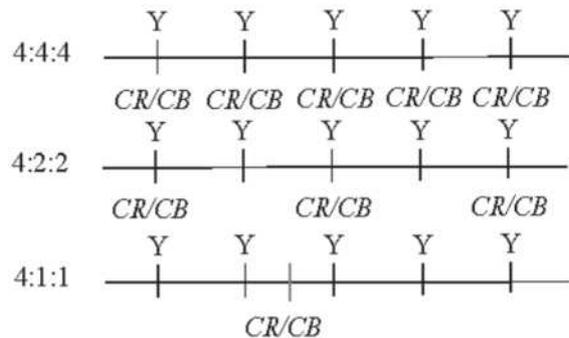
- Verwendet diskrete Zahlenwerte
- Signal wird abgetastet (sampling)
- Frame wird als Pixel-Array gespeichert

## CCIR 601

- Digital Component Video
- Durch Sampling eines analogen Component-Videosignals
- $m:n:l$  ... Multiplikatoren der Sampling-Frequenz (1, 2, 3 oder 4)

### Line Sampling:

- 4:2:2 – Broadcast-Qualität
- 4:1:1 – VHS-Qualität
- 4:2:0 - 2:1 – subsampling in horizontaler und vertikaler Richtung
- 4:1:1, 4:2:2 – MPEG und JPEG



## Video-Kompression

- Verlustlos bzw. -behaftet
- Echtzeit (symmetrisch): brauchen Kompression und Dekompression gleich lang?
- Räumlich bzw. zeitlich: hängt die Kompression auch von Vorgänger- und Nachfolgerframes ab?
- Skalierbar
- Abhängig vom Quellformat

## Einflussfaktoren auf die Videoqualität

- Frame size und depth
- Frame rate, key frame rate
- Algorithmus-Parameter, Dekodierungszeit
- Komprimierte Datenrate und Dateigröße

## I.3 Audio

### Was ist Schall?

- Wellenförmige Ausbreitung von Druckschwankungen

### Mathematische Beschreibung

- Frequenz  $f$  [Hz]: Schwingungen pro Sekunde, Tonhöhe
- Periodendauer  $\tau=1/f$  [s]: Dauer einer Schwingung
- Amplitude  $p_0$  [Pa]: Druckwert bei maximaler Kompression, Lautstärke
- Wellenlänge  $\lambda$  [m]: während einer Periodendauer zurückgelegter Weg
- Schallgeschw.  $v$  [m/s]: abh. vom Ausbreitungsmedium, Luft ca. 330 m/s

### Hörphänomene

- Beugung: Wellen „biegen“ sich um die Ecke
- Interferenz: Überlagerung der direkten und reflektierten Wellen → Lokalisierung der Schallquelle
- Brechung: Wellen ändern die Richtung, wenn sie verschiedene Medien durchlaufen
- Dispersion: Für verschiedene Frequenzen ist die Stärke der Brechung unterschiedlich

## Ton, Klang, Geräusch

- Ton: einzelne Sinusschwingung
- Klang: Überlagerung von Grundton (bestimmt Tonhöhe) und Obertönen (bestimmen Klangfarbe)
- Geräusch: Kein Grundton erkennbar

## Frequenzspektrum

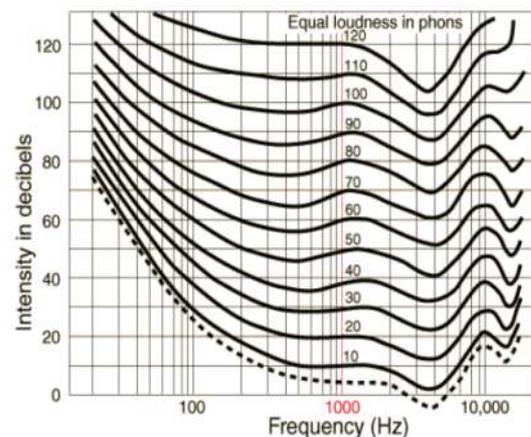
- Jede periodische Schwingung (Klang) kann als eine Überlagerung von Sinusschwingungen dargestellt werden. Daraus lässt sich ein Diagramm mit allen vorkommenden Frequenzen und der jeweiligen Amplitude erstellen.
- Diskretes Spektrum: nur ganzzahlige Vielfache zur Grundfrequenz
- Kontinuierliches Spektrum: unendliche Anzahl an Einzelschwingungen (Geräusche)

## Schalldruckpegel

- $L = 20 \cdot \log(p/p_0)$
- L ... Schalldruckpegel [dB]
- p ... Schalldruck der Welle [Pa]
- $p_0$  ... Bezugsschalldruck: Schalldruck eines 1000Hz-Tones, der gerade noch hörbar ist [Pa]
- Hörbereich ist sehr groß → Abbildung auf kleineren Bereich sinnvoll
- Hörschwelle = 0 dB, Schmerzgrenze = 120 dB
- Menschliches Lautstärkeempfinden entspricht ebenfalls einer logarithmischen Skala

## Kurven gleicher Lautstärke

- Töne mit verschiedenen Frequenzen werden bei gleichem Schalldruck unterschiedlich laut wahrgenommen.
- KGL geben an, wie hoch im Vergleich zum Pegel eines 1000Hz-Tones der Schallpegel eines Tones sein muss, damit dieser gleich laut empfunden wird.



## Lautstärkepegel

- $L_N$  ... Lautstärkepegel [Phon]
- ist gegeben durch den Schalldruckpegel L eines 1000Hz-Tones, der als gleich laut gehört wird
- Wird aus den Kurven gleicher Lautstärke ausgelesen
- Bsp: 100Hz-Ton mit  $L=60\text{dB}$  → 50 Phon

## Grundlagen Digitalisierung

- Input: zeit- und spannungskontinuierliches Signal
- Abtasten / Quantisieren:
  - Bei Audio: zuerst wird abgetastet, danach quantisiert
  - Bei Video: zuerst wird quantisiert, danach abgetastet
- Output: zeit- und spannungsdiskretes Format

## Abtasten

- In periodischen Zeitintervallen wird dem analogen Eingangssignal ein Wert (Sample) entnommen
- Nyquist-Shannon-Theorem:
  - $f_s \geq 2 * f_{\max}$
  - $f_s$  ... Abtastfrequenz
  - $f_{\max}$  ... höchste im Signal vorkommende Frequenz
  - Signale mit einer höheren Frequenz als  $f_{\max}$  können nicht korrekt rekonstruiert werden → es entsteht ein Signal mit falscher Frequenz (Aliaskomponente)

## Quantisierung

- Abgetastete Spannungswerte werden in diskrete Zahlenwerte umgewandelt (gerundet)
- Beim Runden der Werte entsteht ein Quantisierungsfehler → Quantisierungsrauschen
- Dynamikbereich (Signal Noise Ratio): kann durch Erhöhen der Bit depth vergrößert werden
  - $SNR = L_{\max} - L_{\text{noise}}$
- Q ... Größe der Quantisierungsintervalle
- Lineare Quantisierung (für Audio): Q ist konstant
- Nichtlineare Quantisierung:
  - Quantisierungsintervalle sind verschieden groß
  - kleine Werte → kleiner Quantisierungsfehler u. u.
  - für Systeme mit geringer Übertragungsbandbreite

## Codierung

- Audiosignale sind bipolar (negative und positive Werte) → Darstellung als Zweierkomplement → Bei Addition zweier Signale entsteht kein Offset

## Pulscodemodulation (PCM)

- Binärwerte werden seriell als modulierte Spannungspulse über eine einzige Leitung übertragen
- Anwendung z.B. bei Audio-CDs
- Datenrate [bit/s] = Abtastfrequenz \* Bitauflösung
- Bandbreite [Hz] = Datenrate / 2 ... nötiger Frequenzbereich für verlustlose Übertragung

## Digitales Audio – Eigenschaften

- Sampling frequency: Abtasten
- Sample size: Quantisierung
- Number of Channels: z.B. 2 bei Stereo
- Interleaving: Daten für mehrere Kanäle werden hintereinander gespeichert statt separat; besser für Streams geeignet, höhere Hardwareanforderungen
- Sample representation: Negative Werte
- Coding/Compression

## MIDI

- Musical Instruments Digital Interface: Protokoll zur Kommunikation zwischen Computern, Synthesizern, Keyboards und anderen Musikgeräten
- Synthesizer: Sound-Generator, kann verschiedene Features und Bauweisen haben
- Sequencer: Speichert MIDI-Daten; ermöglicht das Erstellen von Loops und Kompositionen
- Track (Spur): zur Organisation von Aufnahmen
- Channel: Trennung von MIDI-Signalen, um verschiedene Geräte über ein Kabel ansprechen zu können
- Voice: Der generierte Sound eines Synthesizers; Ein Synthesizer kann auch mehrere Voices gleichzeitig ausgeben → komplexeres Klangbild
- Key Number: Jede spielbare Note wird durch eine Zahl repräsentiert
- Controller: Verändert Parameter eines MIDI-Geräts
- Patch / Program: bestimmt die Klangfarbe

## Wichtige MIDI-Konzepte

- Timing Clock: bestimmt die Abspielgeschwindigkeit eines Sequencers; Tempo: Beats pro Minute
- MIDI synchronization: synchronisiert verschiedene MIDI-Geräte untereinander
- MIDI Time Code (MTC): zur Synchronisation mit Video

# II. Kompression

## Einteilung

- Entropy Coding
  - Verlustfrei
  - unabhängig von Eigenschaften des Medienformats
  - Daten werden als einfache digitale Sequenz betrachtet
  - z.B.: run-length, Huffman, arithmetic
- Source Coding
  - Verlustbehaftet
  - Bezieht die Semantik der Daten mit ein
  - Stärke der Kompression ist abhängig vom Content
  - z.B.: prediction (DPCM, DM), transformation (FFT, DCT)
- Hybrid Coding
  - z.B.: JPEG, MPEG, px64

## Run-length Encoding

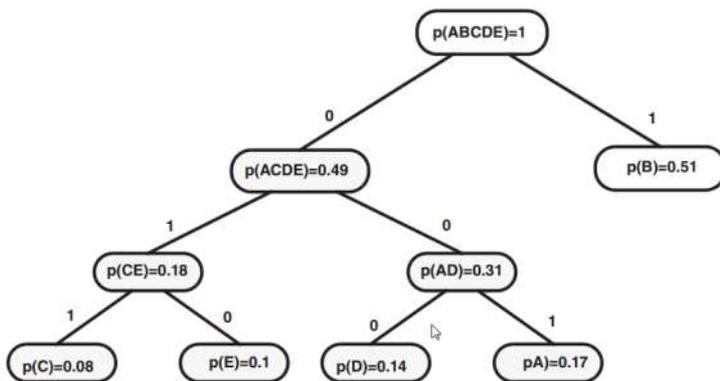
- Entropy Coding Algorithmus
- Ersetzt Sequenzen von wiederkehrenden Bytes (mind. 4 Wiederholungen) durch die Anzahl der Bytes
- Ersetzungen werden durch ‚!‘-Zeichen gekennzeichnet
- z.B.: ABCCCCCCCCCDEFFFFGGG → Codierung → ABC!9DEF!4GGG
- Variationen: zero suppression; text compression; diatomic encoding

## Statistical Encoding

- Frequency dependent encoding; gehört zu Entropy Encoding
- Für jedes vorkommende Symbol  $S_x$  wird die Häufigkeit  $p(S_x)$  ermittelt und danach die minimale Anzahl an Bits pro Symbol gesucht

## Huffman Encoding

- Gehört zu Statistical Encoding
- Anzahl der Bits pro Symbol variiert (häufigere Symbole erhalten kürzere Codes u.u.)
- Encoding-Bsp:
  - $p(A)=0.17$ ,  $p(B)=0.51$ ,  $p(C)=0.08$ ,  $p(D)=0.14$ ,  $p(E)=0.1$
  - Fertiger Huffman-Tree (Schritt-für-Schritt-Anleitung siehe Folien 141-145) und Huffman-Tabelle:



A	001
B	1
C	011
D	000
E	010

- Huffman Coding im Bild- und Videobereich:
  - Einzelne Pixel werden Zeile für Zeile als Bit-Stream interpretiert
  - Kann für einzelne Szenen oder das gesamte Video verwendet werden

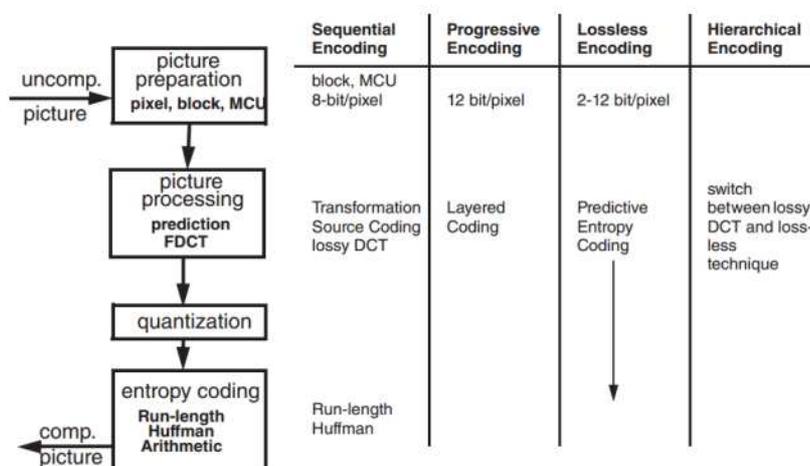
## Differential Encoding

- Source Coding
- Sequenz von Symbolen  $S_x$ , deren Werte nur wenig variieren (und nicht 0 sind) → Speichere nur den Unterschied zum vorigen Symbol
- Bei Bildern:
  - Berechnung der Differenzen von benachbarten Pixeln
  - Kleine Farb- und Helligkeitsunterschiede ergeben niedrige Werte
- Bei Videos:
  - Vorteil z.B. bei ruhigem Hintergrund (Nachrichtensendung, ...)
  - Motion Compensation: 8x8-Pixelblöcke werden verglichen und bei starker Ähnlichkeit nur verschoben (Bewegungsvektoren)

## JPEG – Joint Photographic Experts Group

### Anforderungen

- Hohe Kompression bei gleichzeitig hoher Bildqualität
- Parameter können vom User verändert werden
- Unabhängig vom Ursprungsformat (Abmessungen, Inhalt, Seitenverhältnis, ...)
- Niedrige Hardware- und Softwareansprüche, hohe Kompatibilität
- Unterstützung folgender Einsatzarten:
  - Sequential Encoding: selbe Codierungs- und Scanreihenfolge
  - Progressive Encoding: multiple pass encoding
  - Lossless Encoding
  - Hierarchical Encoding: mehrere Auflösungen



### Bildmodell

- Unabhängig von Bildparametern
- Ursprungsbild besteht aus bis zu 255 Komponenten (Ebenen)
- Innerhalb eines Bildes werden alle Pixel mit derselben Anzahl an Bits kodiert

### Bildvorbereitung

- Bild wird in 8x8-Pixelblöcke unterteilt, DCT (Diskrete Cosinus Transformation) arbeiten auf diesen Blöcken
- Verlustbehafteter Modus → arbeitet mit Blöcken; Verlustfreier Modus → arbeitet mit einzelnen Pixeln
- Abarbeitungsreihenfolgen:
  - von links nach rechts und von oben nach unten
  - interleaved: einzelne Komponenten werden als Minimum Coded Units (MCU) zusammengefasst
- Danach: unkomprimierte Bildteile werden an den JPEG Encoder übergeben

### Bildverarbeitung

- Pixelwerte werden in den Wertebereich [-128, 127] umgewandelt
- Umwandlung vom Zeit- in den Frequenzbereich durch Forward DCT
- $S(u,v)$ -Komponenten:
  - $S(0,0)$  ... „DC-Komponente“: niedrigste Frequenz in beide Richtungen; bestimmt Grundfarbe des Blocks
  - $S(0,1)$  bis  $S(7,7)$  ... „AC-Komponenten“: Andere Frequenzen in beide Richtungen

## Diskrete Cosinus-Transformation

- Tiefe Frequenzen: entsprechen großflächigen Farb- und Helligkeitsverläufen
- Hohe Frequenzen: entsprechen feinen Strukturen und harten Kanten; können stärker komprimiert werden, ohne viel Bildqualität zu verlieren
- Bilddaten müssen der Frequenz entsprechend sortiert werden
- Beispielbilder: siehe Folien 161-165
- Bei JPEG
  - Inverse DCT (I-DCT) bildet DCT-Koeffizienten auf die gesampelten Werte ab
  - DCT und I-DCT können nicht ohne Rundungsfehler berechnet werden → verlustbehaftete Kompression

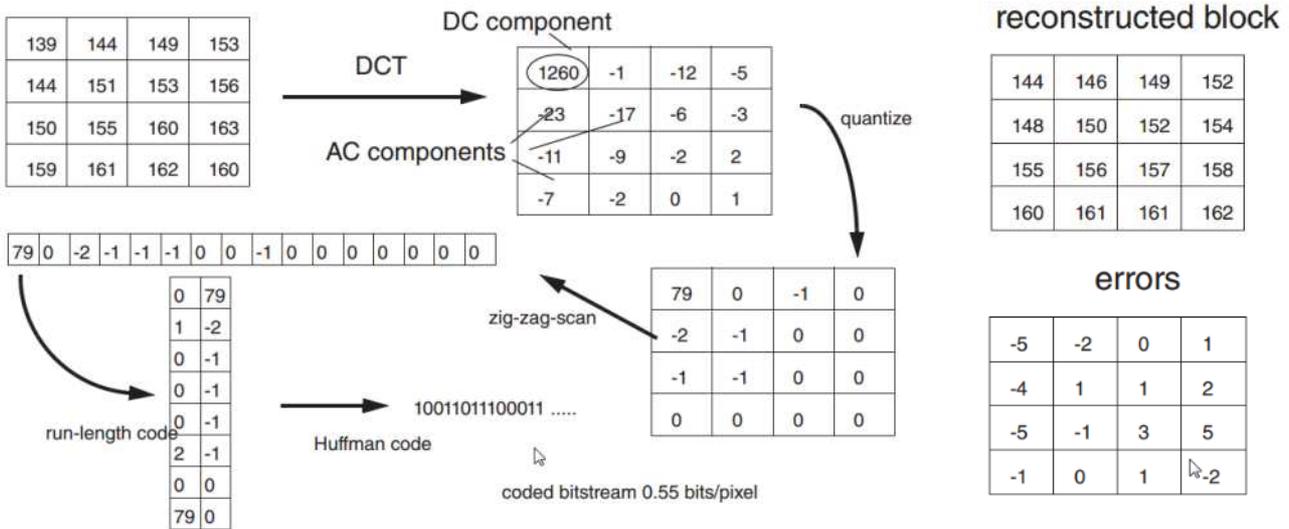
## JPEG – Quantisierung

- Abschneiden von überschüssigen Bits (truncation)
- Verwendung von Quantisierungstabellen:
  - Besteht aus 64 Elementen zu je 8 Bits ...  $Q_{uv}$
  - Neuer Wert  $Sq_{uv} = S_{uv} / Q_{uv}$
  - 2 vordefinierte Tabellen: luminance, chroma

## JPEG – Entropy Encoding

- 8x8-Pixelblöcke werden mittels Zig-Zag-Scan in einen Vektor mit 64 Elementen umgewandelt
- DC-Koeffizienten:
  - Bestimmen die Grundfarbe des Blocks
  - Haben einen hohen Wert, der aber oft ähnlich zum jeweiligen Vorgänger ist (Encode Difference)
- AC-Koeffizienten:
  - Koeffizienten mit niedriger Frequenz werden zuerst verarbeitet (wegen Zig-Zag-Scan)
  - Daher entsteht eine Sequenz von ähnlichen Bytes → effizientere Kodierung
- Algorithmus:
  - Run-Length-Kodierung auf Null-Werte der AC-Koeffizienten anwenden
  - Huffman-Kodierung auf DC- und AC-Koeffizienten anwenden
- Kodierung des DC-Koeffizienten:
  - DC-Werte mittels DC-Code-Tabelle (12 Kategorien) unterteilen: Differenzwerte → SSSS-Werte (Benötigte Bitanzahl pro Wert)
  - SSSS-Werte als Huffman-Symbole interpretieren und Huffman-Baum bilden

**Beispiel JPEG-Encoding/Decoding:**



**MPEG – Motion Picture Experts Group**

**Anforderungen**

- Hohe Kompression bei gleichzeitig hoher Bildqualität
- Unterstützung für symmetrische und asymmetrische (De-)Kompression
- Random-Access-Playback, Fast-Forward und (Fast-)Reverse möglich
- Synchronisation von Audio und Video

**MPEG-1**

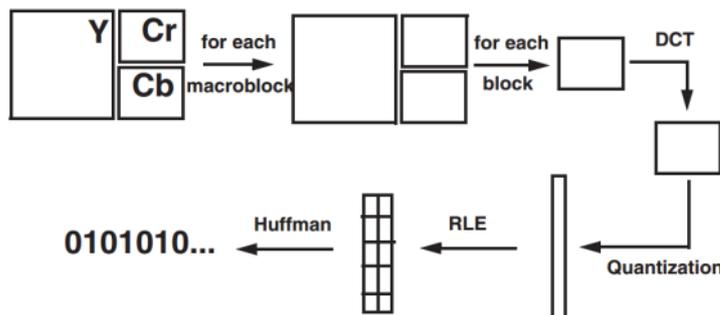
- Auf CD-ROM-Geschwindigkeit angepasst
- Video: Block-Transformationen und Bewegungskompensation bei Inter-Frames
- Audio: Subband-Coder mit Berücksichtigung eines psychoakustischen Modells

**Video-Standard**

- Bilder bestehen aus 3 Komponenten zu je 8 Bit: luminance + 2x chrominance
- Bilder werden in Macroblocks unterteilt (für Bewegungserkennung)
- 4 Arten von Frames → hohe Kompressionsraten durch Redundanzen, schneller random access

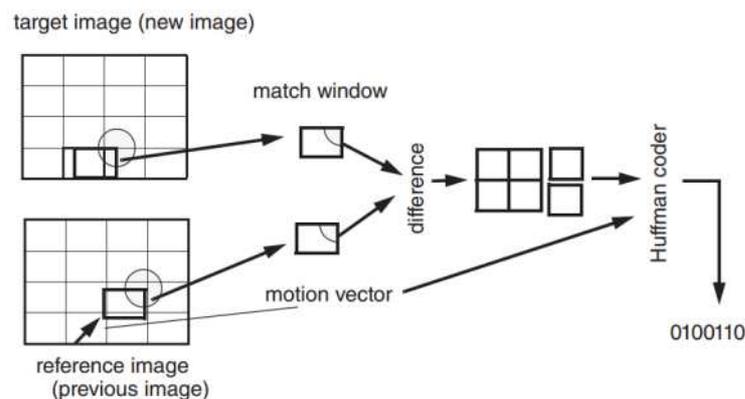
**Intra-coded images (I-frames)**

- Unabhängig von anderen Frames
- JPEG-komprimiert (hohe Qualität)



### Predictive-coded Frame (P-frame)

- Abhängig von vorhergehendem I- bzw. P-Frame
- Kodierung nutzt gleichbleibende oder verschobene Bildbereiche
- Motion Estimation:



- Motion Vector: besteht aus x- und y-Offset
- Motion Computation: innerhalb eines Suchfensters (search window) wird das übereinstimmende Fenster (match window = macroblock) gesucht
- P-Frames bestehen aus I-Frame macroblocks und predictive macroblocks

### Bi-directionally predictive-coded Frame (B-Frames)

- Abhängig von vorhergehenden und nachfolgenden I- bzw. P-Frames

### DC-coded Frames (D-Frames)

- Für Fast-Forward- und Fast-Rewind-Modi
- DC-Parameter werden DCT-kodiert, AC-Koeffizienten werden vernachlässigt
- D-Frames bestehen aus den tiefsten Frequenzen eines Bildes

### MPEG-Decoding

- Wegen B-Frames unterscheidet sich die angezeigte von der Dekodierungs-Bildreihenfolge
- Anzeige-Reihenfolge:
  - Frame-Art: I B B B P B B B I B B B P
  - Frame-Nr.: 1 2 3 4 5 6 7 8 9 10 11 12 13
- Dekodierungs-Reihenfolge:
  - Frame-Art: I P B B B B I B B P B B B
  - Frame-Nr.: 1 5 2 3 4 9 6 7 8 13 10 11 12

### MPEG-Quantisierung

- AC-Koeff. von B- und P-Frames haben hohe Werte, I-Frames haben kleinere Werte
- → MPEG-Quantisierung muss sich anpassen
- Je höher aktuelle Datenrate, desto größer die Quantisierungsstufen

### MPEG-Performance

- Dekodierung ist relativ einfach → via Software in Echtzeit ohne Probleme möglich
- Kodierung ist aufwendig → aktuelle Workstations immer noch zu langsam (bei hoher Qualität)

## MPEG-2

### Standard

- Verwendung von CCIR-601
- Soll als Sammlung von Tools gesehen werden (für verschiedene Ansprüche)
- Erweiterungen zu MPEG-1:
  - Mehr Seitenverhältnisse
  - 4:2:2-, 4:4:4-Macroblöcke
  - Progressive und interlaced frame coding
  - Verbesserung der Bildqualität (Quantisierung, Zig-Zag-Scan)
  - Höhere Bitraten

### Scalable Bit Streams

- 4 skalierbare Modi, um Video in verschiedene Ebenen zu unterteilen:
  - Spatial scalability: Basis-Ebene mit niedrigerer Auflösung
  - Data partitioning: Unterteilt die 64 Koeffizienten in 2 Bitstreams mit unterschiedlicher Priorität
  - SNR scalability: Kanäle haben selbe Datenrate, aber unterschiedliche Quantisierung (und damit Bildqualität)
  - Temporal scalability: Bitstream mit höherer Priorität wird mit niedrigerer Framerate kodiert; I-Frames werden in einem zweiten Bitstream in Abhängigkeit vom ersten Bitstream kodiert

### Profile und Level

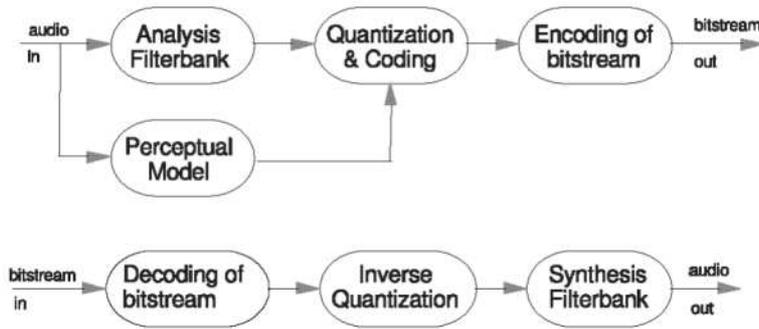
- Profil: vordefinierte Gruppe der gesamten Bitstream-Syntax, sind abwärtskompatibel
- Profile:
  - Nicht skalierbar: simple, main, 4:2:2, SNR
  - Skalierbar: spatial, high, multiview
- Level: wird innerhalb eines Profils definiert und ist eine Menge an Beschränkungen für bestimmte Parameter des Bitstreams, z.B. Auflösung, maximale Bitrate, ...
- Level: low(SIF), main (CCIR-601), high-1440, high (HDTV)
- Tabellen: siehe Folien 206 und 210-211

## MPEG-Audio

### MP3-Standard

- MPEG-1/2 Layer-3
- Offener und sehr gut definierter Standard
- Encoder und Decoder einfach verfügbar
- Weit verbreitet → große Unterstützung durch Hard- und Software
- Advanced Audio Coding (AAC) ... verbesserter Standard

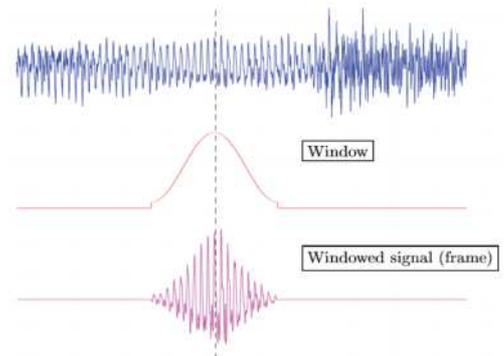
## Encoding-Algorithmus:



- Filterbank: teilt das Signal in Frequenzkomponenten auf
- Psychoakustisches Modell: berechnet Maskierungs-Grenzwert
- Quantisierung und Kodierung
- Kodierung des Bitstreams

## Spektrogramm

- Problem bei Anzeige eines Signals im Frequenzbereich: nicht ersichtlich, wann bestimmte Frequenzen auftreten → Spektrogramm = kombinierte Anzeige
- x-Achse ... Zeit, y-Achse ... Frequenz
- Färbung ... Anteil dieser Frequenz zu diesem Zeitpunkt
- Vorgang:
  - Problem: einzelne Samples haben keine Frequenz → Bereich um das entsprechende Sample wird miteinbezogen („Fensterfunktion“ wird mit dem Signal multipliziert, siehe rechts)
  - Ergebnis wird fouriertransformiert

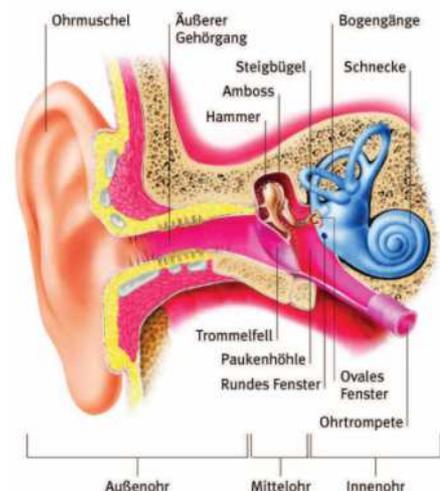


## Hörbereich

- Schmerzschwelle und Sprachbereich für MPEG nicht relevant
- Hörschwelle ist essentiell:
  - Nicht hörbare Töne können weggelassen werden
  - Mindestschallpegel ist frequenzabhängig
- Ruhehörschwelle: Diagramm wird mithilfe von Testpersonen ermittelt

## Das menschliche Ohr

- Innenohr:
  - Mit Flüssigkeit gefüllt
  - Ovale Fenster: Übertragung der Schallwellen
  - Rundes Fenster: Dämpfung der Schwingungen in der Schnecke (Cochlea)
- Schallwellen treffen auf Trommelfell → Übertragung der Schwingung auf Gehörknöchelchen im Mittelohr → Bewegung des Ovalen Fensters → Schwingung der Endolymphe in der Vorhoftreppe



### Hörschnecke

- 3 flüssigkeitsgefüllte Gänge: Vorhoftrappe, Schneckengang, Paukentreppe
- Basilmembran
  - trennt Schneckengang und Paukentreppe
  - ist am Anfang empfindlicher für hohe Frequenzen und am Ende empfindlicher für tiefe Frequenzen (Frequenzselektivität)
  - 24 Abschnitte: „Frequenzgruppen“ → zwei Töne mit sehr ähnlicher Frequenz können nicht differenziert werden
- Haarzellen: befinden sich auf der Basilmembran
  - Äußere Haarzellen: Verstärkung der Schwingungen; Vorfilter
  - Innere Haarzellen: Umwandlung in Nervenimpulse

### Maskierungseffekt

- Zwei gleichzeitig gespielte Töne mit ähnlicher Frequenz → Ohr hört nur den lauterem Ton
- Maskierungsschwellwert: minimale Lautstärke, mit der auch der zweite Ton gehört werden würde
- Maskierung abhängig von: Lautstärke, Frequenz, Zeitintervall und Dauer des maskierenden Tons

### Maskierung und MPEG-Audio

- Audiosignal beinhaltet verschiedene Frequenzanteile → Hörschwelle variiert ständig
- Psychoakustisches Modell: berechnet die Hörschwellenkurve zu jedem Zeitpunkt
- Variable Quantisierung: Quantisierungsrauschen muss immer nur knapp unter der aktuellen Hörschwellenkurve liegen

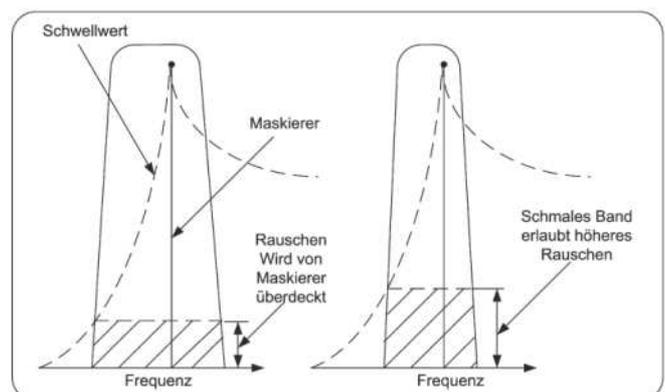
### Subbandkodierung

- Filterbank: teilt Audiosignal in Frequenzbänder auf (ideal: identisch zum menschlichen Gehör)
- Polyphase Filterbank:
  - 32 Subbands mit gleicher Bandbreite
  - Einfacher Aufbau, dafür Überlappungen → Informationsverlust
- Bsp: 4 Subbands, Frequenzbereich = 0-20 kHz, Abtastrate = 48 kHz, Wortlänge = 16 Bit
 

Subband 1:	0 - 5 kHz	Abtastrate = 12 kHz, 16 Bit
Subband 2:	5 - 10 kHz	Abtastrate = 12 kHz, 16 Bit
Subband 3:	10 - 15 kHz	Abtastrate = 12 kHz, 16 Bit
Subband 4:	15 - 20 kHz	Abtastrate = 12 kHz, 16 Bit
- Kritische Abtastung → kein Einfluss auf Datenmenge, kein Datenverlust (da Frequenzbereich in jedem Subband bekannt)

### Subbandkodierung / Kompression

- Für jedes Subband wird berechnet, wie hoch das Quantisierungsrauschen sein darf
- Höhere Subbandanzahl → schmalere Subbands → höhere Kompression



## MPEG Psychoakustisches Modell

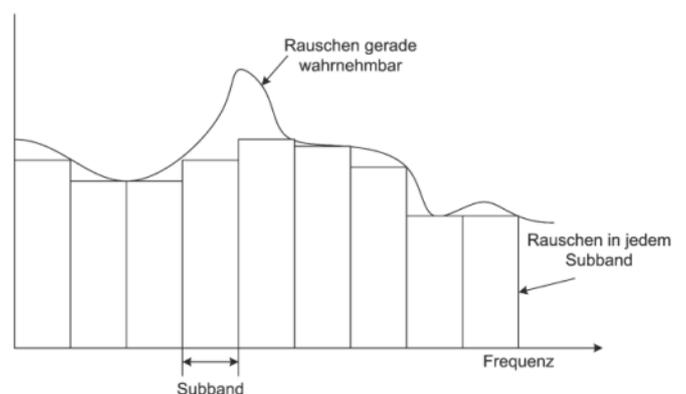
- Analysiert Audiosignal abschnittsweise auf Maskierungseffekte
- Berechnet Maskierungsschwellwert (maximales Quantisierungsrauschen) für jedes Subband
- Hauptverantwortlich für Qualität der Kodierung
- Modell 1: schnelle Berechnung
- Modell 2 (MP3): genaue Berechnung
- Tonale und atonale Komponenten haben unterschiedliche Maskierungseigenschaften → Modell muss unterscheiden können

## Fast Fouriertransformation (FFT)

- Gut geeignet für Signalanalyse: liefert feine Frequenzauflösung und Phaseninformationen
- Zusammenfassung in Frequenzgruppen → Minimierung des Rechenaufwands

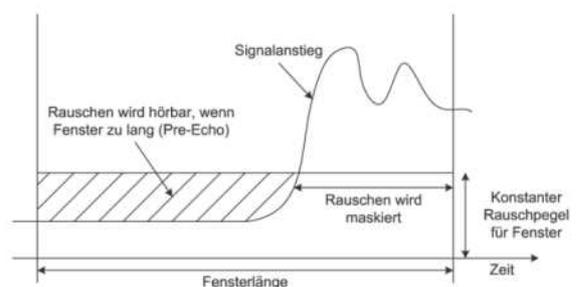
## Beispiel

- Kurve: berechnete Hörschwellenkurve
- Treppenkurve: maximales Rauschen pro Subband



## Kompression bei steilem Signalanstieg

- Fenster mit steilem Signalanstieg schwer komprimierbar
  - Quantisierungsrauschen über gesamter Fensterlänge gleich
  - Rauschen vor Signalanstieg als Pre-Echo hörbar



## Frequenzauflösung / Zeitfenster

- Hohe Frequenzauflösung bei langem Fenster
  - Vorteil: Viele Subbands, hohe Kompression → Subbands gut an Hörschwelle anpassbar
  - Nachteil: Quantisierungsrauschen über lange Zeit nicht änderbar → schlecht bei Signalanstiegen
- Geringe Frequenzauflösung bei kurzem Fenster → Vor- und Nachteile umgekehrt zu oben

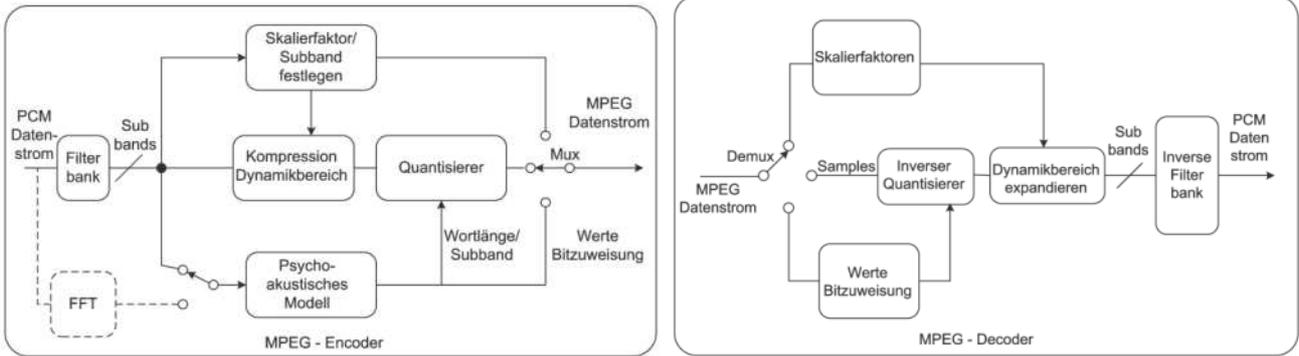
## Skalierfaktor

- Bandrauschunterdrückung
- Kleine Werte vor Kodierung um Skalierfaktor multipliziert, bei Dekodierung mit selben Faktor dividiert → Quantisierungsrauschen um Skalierfaktor gedämpft → besserer SNR für leise Signale

## Bitzuweisung

- Benutzer kann bei Kodierung gewünschte Datenrate wählen
- Zu niedrige Datenrate → Bitstellen reichen nicht aus
  - Hohe Frequenzen werden entfernt
  - Stereo ungenauer aufgelöst
  - Quantisierungsrauschen hörbar

**Einfacher MPEG-Audioencoder und -decoder:**



**MPEG-Datenstrom**

- Header: Beschreibung
- Redundancy Code (CRC): Fehlererkennung
- Bitzuweisung: Wortlänge der Subbandwerte
- Skalierfaktor: siehe oben
- Subbandwerte
- Hilfsdaten: beliebig



**MPEG-1 Audio-Standard**

- Wortlänge: 16 Bit; Sampling-Frequenzen: 32 kHz, 44.1 kHz, 48 kHz
- Kodierung als 1 einzelner Kanal, 2 separate Kanäle oder 1 Stereo-Kanal
- Kompressionsmethoden: Layer 1, Layer 2, Layer 3

**MPEG-2 Audio-Standard**

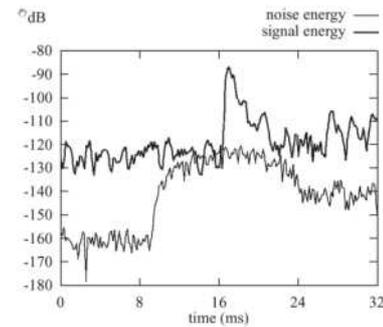
- Erweiterungen zu MPEG-1:
  - Multichannel-Kodierung mit Abwärtskompatibilität
  - Niedrigere Sampling-Frequenzen möglich
- AAC ... neuer effizienter Algorithmus, nicht abwärtskompatibel

**MPEG Layer-3 Algorithmus**

- Modi: Single Channel, Dual Channel, Stereo, Joint Stereo
- 2 Filterbänke:
  - Polyphase Filterbank (wie in Layer-1 und 2)
  - Zusätzlich: Modified DCT (MDCT)
- Wahrnehmungsmodell:
  - Maskierungsberechnungen
  - Output: Rauschgrenze für jedes Subband
- Quantisierung und Kodierung:
  - Besteht aus innerer Schleife (rate loop) und äußerer Schleife (noise control loop)
  - „Power-Law“-Quantisierung: höhere Werte werden gröber quantisiert
  - Danach: Huffman-Kodierung

## Artefakte

- Verlust von Bandbreite
- Pre-Echoes: Rauschen ist vor großen Signalanstiegen hörbar (siehe rechts)
- Roughness, double-speak



## Qualitätsmessung

- Einzige Möglichkeit: Hörtests
- ABC/HR-Test
  - Zuerst wird originales Signal (A) abgespielt, danach das Original und das kodierte Signal (BC) in zufälliger Reihenfolge („triple stimulus“)
  - Zuhörer muss entscheiden, welches Signal kodiert ist („hidden reference“)
- CCIR Impairment Scale Test
  - Originales und kodierte Signal wird in zufälliger Reihenfolge abgespielt
  - Zuhörer bewertet die Qualität nach einem Notensystem
- Qualität nicht objektiv messbar (z.B. durch SNR, Bandbreite, ...)
- Wahrnehmungsabhängige Messtechniken (pmt): noch nicht weit genug entwickelt

## III. Multimedia Environments

### Videodisc (speziell: LaserVision (LV))

- „Schallplattengroße Audio-CD“, enthält aber nur analoge Information
- 2 Formate:
  - CAV (constant angular velocity): flexibleres Playback
  - CLV (constant linear velocity): doppelte Kapazität
- Aspen Movie Map: mit LV realisiertes Hypermediasystem

### CD-Familie

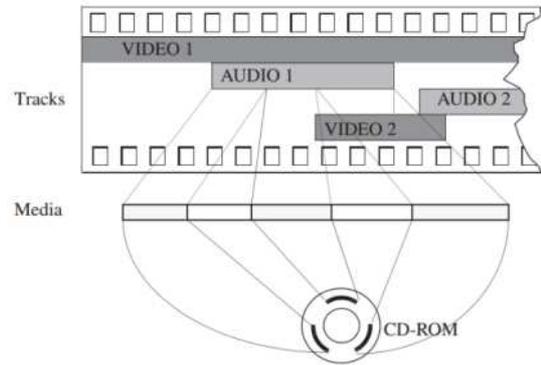
- Unterschiede zu LV: Größe, voll digital, nur CLV
- Frames: kleinste Dateneinheiten
- CD-ROM:
  - Zusätzliche Fehlererkennung und -korrektur
  - Disk ist in Datenblöcke unterteilt
- CD-i (Compact Disc - Interactive)
  - Erste Multimedia-Plattform
  - Video, Bilder, Audio, Text

### DVI (Digital Video Interactive)

- Formate und Codecs für Audio und Video, entsprechende Hard- und Software
- Video, Bilder, Audio, Text

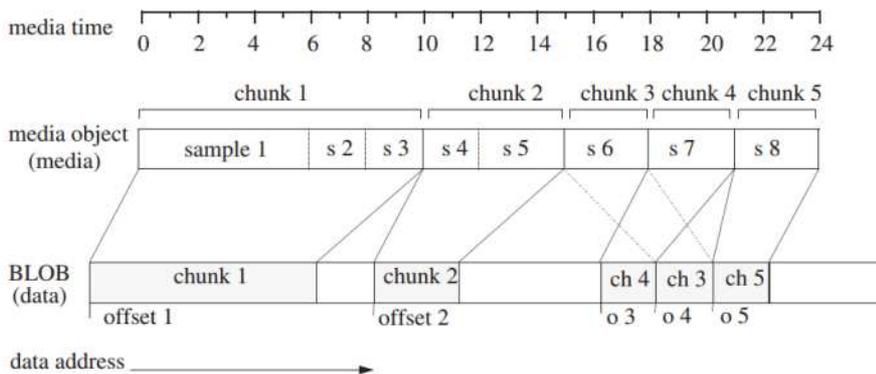
### Quicktime

- Offenes Environment
- „Movie“: besteht aus verschiedenen Tracks (Audio bzw. Video, siehe rechts)
- „Component“: Software mit einem bestimmten Interface



### Medien-Organisation

- Zeit in Quicktime:
  - Timing ist eindeutig
  - Time coordinate system (TCS)
- Data entity: repräsentiert den Speicher der Audio/Video-Daten
- Media entity: zeitliche Abfolge von Daten, hat einen Typ (Audio oder Video)
- Track entity: Anordnung von Media-Entities
- Movie entity: Gruppe von Track-Entities
- Speicherung auf Datenträger durch „Atome“ (kleinste Dateneinheit)



The mapping of media time to data address

Time-to-sample

sample number	sample span	sample start time	sample duration
1	1	0	6
2	3	6	2
5	4	12	3

Media Value

Ch 1	sample 1
	sample 2
	sample 3
Ch 2	sample 4
	sample 5
Ch 3	sample 6
Ch 4	sample 7
Ch 5	sample 8

Sample-to-chunk

chunk number	chunk span	start sample in chunk	samples /chunk	encoding
1	1	1	3	E1
2	1	4	2	E1
3	2	6	1	E1
5	1	8	1	E2

## IV Multimodal Information Retrieval

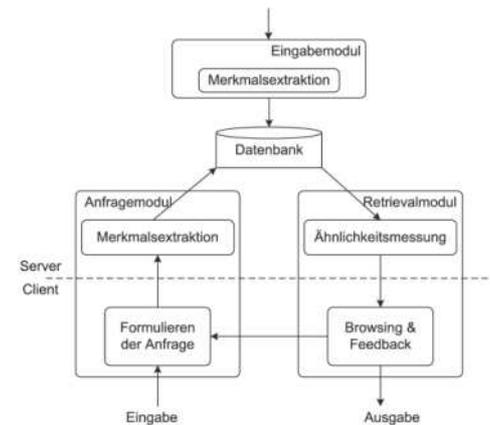
### IV.1. Grundlagen

#### Definition

- Information Retrieval (IR): Informationswiedergewinnung aus unstrukturierten Datenmengen
- Multimodal IR: Inhaltsbasiertes Multimedia Information Retrieval

#### Grundprinzip

- Merkmalsextraktion: Aus einem Medien- oder Multimediaobjekt werden inhaltsbasierte Merkmale extrahiert (feature extraction) und in einem Merkmalsvektor zusammengefasst
- Anfrage:
  - Referenz-Medienobjekt wird übergeben, ähnliche Objekte werden als Ergebnis zurückgeliefert (Vergleich von Anfragevektor und Kollektionsvektoren)
  - Ranking der Ergebnisse
  - Iteration durch Modifikation der Anfrage oder Relevanz-Feedback



#### Retrieval-Modelle

- Legt Realisierung folgender Komponenten fest: Interne Dokumentendarstellung, Anfrageformulierung, Vergleichsfunktion

#### Boolesches Modell

- Einfaches Modell, basiert auf Mengentheorie und boolescher Algebra
- Dokumente werden als Mengen von Indextermen mit Gewicht „1“ repräsentiert
- Anfragen: Angabe von Termen mit booleschen Junktoren (and, or, not)
- Vergleichsfunktion: prüft, in welchen Dokumenten die Anfrageterme enthalten sind
- Bsp:
  - 3 Dokumente:  $d_1 : \{\text{Sardinien, Strand, Ferienwohnung}\}$   
 $d_2 : \{\text{Korsika, Strand, Ferienwohnung}\}$   
 $d_3 : \{\text{Korsika, Gebirge}\}$
  - Anfragen und Ergebnisse:
    - „Korsika“ liefert  $\{d_2, d_3\}$
    - „Ferienwohnung“ liefert  $\{d_1, d_2\}$
    - „Ferienwohnung and Korsika“ liefert  $\{d_2\}$
    - „Ferienwohnung or Korsika“ liefert  $\{d_1, d_2, d_3\}$
    - „Ferienwohnung and not Korsika“ liefert  $\{d_1\}$
- Anfrage muss in disjunktive bzw. konjunktive Normalform (DNF bzw. KNF) umgewandelt werden
- Nachteile und Milderung:
  - Exaktes Modell: keine Ähnlichkeitssuche → Einführung von Relevanzstufen
  - Oft zu viele oder gar keine Dokumente → zweistufiges Suchverfahren zur Verfeinerung der Anfrage
  - Boolesche Junktoren für Laien schwer anwendbar → Junktoren sollten umformuliert werden (*all* statt *and*, ...)

**Fuzzy-Modell**

- Erweiterung des booleschen Modells → Abmilderung der zu scharfen Enthaltenseinsbedingung
- Prüft, wie stark der Term ein Dokument charakterisiert (Wert von 0 bis 1)
- Anfragen mithilfe von booleschen Junktoren (analog zu booleschem Modell) und Mengenoperationen
- Bsp:

- Fuzzy-Mengen:  $Korsika = \{ \langle d1; 0,1 \rangle, \langle d2; 0,6 \rangle, \langle d3; 1 \rangle \}$   
 $Strand = \{ \langle d1; 0,3 \rangle, \langle d2; 0,2 \rangle, \langle d3; 0,8 \rangle \}$

- Zugehörigkeitsfunktionen:

Anfrage	$\mu$	d1	d2	d3
	$\mu_{Korsika}$	0,1	0,6	1
	$\mu_{Strand}$	0,3	0,2	0,8
1	$\mu_{Korsika \cap Strand}$	0,1	0,2	0,8
2	$\mu_{Korsika \cup Strand}$	0,3	0,6	1
3	$\overline{\mu_{Korsika}}$	0,9	0,4	0

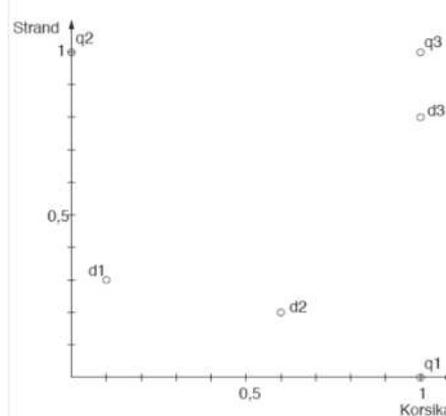
- Zugehörigkeitswerte: Berechnung z.B. durch Term-zu-Term-Korrelationsmatrix

**Vektorraummodell**

- Dokumente werden als Vektoren einen Vektorraums aufgefasst → Lineare Algebra anwendbar
- Unterstützt Ähnlichkeit
- Berechnet die Ähnlichkeit zwischen Anfragevektor und Dokumentenvektor (z.B. durch Cosinusmaß oder Distanzfunktionen)
- Bsp: 3 Dokumente, 2 Terme, 3 Anfragen

Dimension	d1	d2	d3
Korsika	0,1	0,6	1
Strand	0,3	0,2	0,8

Dimension	q1	q2	q3
Korsika	1	0	1
Strand	0	1	1

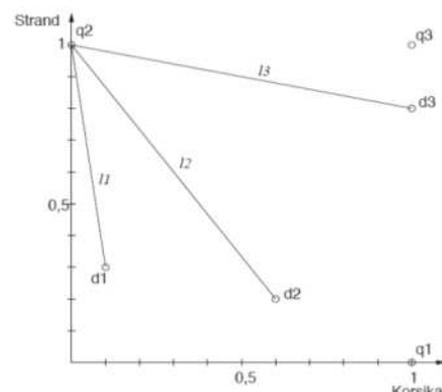
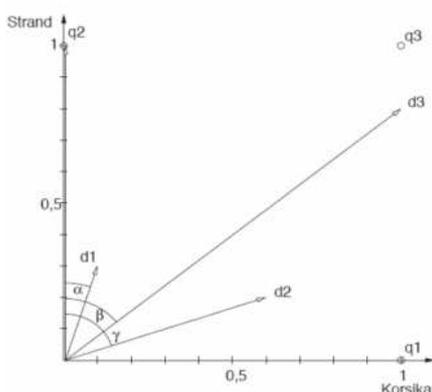


- Cosinusmaß:

$sim_{cos}$	d1	d2	d3
q1	0,3162	0,9487	0,7809
q2	0,9487	0,3162	0,6247
q3	0,8944	0,8944	0,9939

Euklidische Distanz (Distanzfunktion):

$dissim_{L_2}$	d1	d2	d3
q1	0,9487	0,4472	0,8
q2	0,7071	1	1,0198
q3	1,1402	0,8944	0,2



## Relevance Feedback

### Motivation

- Erste Ergebnisliste für den Suchenden oft nicht zufriedenstellend → Anfrageverfeinerung durch Nutzerinteraktion
- Gründe für Anfragemodifikation: vage Vorstellung über Ergebnis, schlechte Anfrageformulierung, unbekannte Datenkollektion
- Arten der Nutzerreaktion: Sequentielle Suche (oft wenig sinnvoll), manuelle Anfragemodifikation, Relevance Feedback
- Relevance Feedback: Bewertung der gefundenen Dokumente durch den Nutzer, IR-System modifiziert daraufhin die Anfrage
- Bsp:
  - $q$  ... theoretisch ideale Anfrage (dem Nutzer nicht bekannt)
  - $q_0, q_1, q_2$  ... Anfrage-Iterationen mit Bewertungen durch Nutzer

Anfrage	Ergebnisdokumente			
	1	2	3	...
$q$	$d_0$	$d_1$	$d_2$	...
$q_0$	$d_4$	$d_1 (+)$	$d_5 (-)$	...
$q_1$	$d_1 (+)$	$d_3 (+)$	$d_4 (-)$	...
$q_2$	$d_3$	$d_1$	$d_0$	...

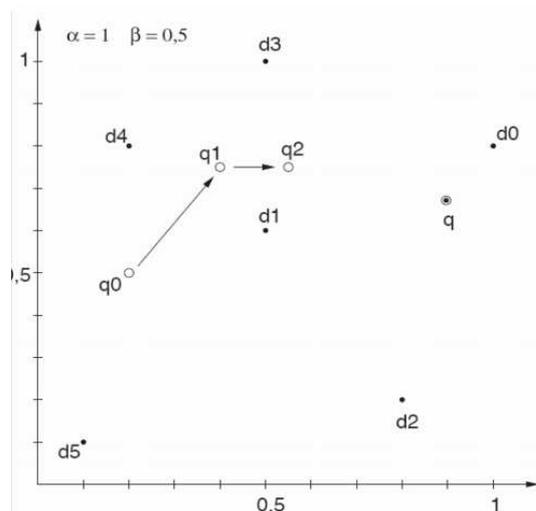
### Bewertung von Dokumenten

- Nur sinnvoll wenn:
  - Anzahl der zu bewertenden Dokumente < 10
  - Reduzierte Darstellung der Ergebnisdokumente
- Arten von Bewertungssystemen:
  - Relevant und keine Bewertung
  - Relevant, irrelevant und keine Bewertung
  - Gestufte Relevanzwerte
- Bewertungsgranulat: Ähnlichkeit zu mehreren Anfrageobjekten
- Pseudorelevanz: automatische Bewertung
- Bewertungsauswertung: Modifikation der Anfrage, von Nutzerprofilen, von Dokumentenbeschreibungen, des Suchalgorithmus oder von Anfragetermgewichten

### Verfahren von Rocchio

- Modifikation von Termgewichten des Anfragevektors
- Termgewichte relevanter Dokumente werden verstärkt u.u.
- Verschiebung des Anfragevektors in Richtung der relevanten Dokumente

Anfrage	Ergebnisdokumente			
	1	2	3	...
$q$	$d_0$	$d_1$	$d_2$	...
$q_0$	$d_4$	$d_1 (+)$	$d_5 (-)$	...
$q_1$	$d_1 (+)$	$d_3 (+)$	$d_4 (-)$	...
$q_2$	$d_3$	$d_1$	$d_0$	...

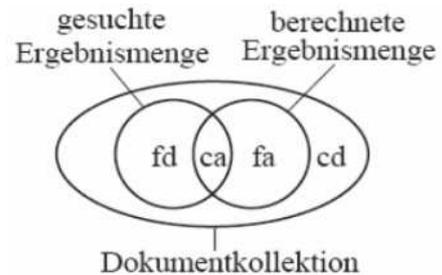


## Bewertung von Retrieval-Systemen

### Definitionen

- Correct alarms: korrekt als relevant zurückgeliefert
- Correct dismissals: korrekt als irrelevant eingestuft → nicht zurückgeliefert
- False alarms: irrtümlich als relevant zurückgelieferte Dokumente → möglichst zu vermeiden!
- False dismissals: irrtümlich als irrelevant eingestuft → nicht zurückgeliefert → möglichst zu vermeiden!

$$\begin{aligned}
 |\text{gesuchte Ergebnismenge}| &= fd + ca \\
 |\text{berechnete Ergebnismenge}| &= ca + fa \\
 |\text{Dokumentkollektion}| &= fd + ca + fa + cd
 \end{aligned}$$

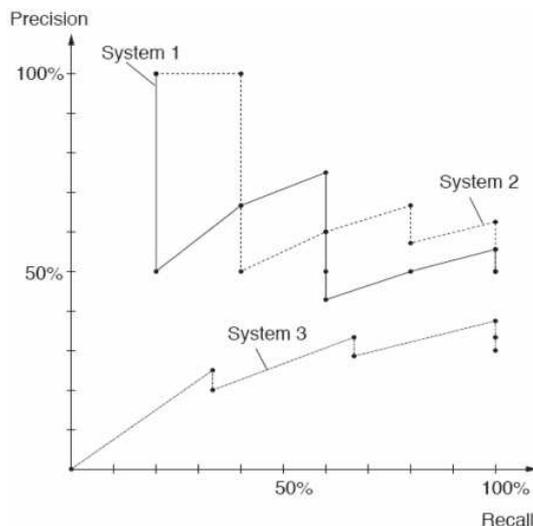


### Precision, Recall, Fallout

- Precision:  $P_q = \frac{ca}{ca + fa}$       Recall:  $R_q = \frac{ca}{ca + fd}$       Fallout:  $F_q = \frac{fa}{fa + cd}$
- Precision-Recall-Paare:
  - Abhängigkeit von Precision und Recall von der Größe der Ergebnismenge
  - Kann als Linie in einem Precision-Recall-Diagramm dargestellt werden
  - Inkrementelles Vergrößern der Ergebnismenge
  - Bsp 1:

Anzahl	1	2	3	4	5	6	7	8	9	10
$P_1$	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{5}$	$\frac{3}{6}$	$\frac{3}{7}$	$\frac{4}{8}$	$\frac{5}{9}$	$\frac{5}{10}$
$R_1$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{5}{5}$
$P_2$	$\frac{1}{1}$	$\frac{2}{2}$	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{3}{5}$	$\frac{4}{6}$	$\frac{4}{7}$	$\frac{5}{8}$	$\frac{5}{9}$	$\frac{5}{10}$
$R_2$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$
$P_3$	$\frac{0}{1}$	$\frac{0}{2}$	$\frac{0}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{2}{6}$	$\frac{2}{7}$	$\frac{3}{8}$	$\frac{3}{9}$	$\frac{3}{10}$
$R_3$	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$

- Bsp 2:



## IV.2. Content-based Image Retrieval

### Klassifizierung

- RBR (retrieval by browsing): Suche wird auf die tatsächlichen Bilder oder Thumbnails ausgeführt
- ROA (retrieval by objective attributes): Anfrage mithilfe von Meta-Daten und logischen Attributen
- RSC (retrieval by spatial constraints): Anfrage mithilfe von relativen räumlichen Beziehungen, z.B. Richtungen, Überschneidungen, ...
  - Relaxed RSC queries: unscharfe Suche mit Ranking der Ergebnisse
  - Strict RSC queries: scharfe Suche, alle Beziehungen müssen erfüllt sein
  - Verwendung eines „Sketch Pad Windows“: Symbole der gewünschten Objekte werden in einem Fenster platziert, daraus wird die Suchanfrage berechnet
- RSA (retrieval by semantic attributes)
- RFS (retrieval by feature similarity): Suche mithilfe von Beispielbildern (→ „content-based“ retrieval)

### Motivation

- Konventionell: Inhaltliche Erschließung sehr aufwändig und manchmal unmöglich (im Internet)
- Content-based IR: Bilder aufgrund des dargestellten Inhalts finden

### Anfragearten

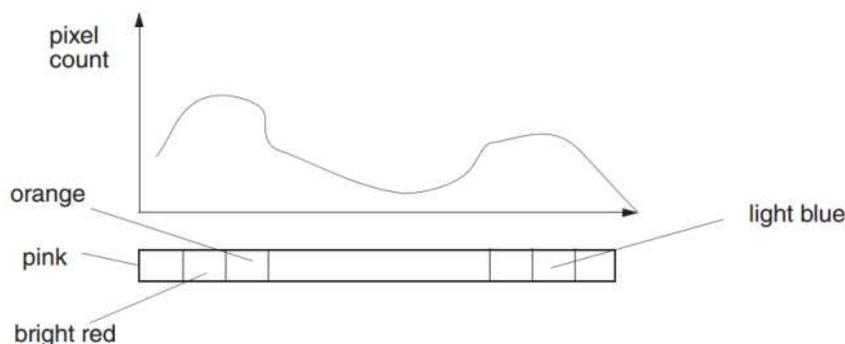
- Browsing: suche in vorgegebenen Kategorien
- Textsuche: mittels Schlagwörtern
- Visuelle Beschreibung:
  - Query-by-example: Vorgabe eines Beispielbildes
  - Query-by-sketch: Benutzer skizziert Vorgabebild
  - Query-by-template: Benutzer wählt Farb- und Texturkomponenten

### Merkmale

- Primitive Merkmale (Wahrnehmungsmerkmale): automatische Extrahierung, z.B. Farbe, Textur, Form
- Semantische Merkmale: z.B. Erkennung von Objekten und Szenen

### Farbmerkmale

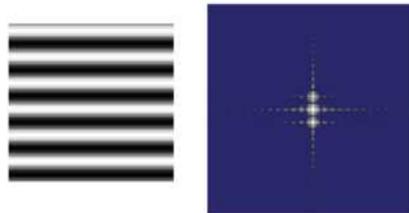
- Globale Farbverteilung: durch Farb-Histogramm beschrieben



- RGB → Distanzmaße nicht zufriedenstellend → uniformes Farbmodell notwendig (z.B. LUV)
- Einfache Repräsentation und Berechnung, unempfindlich bezüglich Bildauflösung, Rotation, Zoom
- Distanzmaße: L1- und L2-Maß (Euklidische Distanz), Histogram Intersection, Weighted Euclidean Distance
- Lokale Objektfarben: Benachbarte ähnliche Pixel werden zusammengefasst (color-based segmentation)

### Texturmerkmale

- Schwer beschreibbar → retrieval-by-example
- Auch bei Graustufenbildern aussagekräftig
- Eigenschaften: z.B. Körnigkeit (Feinheitsgrad), Periodizität (Regelmäßigkeit), Ausrichtung im Raum
- Verfahren
  - Strukturell: Lage und Ausrichtung, besonders für regelmäßige Texturen
  - Statistisch: Verteilung der Helligkeitswerte, Autokorrelationsfunktion (Körnigkeit), Co-Occurrence-Matrix (Kontrast)
  - Markov Random Fields: Pixel werden abhängig von ihren Nachbarpixeln modelliert
  - Fraktale Modelle: für unregelmäßige Texturen (z.B. Küstenlinien)
- Texturanalyse im Frequenzbereich:
  - Frequenzbereich: Räumliche Verteilung der Helligkeitswerte  
Häufige Helligkeitsunterschiede → hohe Frequenzen u.u.  
Errechnet durch Fourier-Transformation
  - Frequenzbild: Position → Frequenz (je weiter entfernt vom Mittelpunkt, desto höher die Frequenz)
  - Bsp: Textur → Frequenzbild:



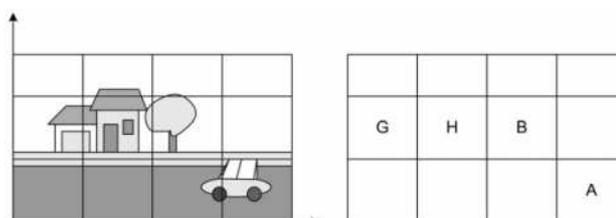
- Tamura-Modell: 3 Parameter (Kontrast, Körnigkeit, Ausrichtung)
- Brodatz-Datenbank:
  - Besteht aus Graustufenbildern von verschiedenen Texturen
  - Wird verwendet, um Textur-Algorithmen zu testen
- Testvorgang:
  - Jedes Brodatz-Bild wird in 9 Sub-Bilder aufgeteilt, auf diese wird der Algorithmus angewandt
  - Ergebnisse: Retrieval Rate, Average Retrieval Rate
- Edge Histogram Descriptor: beschreibt räumliche Anordnung von Kanten im Bild

### Formmerkmale

- 2 Arten:
  - Umrissbasiert (contour shape): Form eines Objekts durch Kantenextraktion
  - Bereichsbasiert (region shape): zusammenhängende oder unterbrochene Regionen
- MPEG-7 Shape Descriptors: Contour Shape, Region Shape, Shape 3D Spectrum, Shape 2D/3D

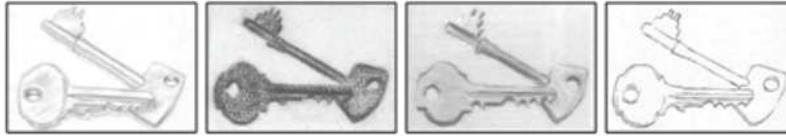
### Weitere Merkmale

- Räumliche Anordnung von Objekten in einem Bild
  - Bsp: 2D-String



→ Projektion auf die x-Achse: „h:  $G < H < B < A$ “, Projektion auf die y-Achse: „v:  $A < G = H = B$ “

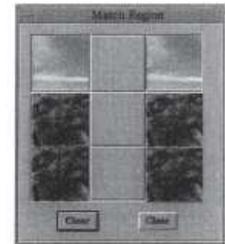
- Kantenbilder (query-by-sketch)
  - Bildsuche mittels handgemalter Skizze
  - Bsp:



- Beziehung zwischen Objekten:
  - Richtungsbezogen: Position zueinander, Entfernung, Winkel
  - Topologie: Berührung, Überschneidung, etc.
- Gesichtserkennung
  - Gesichtsdetektion: Erkennen von Gesichtern in Bildern
  - Gesichtsidentifikation: Wiedererkennen eines Gesichts
  - Anwendung: z.B. Zugangskontrollen, Analyse von Überwachungsvideos

### User Interface

- Query-by-template: Benutzer wählt verschiedene Sample-Bilder (bestimmte Farben bzw. Texturen) aus und platziert sie in einem Grid („Template-Map“)



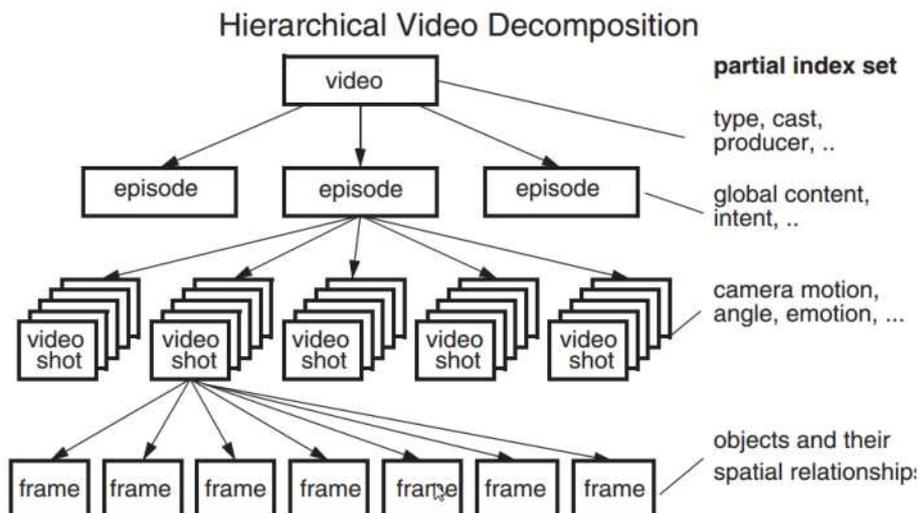
- Query-by-sketch:



- Query-by-example: Benutzer formuliert simple Anfrage → vorläufiges Ergebnis wird zurückgegeben → Benutzer wählt aus den Ergebnisbildern jene Bilder aus, die am ähnlichsten zu dem endgültigen Ergebnis sein sollen  
Beispielsystem: „Blobworld“ (siehe Folien 465-466)

## IV.3. Video Retrieval

### Schnitterkennung (Video Segmentation)



**Definitionen**

- Shot (Take): Sequenz von Video-Frames ohne Übergänge
- Video Segmenation: Bestimmung der Grenzen zwischen Shots → leichtere Organisation von Video-Material für Browsing und Content-based IR
- Unterscheidungen von Video Segmentation:
  - Frame-Unterschiede: Pixel-Vergleiche, Histogramm-Vergleiche
  - Camera-Operationen
  - Spezialfall: Komprimierte Videos

**Pixel-Vergleichsmethoden**

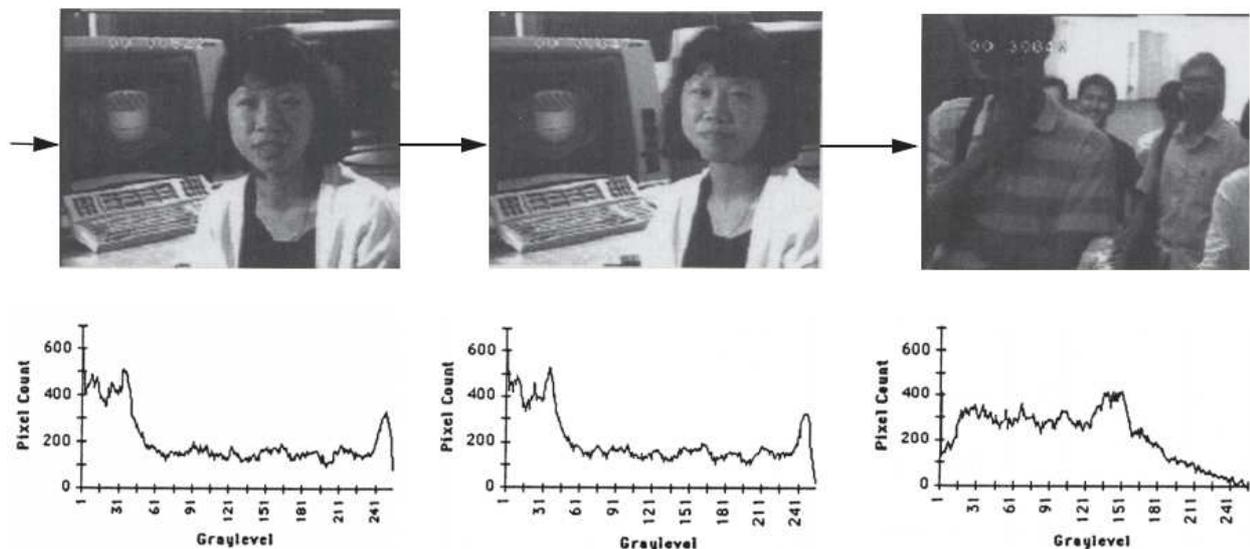
- Paarweiser Pixel-Vergleich: Vergleich der Pixelwerte in 2 aufeinanderfolgenden Frames



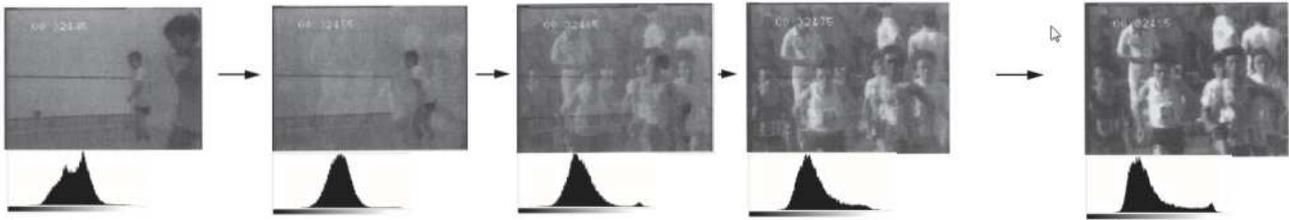
- Segment Boundary: prüft, ob sich ein bestimmter Prozentsatz an Pixeln geändert hat
  - Nachteil: anfällig für Fehler bei Bewegung von Kamera bzw. Objekten, Beleuchtungsänderung etc.
- Likelihood Ratio: Frames werden in Bereiche aufgeteilt, die miteinander verglichen werden
  - Erkennt Kameraschnitt, wenn der LR von genügend Bereichen den Grenzwert überschreitet
  - Vorteil: höhere Toleranz für sich langsam bewegende und kleine Objekte
  - Nachteil: erkennt Kameraschnitt nicht, wenn Mittelwert und Varianz der beiden Bereiche gleich bleiben (sehr unwahrscheinlich)

**Histogramm-Vergleichsmethoden**

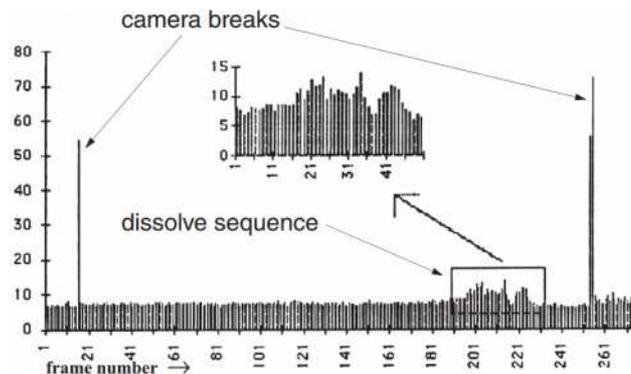
- Weniger empfindlich für Objekt-Bewegungen
- Bsp: harter Schnitt (Camera Break)



- Bsp: weicher Übergang, Überblendung (Dissolve Sequence)



- Frame-to-Frame-Unterschiede:
  - 2 Grenzwerte zur Unterscheidung von harten Schnitten und weichen Übergängen



**Motion Continuity**

- Bewegungsvektoren: repräsentieren die Geschwindigkeit eines Pixelblocks im Bild; werden mittels Block Matching berechnet
- Gleichmäßigkeit (smoothness) der Bewegung wird ebenfalls gemessen

**Weiche Übergänge (Special Effects):**



- Twin-Comparison-Methode (siehe auch Frame-to-Frame-Unterschiede oben)
  - 2 Grenzwerte für Erkennung von harten Schnitten und Special Effects
  - Immer wenn ein Histogramm-Differenzwert zwischen diesen Grenzen liegt, beginnt der Übergang
  - Der Übergang endet, sobald einer der folgenden Frames den Bereich zwischen den Grenzen verlässt
  - Bsp: siehe Folien 488-490
- Multi-Pass-Methode:
  - Vorübergehend niedrigere Frames pro Sekunde → kürzere Verarbeitungszeit

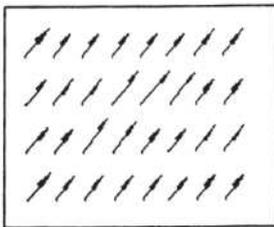
## Bewegungsanalyse

### Camera Operation Detection

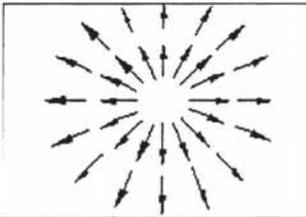
- Unterscheidung zwischen Special Effects und Kamerabewegung notwendig
- Zwei Methoden: Bewegungsvektor-Analyse, Video X-Ray

### Bewegungsvektor-Analyse

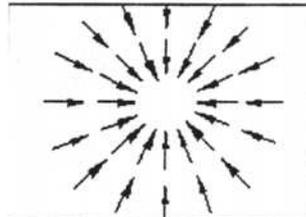
- Frame wird in Blöcke unterteilt
- Block Matching: jedem Block wird ein Vektor zugewiesen, der den gesamten Block verschiebt
- Optischer Fluss (optical flow): Vektorfeld, wo jeder Vektor die Geschwindigkeit des zugehörigen Blocks repräsentiert
- Erkennung von: Schwenks (panning), Neigung (tilting), Zoom, bestimmte Bewegungsmuster
- Panning & Tilting:
  - Optischer Fluss: gleiche Richtungen (parallel, siehe rechts)



- Zooming
  - Optischer Fluss: Vektoren gehen vom Mittelpunkt des Frames aus
  - Zoom in:
  - Zoom out:



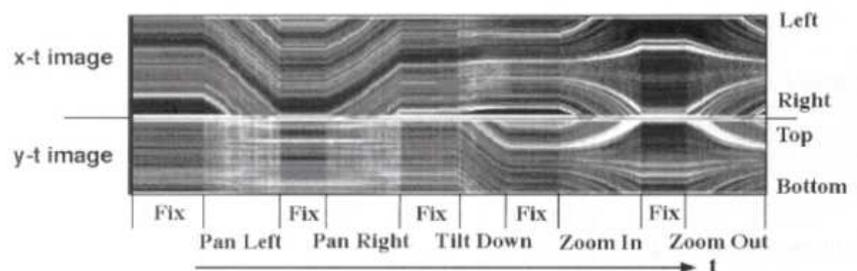
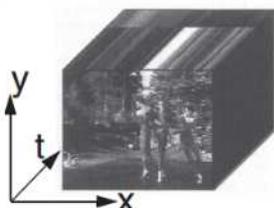
Zoom out:



- Objektbewegung: Menschen und Tieren (laufen, springen, ...), Fahrzeuge (fahren, fliegen,...), etc.

### Video X-Ray

- Bild mit Raum- und Zeitebene
- Panning (Schwenk): schräge Linien auf der Top-Ansicht
- Tilting (Neigung): schräge Linien auf der Seitenansicht
- Zoom: gebogene Linien auf beiden Ansichten



## Video Representation

- Möglichkeit gesucht, mit der man teils stundenlanges Videomaterial in wenigen Minuten zusammenfassen kann
- Video Content Abstraction: Extrahieren der wichtigsten Informationen, des generellen Stils und der Hauptthemen via Video Icon construction oder Key-Frame extraction

### Video Icons

- Verschiedene Ansätze:
  - Movie Icon (Micon): Repräsentation als 3D-Objekt
  - Interactive Micons
  - Paper-Video: Diagramm-ähnlicher Browser
  - Video Panorama: ähnlich einem „Panoramafoto mit Zeitkomponente“
  - Videoscope: Content analyzer
  - Sound Browser: Unterteilung in Sequenzen mit Musik, Sprache, etc.
  - Salient Stills: wichtige Szenen in einem Bild zusammengefasst
  - Videoscape Icon
  - Videomap: Zeitleiste mit Referenz-Shots und zugehörigen Histogrammen, X-Ray, etc.
  - Beispiele: siehe Folien 509-517

### Key-Frame Extraction

- Einfache Berechnung: bei jeder größeren Veränderung des Contents wird ein Keyframe gesetzt
- Nachteil: nicht sehr repräsentativ

## Segmentierung von komprimierten Videos (MPEG)

- Hardware-Dekompression: alle obigen Verfahren können ebenfalls angewendet werden
- Software-Dekompression: viel weniger effizient
- Falls HW-Dekompression nicht verfügbar → Eigenschaften verwenden, die direkt auf komprimierten Daten anwendbar sind, z.B. DCT-Koeffizienten, MPEG motion vectors

### Algorithmen mit DCT-Koeffizienten

- Nur I-Frames haben DCT-Koeffizienten
- Vorteil: nur wenige Frames sind I-Frames → kürzere Rechenzeit
- Nachteil: Verlust von zeitlicher Auflösung → möglicherweise falsche Positive
- DCT Coefficients Correlation:
  - DCT-Koeffizienten von aufeinanderfolgenden Frames wird verglichen
- DCT Block Comparison:
  - Paarweiser DCT-Block-Vergleich
- Nicht gut geeignet für harte Schnitte

### Algorithmus mit Bewegungsvektoren

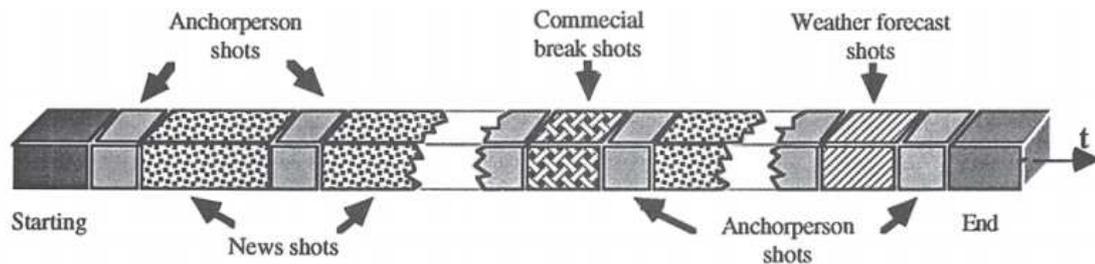
- P-Frames: Einzelnes Set an Bewegungsvektoren
- B-Frames: 2 Sets an Bewegungsvektoren (zum vorigen und nachfolgenden Frame)
- Während eines Shots: relativ gleichmäßige Änderungen
- Bei einem harten Schnitt: Gleichmäßigkeit wird unterbrochen
- Nicht gut geeignet für weiche Übergänge

### Hybrider Algorithmus

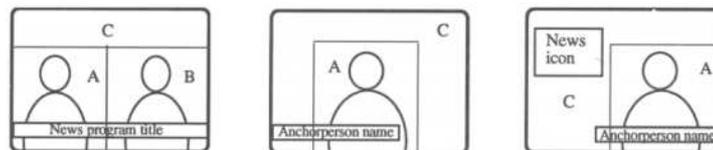
- Erster Durchgang: DCT-Vergleich (I-Frames), Erkennung potentieller Schnitte, Übergänge, ...
- Zweiter Durchgang: DCT-Vergleich mit kleinerem skip factor, Erkennung von falschen Positiven
- Weitere Durchgänge: Bewegungsbasierte Vergleiche der erkannten Sequenzen, erhöhte Genauigkeit
- Genauester und effektivster Ansatz

### Beispiel: TV-Nachrichten

- Aufgabe: Automatische Extraktion von semantischen Informationen



- Zeitliche Segmentierung: via Video Segmentation
- Einteilung der Shot-Arten:
  - z.B. Modellierung verschiedener Nachrichtensprecher-Shots:



- Intro, Outro, Werbespots, Wetterbericht, etc.
- Abstraktion: Erzeugung von Key-Frames für jeden Shot

## IV.4. Audio-Retrieval

- Siehe auch: „Spektrogramm“ im Kapitel II. Kompression

### Audio-Klassifikation

- Sprache: männlich/weiblich
  - Geringe Bandbreite (100 – 7000 Hz) → niedriger Zentroid
  - Häufige Pausen, höherer Anteil an Stille
  - Nulldurchlaufrate (Zero Crossing ZC) variiert stark
- Musik
  - Hohe Bandbreite (16 – 20000 Hz) → höherer Zentroid
  - Niedriger Anteil an Stille
  - Nulldurchlauf variiert nicht so stark
  - Regelmäßiger Rhythmus
  - 3 Arten: strukturiert, synthetisch (MIDI), aufgezeichnet
- Umgebungsgeräusche

## Vorgang Klassifikation

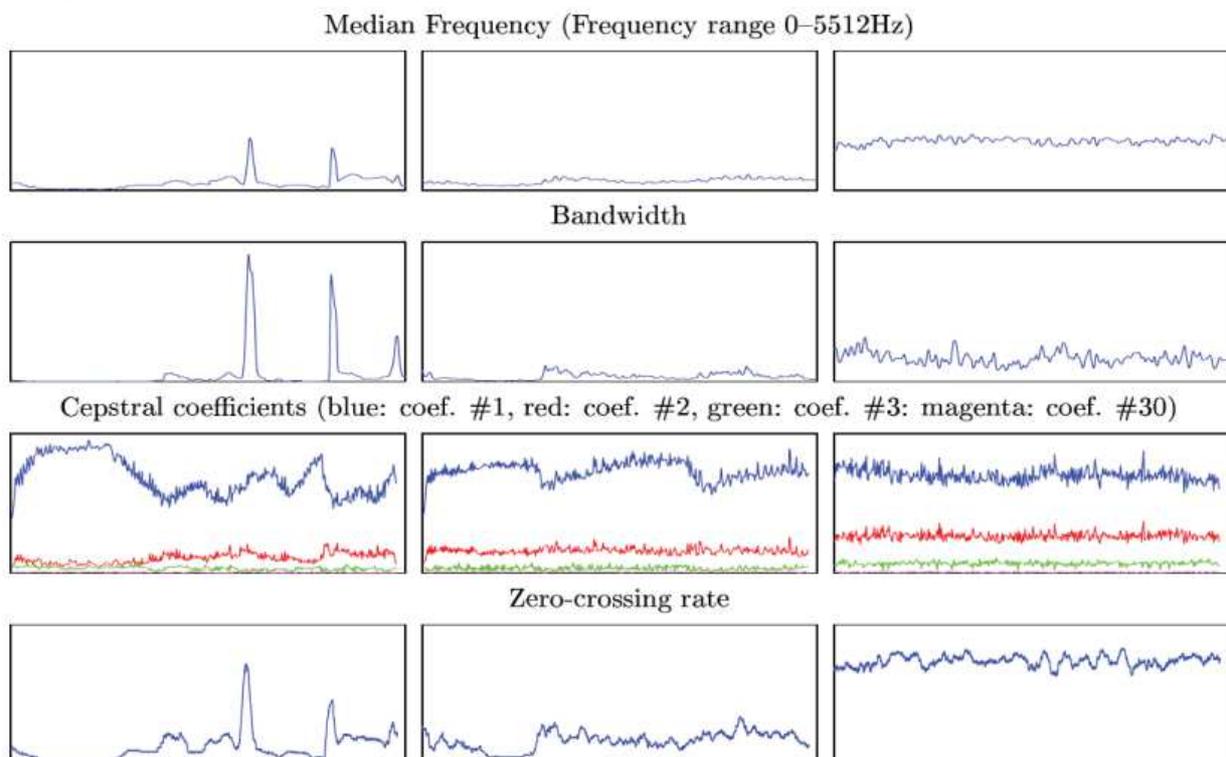
- Ein Merkmal nach dem anderen (hohe Differenzierung zuerst → einfacher zu berechnen)
- Bsp:
  1. Zentroid: wenn hoch → Musik
  2. Anteil der Stille: wenn niedrig → Musik
  3. ZC-Variabilität: wenn niedrig → Solo-Musik, ansonsten → Sprache

## Indexierung strukturierter Musik

- Keine Merkmalsextraktion erforderlich
- Ähnlichkeit schwierig zu definieren
- Möglichkeit: nur Tonhöhenwechsel berücksichtigen (Up, Down, Repeat) → Zeichenkettenvergleich

## Audio-Merkmale im Spektrogramm-Bereich

- Anteil der Stille
  - Silence Ratio: Anteil der Messwerte, die einer Periode der Stille angehören
  - Stilleperiode: Bestimme Anzahl an aufeinanderfolgenden Messwerten, die unter einem bestimmten Amplituden-Grenzwert liegen
- Harmonie: Frequenzen der dominante Komponenten sind ein Vielfaches einer Grundfrequenz, v.a. bei Musik
- Tonhöhe: nur bei periodischen Klängen
- Nullkreuzungsrate (Zero Crossing Rate): Anzahl der Vorzeichenwechsel in einem Frame
- Problem: außer bei ZCR immer Spektrogramm notwendig → Fensterfunktion und -länge hat hohe Bedeutung
- Vergleich von Merkmalen (Redundanzen):



## IV.5. Distanz und Ähnlichkeit

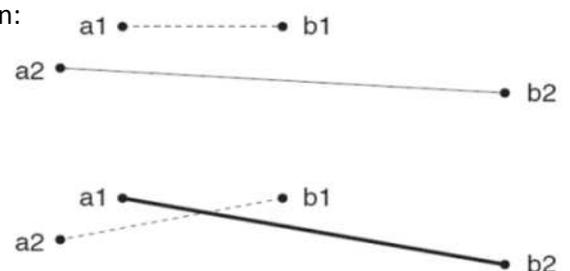
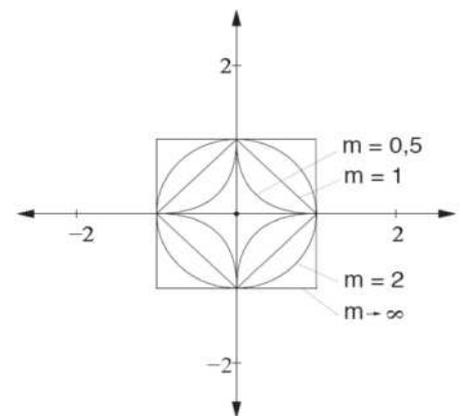
### Distanzfunktionen

- Vergleichen die Merkmalswerte zweier Medienobjekte
- Invarianz: drückt aus, welche Merkmale nicht verglichen werden sollen
- Distanzfunktion:
  - Binäre Funktion
  - 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):*  $\forall o_1, o_2, o_3 \in O : d(o_1, o_3) \leq d(o_1, o_2) + d(o_2, o_3)$

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

### Arten von Distanzfunktionen

- Einfache Distanzfunktion:  $d_{abs}$
- Euklidische Distanzfunktion:  $d_{L2}$
- Minkowski-Distanzfunktion:  $d_{Lm}$ 
  - $m = 1 \rightarrow$  Manhattan- oder Blockdistanz
  - $m = 2 \rightarrow$  euklidische Distanz
  - $m = \text{unendlich} \rightarrow$  Max- oder Tschebyscheff-Distanz
  - translationsinvariant, aber nicht rotations- (außer  $m=2$ ) und skalierungsinvariant
  - m-Einheitskreise: je kleiner  $m$ , desto größer die Distanz zwischen 2 Punkten (siehe rechts)
  - häufig als gewichtete Variante verwendet
- Quadratische Distanz:  $d_q$ 
  - Erweiterung der gewichteten euklidischen Distanz durch Drehung
  - Arten: Einheitsmatrix, Diagonalmatrix, orthonormale Matrix, symmetrische Matrix
- Mahalanobis-Distanzfunktion:  $d_M$
- Quadratische Pseudodistanz:  $pd_q$ 
  - Abstand 0 auch für nichtidentische Punkte
- Bottleneck-Distanz:  $d_b$ 
  - Distanzfunktion auf (gleich große) Mengen
  - Sucht das Minimum der maximalen Elementepaarabstände:

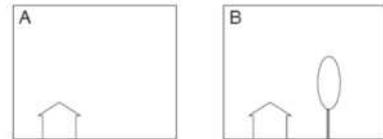


## Ähnlichkeitsmaße

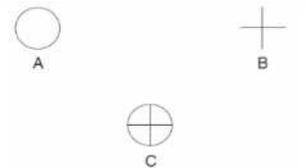
- Problem: keine allgemein akzeptierte Definition von Ähnlichkeit
- Ähnlichkeitsmaß: Funktion, die 2 Objekten eine Zahl zwischen 0 und 1 zuordnet (1 ... max. Ähnlichkeit)
- Viele Ansätze verwenden Distanzfunktion auf Featurewerte
  - Distanzwerte werden auf  $[0,1]$  abgebildet
  - Distanz nicht für alle Anwendungen geeignet

### Probleme

- Selbstidentität: gilt nicht grundsätzlich
- Positivität: keine allgemeine Bedingung für menschliches Ähnlichkeitsempfinden
- Symmetrie: Rollentausch macht Unterschied:  
(„Was sind die wichtigen Merkmale?“)



- Dreiecksungl.: Unterschiede zwischen 2 Objekten werden zu hoch bewertet, wenn kein drittes Vergleichsobjekt vorhanden ist → Unähnlichkeit zwischen A und B wird stärker eingeschätzt als Summe der Unähnlichkeiten zu C



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

### Ähnlichkeitsabstand

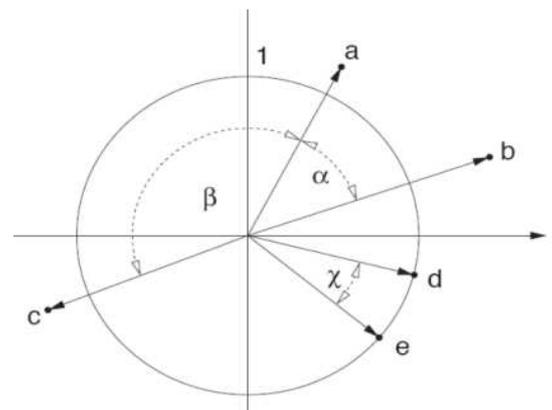
- Eigenschaften des Ähnlichkeitsabstands sind allgemeiner als die Distanzeigenschaften (z.B. Symmetrie nicht gefordert)
- Bei Anwendung einer monoton wachsenden Funktion bleiben Eigenschaften erhalten

### Ähnlichkeitsmaße

- „Weltwissen“ spielt bei Ähnlichkeitsempfindung eine Rolle
- Ebenen der Inhaltsverarbeitung:
  - Syntaktische Ebene: ohne Bedeutung der Objekte
  - Semantische Ebene: Ähnlichkeitsvergleich
  - Pragmatische Ebene: Interpretation, thematische Kategorien
- Pre-Attentive vs. Attentive Wahrnehmung:
  - Pre-Attentive: in den ersten 250 ms; ohne Interpretation (Weltwissen)
  - Pre-Attentive Features: z.B. Größe und Anzahl von Objekten, Farbe, Intensität
- Beispiele für Ähnlichkeitsmaße: Repräsentationssatz, Kosinusmaß

### Kosinusmaß

- Weit verbreitet, basiert auf Vektorraummodell
- Skalarprodukt von Vektoren (siehe rechts)
- Ergebnis: Semi-Pseudo-Distanzfunktion

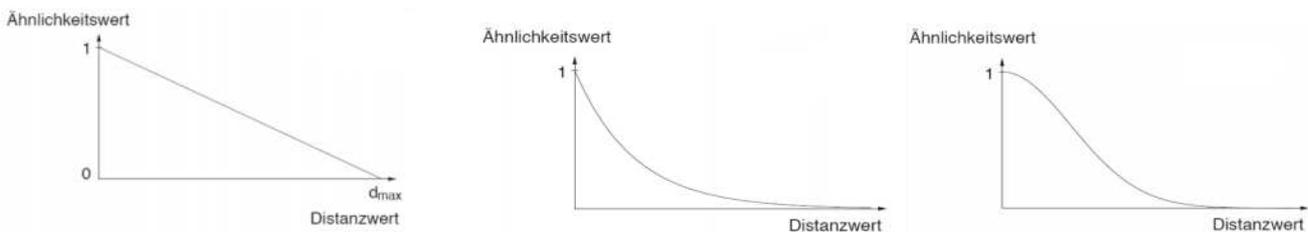


### Aggregation von Werten

- Anforderungen an eine Aggregatfunktion: Ähnlichkeitswerte, (strikte) Monotonie, Stetigkeit, Idempotenz, Unabhängigkeit von der Reihenfolge
- Generalisiertes Mittel

### Umwandlungsfunktionen

- Linearkombination der Grenzbedingungen (unten links)
- Dynamische Sensibilität → hohe Sensibilität gegenüber geringen Distanzen (unten Mitte)
- Modifikation: Abschwächung bei sehr geringen Distanzen (unten rechts)
- Parametrisierbare Funktion: Sensibilität selbst bestimmbar



## V. Content Description (MPEG-7)

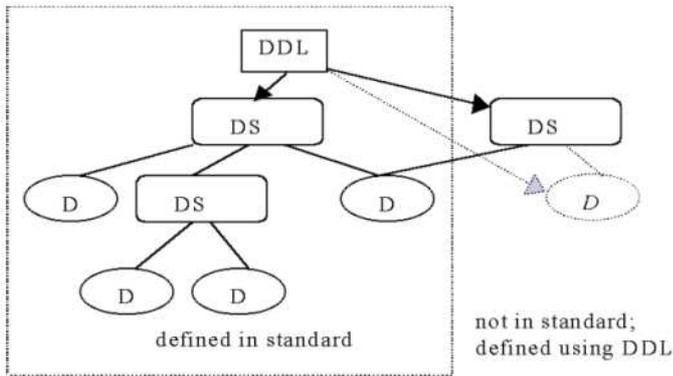
### Motivation

- Stetig steigende Menge an Multimedia-Content → unübersichtlich
- Internet → hohe Anzahl an Produzenten und Konsumenten
- Wert der Information in Multimedia-Content auch abhängig davon, wie einfach diese gefunden werden kann
- Lösung → MPEG-7
  - System zur manuellen oder automatischen Beschreibung von multimedialen Inhalten
  - Flexibilität beim Daten-Management
  - Verbesserung bei Zusammenarbeit zwischen Daten-Ressourcen
  - Standardisierung von „Description Schemes“, „Descriptors“, „Description Definition Language“ (DDL)

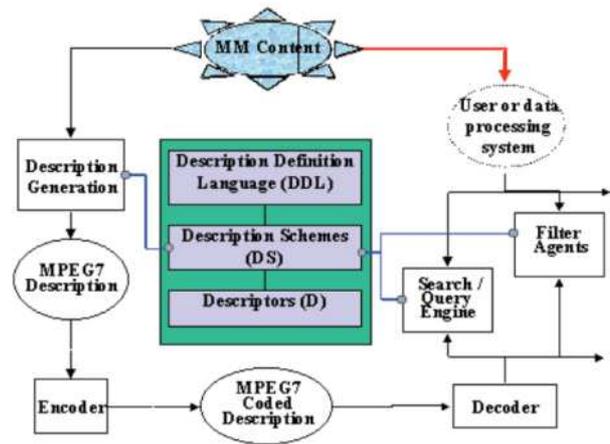
### Definitionen

- Feature: charakteristische Eigenschaft eines Datensets, z.B. die Farbe eines Bildes, Rhythmus eines Audio-Segments, Genre eines Musikstücks, Titel eines Films, ...
- Descriptor: Repräsentation eines Features mit definierter Syntax und Semantik, z.B. „Color: string“
- Ein Feature kann auch durch mehrere Descriptoren beschrieben werden, z.B. Farb-Histogramme
- Descriptor Value: Instanz eines Descriptors mit entsprechendem Wert
- Description Scheme: spezifiziert die Struktur und Semantik von Beziehungen zwischen Descriptoren und/oder weiteren Description Schemes (siehe rechts)
- Description: Kombination aus Descriptor Value und Description Scheme
- Coded Description: kodierte Description
- Description Definition Language (DDL): ermöglicht die Erzeugung und Modifikation von Description Schemes und Descriptoren

**Description Scheme:**



**MPEG-7 in der Praxis:**



**DDL-Anforderungen:**

- Compositional Capabilities:
  - Erzeugung neuer Description Schemes und Descriptoren
  - Ein Description Scheme kann aus mehreren Description Schemes bestehen
- Transformational Capabilities:
  - Wiederverwendung, Erweiterung und Vererbung von vorhandenen Description Schemes und Descriptoren
- Unique Identification:
  - Eindeutige Referenzierung für jedes Description Scheme bzw. Descriptor
- Data Types: primitive Datentypen (Text, Integer, ...)
- Beziehungen zwischen bzw. innerhalb eines Description Schemes
- Beziehungen zwischen Description und Daten

**Terminal**

- Kann eine Stand-Alone-Software oder Teil eines größeren Systems sein
- Arbeitet mit dem Multimedia-Content
- Besteht aus mehreren Teilen: Application, Compression Layer, Delivery Layer, Transmission/Storage Medium

**Transmission / Storage Medium**

- Entsprechen den tieferen Ebenen der sendenden Infrastruktur
- Übermittelt multiplexed streams an den Delivery Layer

**Delivery Layer (DL)**

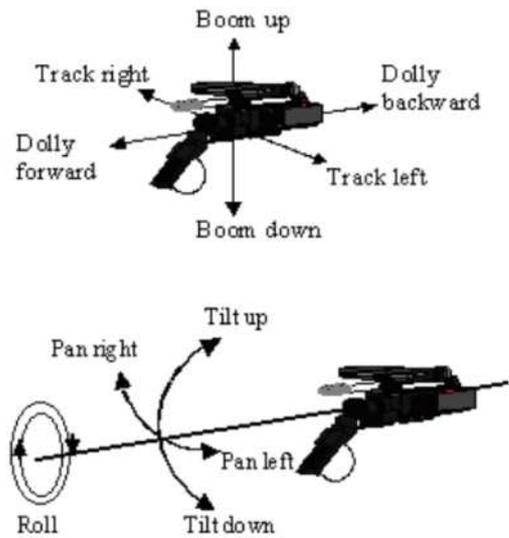
- Enthält Funktionen u.a. Synchronisation, Framing und Multiplexing von MPEG-7-Content
- MPEG-7-Content kann gemeinsam mit dem Multimedia-Content geliefert werden, der beschrieben wird (ist nicht immer der Fall)
- Access Unit: Einzelne MPEG-7-Datenblöcke, auf die individuell zugegriffen werden kann
- MPEG-7 elementary streams: enthalten verschiedene Informationen
  - Schema information: definiert die Struktur der MPEG-7-Description
  - Descriptions information: gesamte oder teilweise Beschreibung des Multimedia-Contents

**Compression Layer**

- Access Units werden aufgeschlüsselt und die Content-Beschreibung wird wiederhergestellt

### Deskriptoren für Kamera-Bewegung

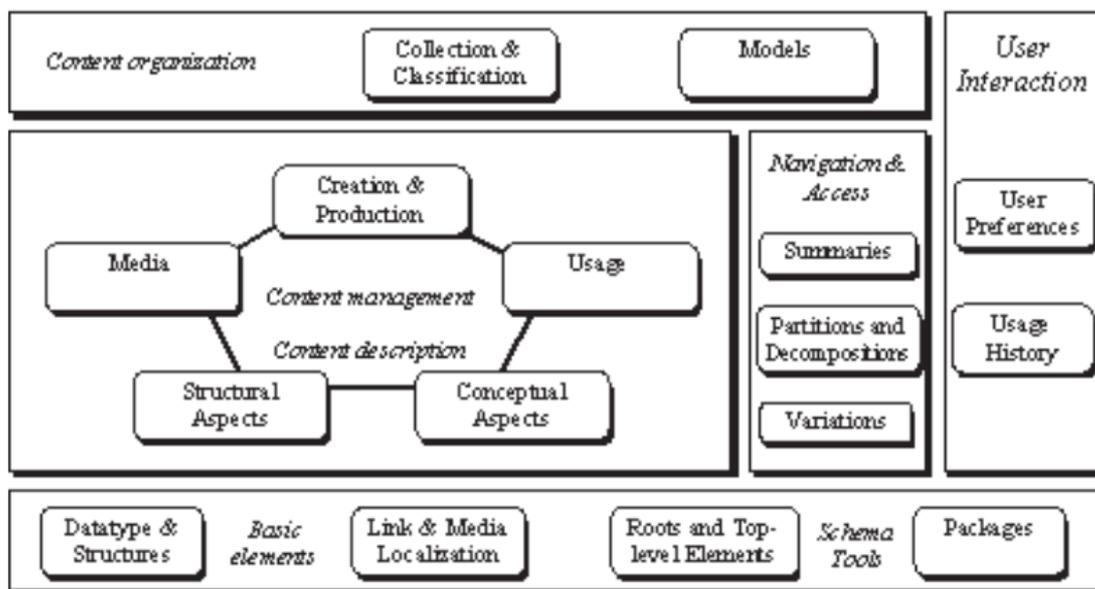
- Motion Trajectory:
  - „Flugbahn“ eines Objekts
  - Position eines Objekts in Zeit und Raum
  - Definiert als Liste von Keypoints (x,y,z,t)
- Parametrische Bewegung
- Activity Descriptor: beschreibt die „Intensität“ in einem Video-Segment, z.B. Torschuss bei einem Fußball-Match → „hohe Intensität“



### Lokalisierung

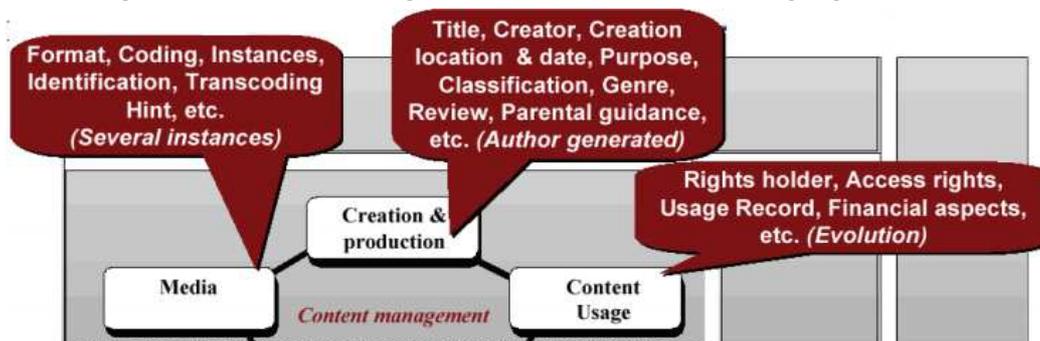
- Region Locator: Jedes Objekt wird als Box spezifiziert
- Spatio-Temporal Locator: beschreibt die Regionen in einer Video-Sequenz in Zeit und Raum

### MPEG-7 Multimedia Description Schemes (MDS)

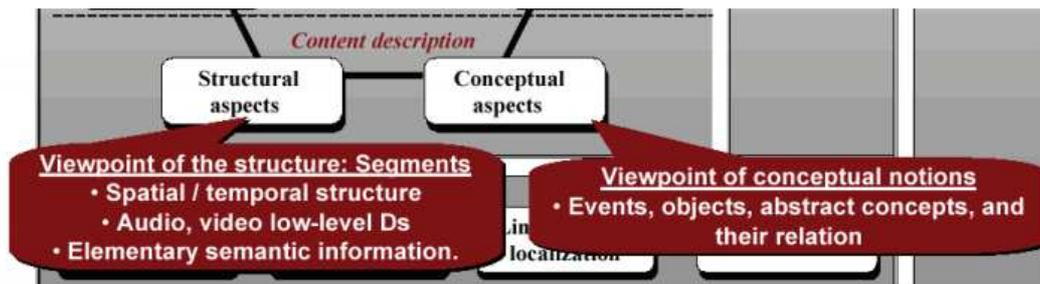


### Content Management (CM)

- CM-Tools ermöglichen die Beschreibung von Content (von seiner Erzeugung bis zum Konsum)



## Content Description



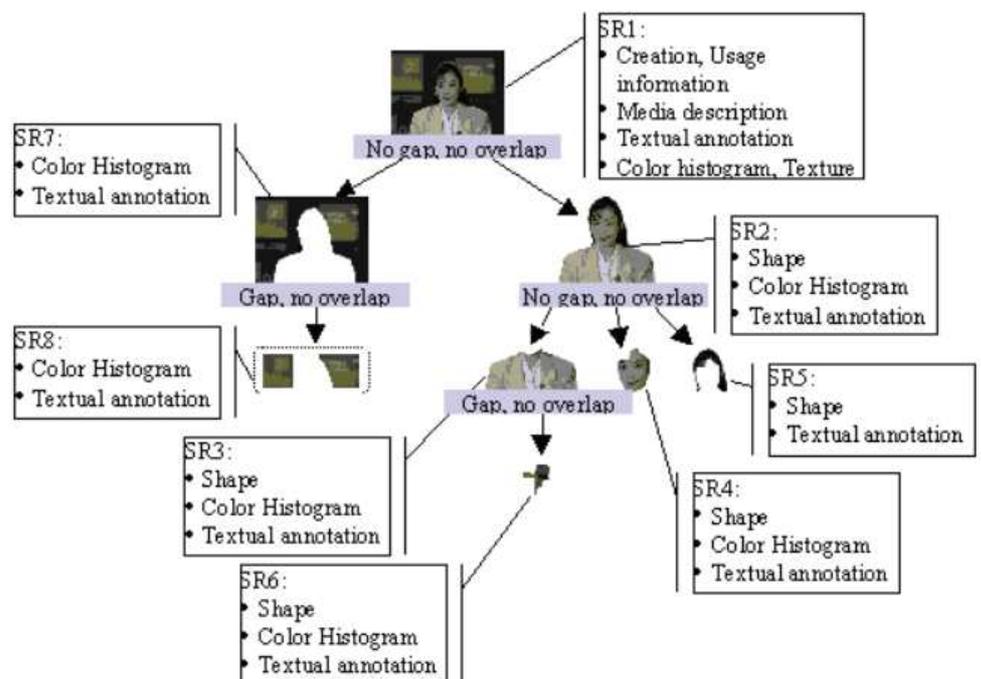
- Beschreibung der strukturellen Aspekte
- Wichtigstes Element: „Segment DS“

### Segment Description Scheme

- Beschreibung der physikalischen und logischen Aspekte von Multimedia-Content
- Segment: repräsentiert einen zusammenhängenden Abschnitt eines Multimedia-Contents
  - Temporal Segment: zeitlich zusammenhängend, z.B. Video-Segment, Audio-Segment
  - Spatial Segment (Still Region): Gruppe von benachbarten Pixeln
  - Spatio-Temporal Segment (Moving Region): zeitlich und räumlich zusammenhängend
- Segment DS hat 5 Subklassen zur weiteren Unterteilung:
  - AudioVisual Segment DS
  - Audio Segment DS
  - Video Segment DS
  - Still Region DS
  - Moving Region DS

### Segment Description

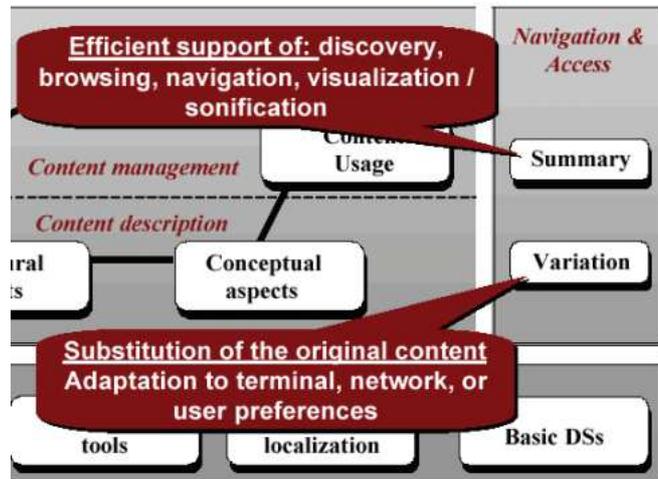
- Jedes Segment kann durch folgende Informationen beschrieben werden:
  - Creation information
  - Usage information
  - Media description
  - Textual annotation
  - Specific features
- Bsp Image Description: (siehe rechts)



- Segment Tree: effizienter Zugriff, leicht skalierbar
- Segment Graph: in Fällen, wo Segment Tree einschränkend wäre
  - Bsp: siehe Folien 691-693

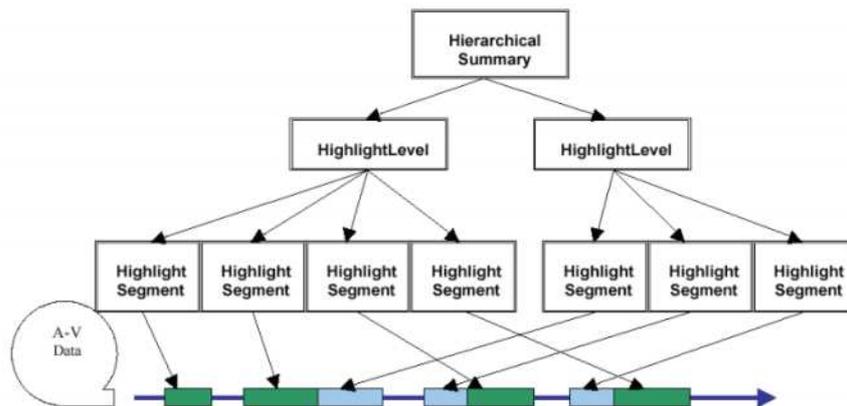
### Navigation & Access

- Summaries: ermöglichen effizienten Browsing, Navigation, Visualisierung
- Views & Partitions: zeigen Multimedia-Daten im Zeit- oder Frequenzbereich
- Variations



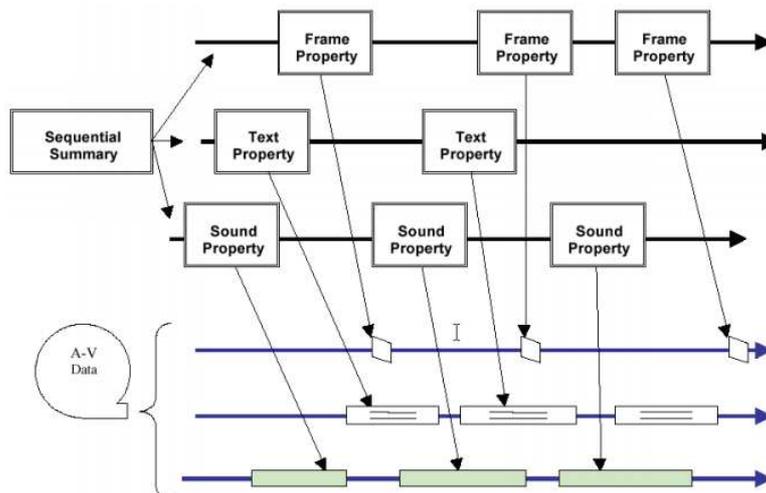
### Hierarchical Summary DS

- Unterteilt den Content nach Feinheitgrad in verschiedene Ebenen
- Highlightlevel DS: spezifiziert die Elemente der Hierarchical Summary DS



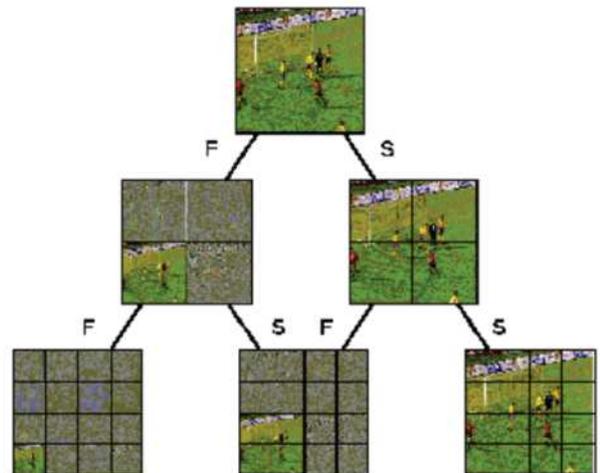
### Sequential Summary DS

- Sequenz von Audio-Clips, Bildern oder Video-Frames → Slideshow
- Kann unabhängig vom beschriebenen AV-Content (Audio-Video-Content = Multimedia-Content) gespeichert werden → schnellere Navigation und Zugriff, dafür höherer Speicherbedarf



**Partitions & Decompositions**

- Knoten-Arten:
  - Views in space: räumliche Segmente
  - Views in frequency: „Wavelet“-Subbänder
  - Views in space and frequency
- Ermöglicht effizienten „Multi-resolution“Zugriff



**Variation des Contents**

- Server, Proxy oder Terminal bestimmt die für die jeweilige Situation passendsten Version des Contents
- Variation DS:
  - spezifiziert die unterschiedlichen Variationen, z.B. verschiedene Auflösungen, Sprachen, ...
  - Variation fidelity value: Qualität der Variation im Gegensatz zum Original

**Content Organisation**



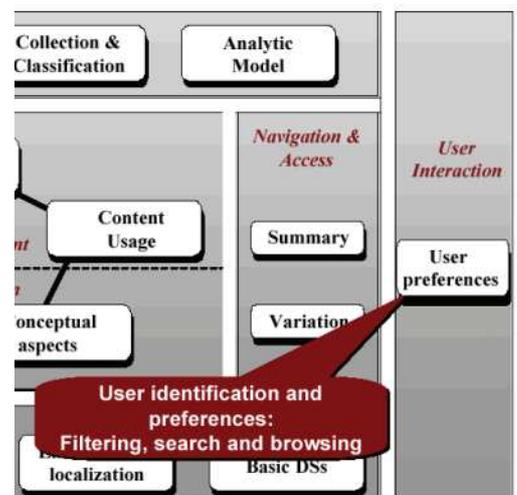
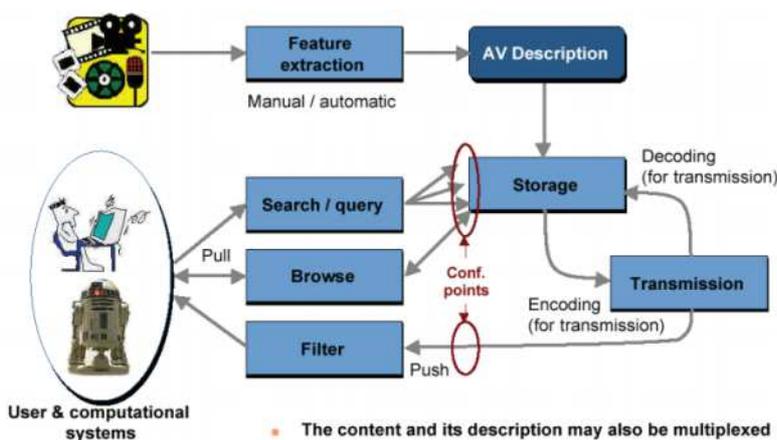
**Collection Structure DS**

- Gruppiert av-Content, Segmente, Events oder Objekte
- Spezifiziert die typischen Eigenschaften dieser Elemente und die Beziehungen zwischen Collections

**User Interaction**

- User Preference DS: Personalisierung von av-Content → effizienterer Zugriff (siehe rechts)

**Zusammenfassung – Information Flow:**



## VI. MPEG-4

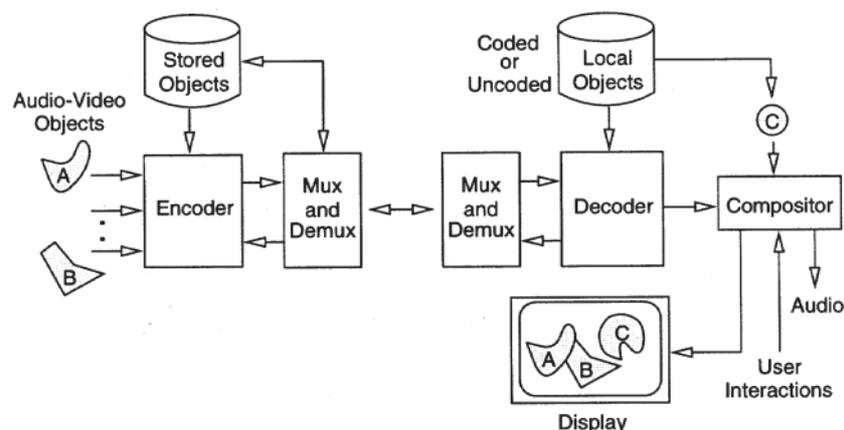
### Motivation

- Verbindung von 3 Service-Modellen: Communications, Interactivity und Broadcasting
- Standardisierte Algorithmen zur Kodierung von MM-Content
- Ermöglicht Interaktivität, hohe Kompression, Skalierbarkeit von Audio und Video
- Support für natürlichen und künstlichen MM-Content
- Nutzen für...
  - Autoren: bessere Wiederverwertbarkeit und Flexibilität, Schutz der Eigentumsrechte
  - Service Provider: für Netzwerke geeignet
  - Nutzen für End User: viele Funktionen, auf die einfach zugegriffen werden kann

### Prinzipien

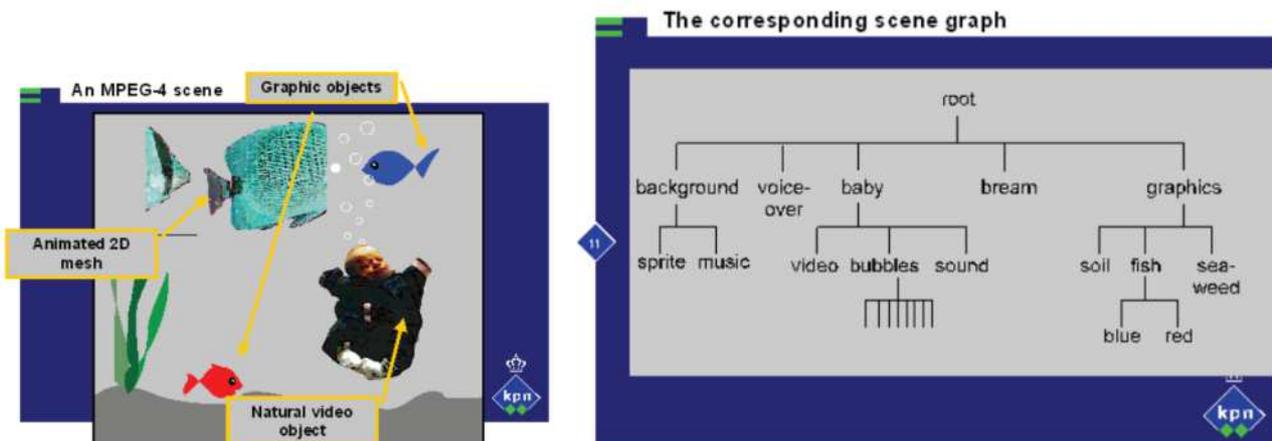
- Audio-Visual Scenes:
  - Zusammensetzung aus av-Objects entsprechend einer Scene Description
  - Erlaubt Interaktionen zwischen den Elementen
  - Einfache Wiederverwendbarkeit
- Arten von Audio-Visual Objects:
  - Natürlich (natürliche Audio- und Video-Aufzeichnungen) oder synthetisch (Text, Grafiken, ...)
  - 2D (z.B. Website), 3D (z.B. virtuelle Welt)
  - Stream oder Download
- Scene Description:
  - Beschreibt die Beziehungen zwischen den Objekten in Raum und Zeit und deren Verhalten
- Unabhängig von der Bitrate, werden komprimiert
- MPEG-4 stellt zur Verfügung:
  - Coding: Repräsentation von Audio-, Video- und av-Einheiten (Media Objects)
  - Composition: Erzeugung von av-Szenen
  - Multiplex: Multiplexing und Synchronisierung der Daten zum Senden über Netzwerke
  - Interaction: Interaktion des Empfängers mit der av-Szene

### MPEG-4-Architektur:



## MPEG-4 Systems

- Definiert das Framework zur Integration von MM-Komponenten
- Integriert die grundlegenden Decoder
- Stellt die Spezifikationen der Kompositions- und Multiplex-Teile des System zur Verfügung
- Composition:
  - Modell zur Beschreibung komplexer MM-Szenen
  - Basiert auf der Virtual Reality Modeling Language (VRML)
  - Abstufung zwischen 2D- und 3D-Content
  - Schnittstellen und Synchronisation von gestreamten Medien
  - Szene kann in binärer Darstellung kodiert werden
  - Szenen werden als hierarchisch geordnete Graphen repräsentiert (Blätter ... Media Objects)



### Spatial Composition

- Composition stream: wird speziell behandelt, da er Informationen zum Aufbau der Szene beinhaltet
- Spatial relationships:
  - Jedes Media Object hat sein eigenes internes Koordinatensystem
  - Media Objects werden entsprechend ihren Beziehungen im Scene Graph gruppiert

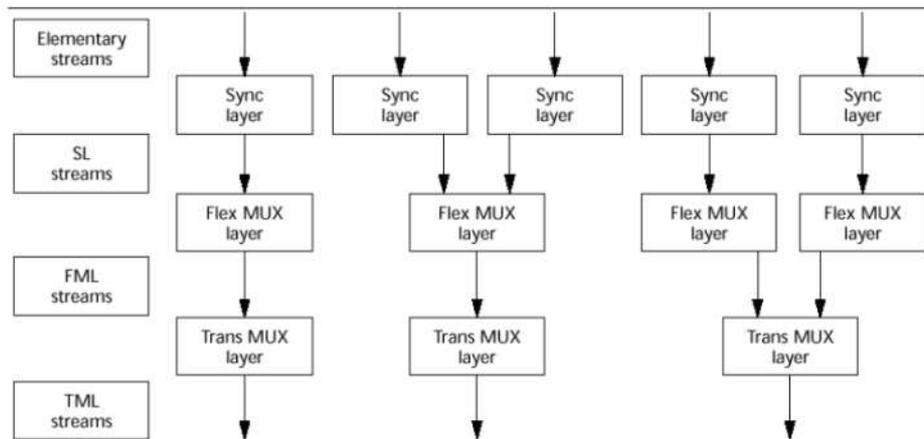
### Temporal Composition

- Composition stream (BIFS): hat eine eigene Zeitbasis
- Zeitbasen der Komposition und des Data streams müssen übereinstimmen
- Time stamps:
  - Decoding time stamp (DTS): zeigt an, wann ein Media Object beim Decoder Input bereit sein soll
  - Composition time stamp (CTS): zeigt an, wann ein Media Object beim Compositor Input bereit sein soll
- Time value: stellen Zeitdauern oder Zeitpunkte dar

### Multiplex

- Netzwerke haben stark unterschiedliche Performance-Eigenschaften
- Unterteilung von:
  - Synchronisation Layer: Timing- und Synchronisationsinformationen
  - Flexible Multiplex Layer: Multiplexen von Streams mit unterschiedlichen Eigenschaften
  - Transport Multiplex Layer: passt den multiplexed stream an das aktuelle Netzwerk an

## Multiplex-Struktur



### Synchronisation Layer

- Elementare Streams werden in Pakete unterteilt und Header-, Timing- und Synchronisations-Informationen hinzugefügt
- Header enthält: Sequence Number, Bit Rate, Object Clock Reference (OCR), Decoding Time Stamp (DTS) und Composition Time Spamp (CTS)

### Flexible Multiplex Layer

- Streams mit ähnlichen QoS-Anforderungen (Quality of Service) werden multiplexed
- Optional: „Intermediate Flexible Multiplex Layer“ für niedrige Bitraten

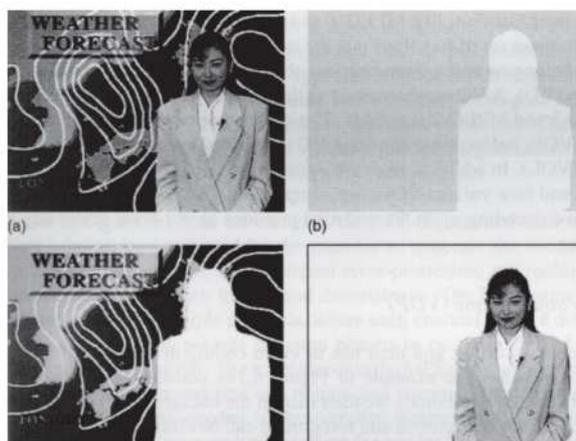
### Transport Multiplex Layer

- Fehlererkennung, Qualitätssicherung, ... → Transport Multiplex Layer (TML)
- Hängt vom Übertragungs- bzw. Speicherungssystem ab, an das die kodierten Daten geschickt wird

## MPEG-4 Video

### Video Codec

- Unterstützt sowohl rechteckige als auch beliebig geformte Video-Objekte und Skalierbarkeit
- Bit stream wird in einen Stand-alone base layer und mehrere Enhancement Layer aufgeteilt
- Jeder Frame wird in mehrere Regionen unterteilt → „Video Object Planes“ (VOP)
- Video Object (VO): Aufeinanderfolgende VOPs, die zum selben Objekt in einer Szene gehören
- Video Object Layer (VOL): enthält die Informationen zur Form, Bewegung und Textur einer VOP
- Bsp VOP:



## Interactivity

- Zwischen User und Encoder-Level: Änderung der Bitrate, des Multiplexings, des Composition scripts, etc.
- Zwischen User und Decoder-Level: Änderung eines Teils des Bit Streams, der Position von Video-Objekten, etc.

## Media Integration of Text and Graphics (MITG)

- Bilder und Video Objects können auf verschiedene Art im Scene Graph platziert werden
- Scene Graph kann Audio-Quellen enthalten
- Text kann eingefügt und angepasst werden (Schriftart, Größe, etc.)
- 3D-Graphiken

## Gesichts-Animation

### Facial Animation Parameters (FAPs)

- Vordefinierte Parameter, unabhängig vom verwendeten Gesichtsmodell
- Beschreiben Bewegungen der Gesichtsmerkmale
- Viseme: definieren die Position des Mundes beim Sprechen von bestimmten Lauten
- Ausdrücken von Emotionen
- Facial Definition Parameters (FDPs): werden zum Kalibrieren von Gesichtsmodellen verwendet

### MPEG-4 Text-to-Speech (M-TTS)

- Enthält Informationen wie: Geschlecht, Alter, Sprechgeschwindigkeit, etc.
- Unterstützt fast-forward, pause, play, rewind, ...
- Wird an die Gesichtsanimations-Engine im MPEG-4-Player übergeben → Face Animation

## B-Frames

- MPEG-2-Modi:
  - Vorwärts-Modus: Referenzframe ist vorheriger I- bzw. P-Frame → Vektor + Fehlerbild
  - Rückwärts-Modus: Referenzframe ist nachfolgender I- bzw. P-Frame → Vektor + Fehlerbild
  - Interpolations-Modus: Vorwärts- und Rückwärts-Vektoren werden übertragen; Fehlerbild resultiert aus Interpolation beider Referenzwerte
- Erweiterung durch MPEG-4:
  - Direkter Modus: Vorwärts- und Rückwärts-Vektoren werden von einem einzelnen Vektor abgeleitet → „Delta-Vektor“

## Adaptive Quantisierung

- Jedem Makroblock wird individuell quantisiert (abhängig von psychovisuellen Aspekten)
- Bsp: Artefakte bei sehr hellen/dunklen Bereichen weniger sichtbar → stärkere Quantisierung

## Viertel-Pixel Bewegungskompensation

- „Ein-Pixel“-Genauigkeit:
  - Makroblöcke können sich nur um ganzzahlige Pixelwerte verschieben
  - Hoher Prädiktionsfehler, schlechte Kompression
- MPEG-4 → „Viertel-Pixel“-Bewegungskompensation
  - Virtuelle Viertel-Pixel werden mittels Interpolation berechnet
  - Bewegungsvektoren haben größere Genauigkeit
  - Realitätsnaher, bessere Qualität
- Auch möglich: „Halb-Pixel“-Genauigkeit

## Sonstige MPEG-4 Erweiterungen

- Globaler Bewegungsausgleich
  - Für eine Video Object Plane (VOP) werden bis zu 4 Bewegungsvektoren berechnet
  - Effektivere Kompression bei Kamerabewegungen
  - Encoder entscheidet zwischen Makroblock-Vektoren oder globalen Vektoren
- Alternative Scan-Modi
  - MPEG-4: Alternativer horizontaler und alternativer vertikaler Scan, siehe Folien 762,763
- Quantisierungsmethode
  - User kann zwischen H.263, MPEG und eigener Quantisierungstabelle wählen

# VII. Multimedia Programming Abstractions

## Programming Environments

- Auswahl des Frameworks:
  - Welche Programmier Techniken, Methoden, etc. benötigt werden
  - Welche Services angeboten werden
  - Welche Medientypen und Standards unterstützt werden

## Objektorientierte Multimedia-Programmierung

- Kapselung, Erweiterbarkeit, Cross-Platform Development

## Framework-Entwicklung

- Definition von „Abstractions“: Funktionen aus der Audio- und Videoproduktion
- Collaborating Objects: Strukturierung von typischen Komponenten
- Architecture: Spezifikation der Zusammenarbeit von Objekten
- Classes: für unterschiedliche Zwecke und die jeweilige Hardware-Plattform

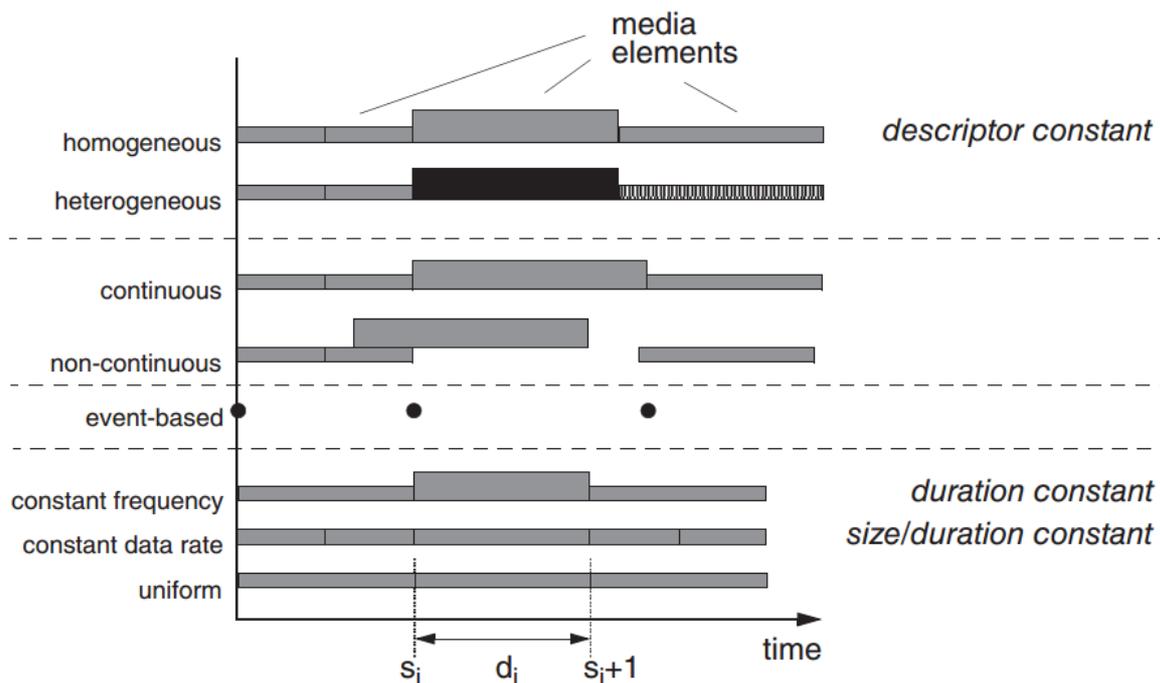
## Abstractions

### Time-Object

- Überbegriff für Media processing elements und zeitabhängige Medien
- Sollte vom System unterstützt werden
- Continuous time value: kontinuierlicher Zeitwert
- Discrete Time Coordinate System (DTCS): Abbildung von diskreten auf kontinuierliche Zeitwerte

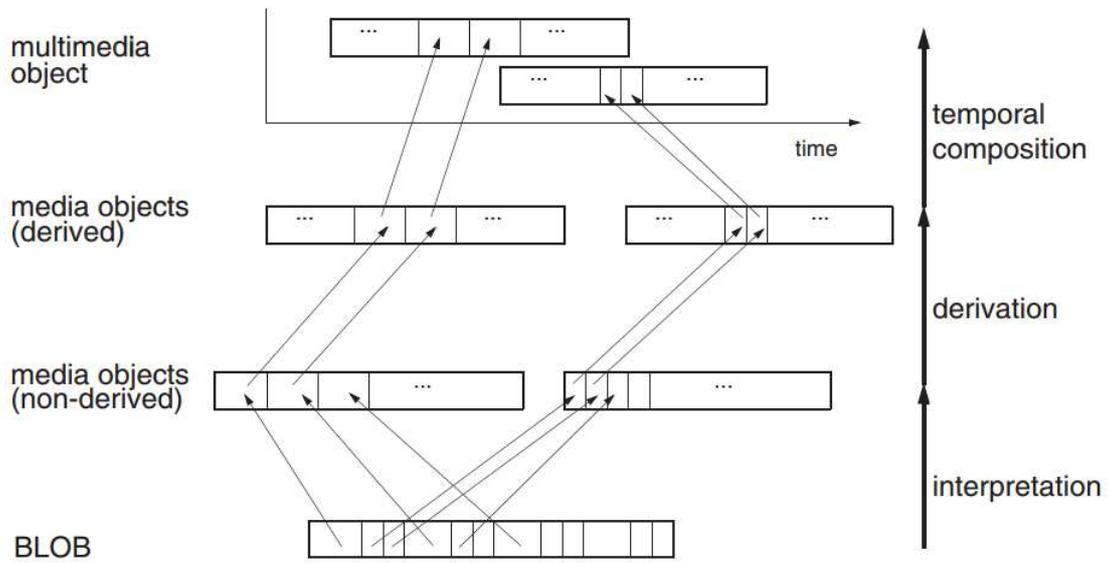
### Timed Streams

- Kommunikationskanal mit Echtzeit-Garantie und kontinuierlicher Übertragung von Medienelementen (samples)
- Sollte als Multicast realisiert sein und sowohl asynchrone und isochrone Übertragung unterstützen
- Media Type: spezifiziert die Kodierungs-Daten für ein Medium, z.B. Audio, Video, Image
- Media Element: eigenständige Dateneinheit; hat einen Media Type, eine bestimmte Größe (in Bytes) und einen Descriptor
- Timed Stream (Media Object): eine endliche Sequenz an Mediendaten
- Timed-stream Type: ein Tupel  $T = \langle M, D, R \rangle = \langle \text{Media Type, DTCS Type, Einschränkungen (Rules)} \rangle$
- Eigenschaften: Homogenität, Kontinuität, Frequenz, Datenrate, Uniformität
- Bsp:



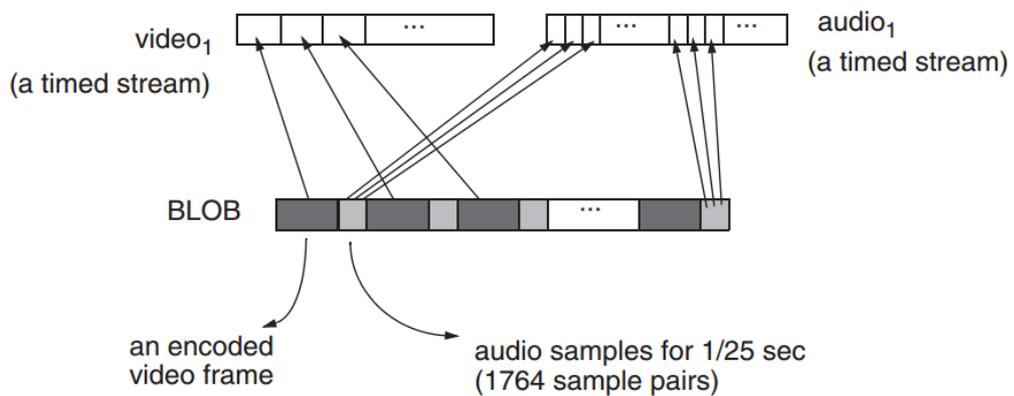
- Interpretation von zeitabhängigen Medien als:
  - Pools: für Medienunabhängige Operationen, z.B. Kopieren
  - Streams: für Medienabhängige Operationen, z.B. Bearbeitung; Timing muss synchron sein

### Interpretation - Derivation - Composition



#### Interpretation

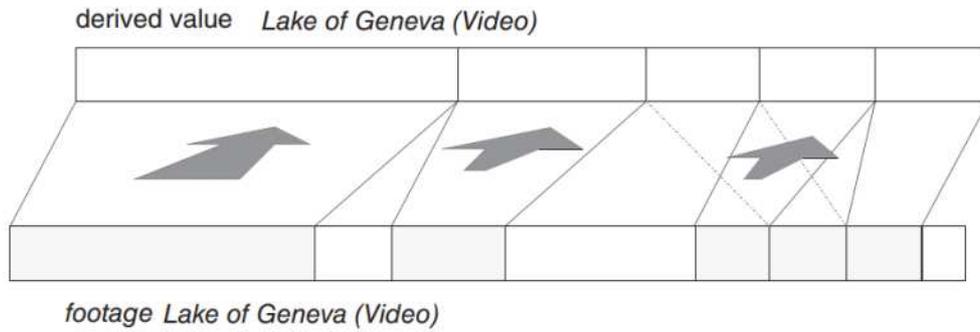
- Interpretation eines Streams (BLOB): Abbildung des Streams auf eine Menge von Medienelementen
- Problemfaktoren: Heterogenität (verschieden große Elemente), Interleaving und Padding, etc.
- Struktur-Information darf nicht von den Daten getrennt werden
- Bsp:



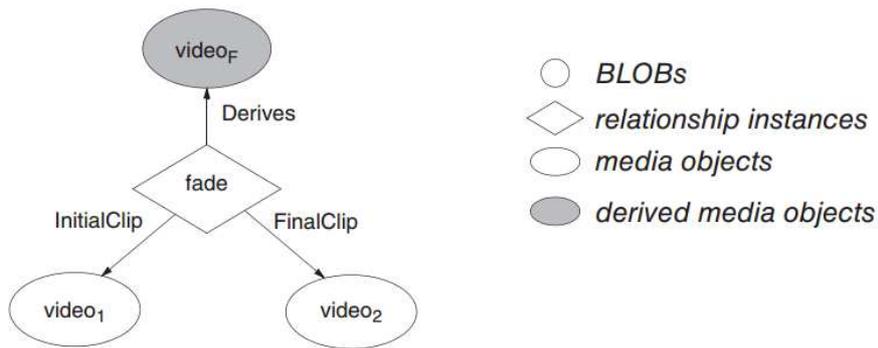
- Mapping:
  - video1(elementNumber, elementSize, blobPlacement)
  - audio1(elementNumber, blobPlacement)

**Derivation**

- Ableitung eines Objekts von einem anderen Objekt
  - Derivation: Herkunft
  - Derivation Object: Ursprungs-Objekt
  - Derived Objekt: abgeleitetes Objekt

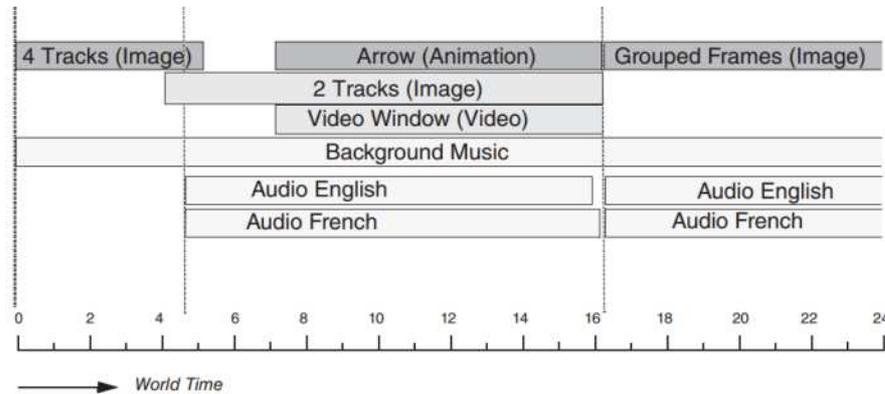


- Vorteile: geringerer Speicherplatz, Datenunabhängigkeit, effizientere Modifikation
- Datendiagramm:
  - Zur Visualisierung und Optimierung

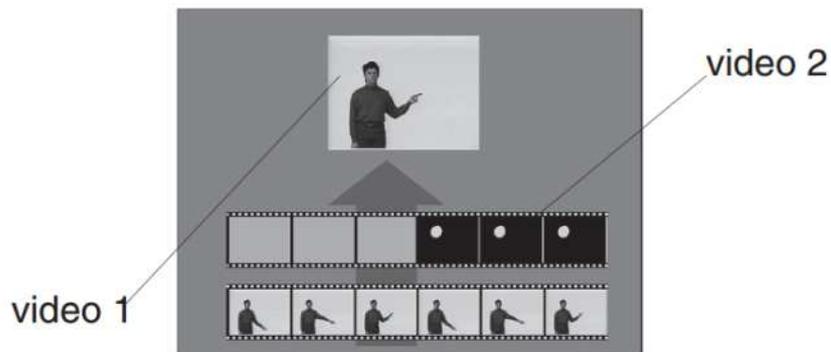


**Composition**

- Anordnung von Media-Object-Gruppen (Komponenten) in Raum und Zeit
- Sollte intuitiv, einfach und flexibel anwendbar sein
- Ergebnis der Composition: „Multimedia Object“
- Temporal Composition:

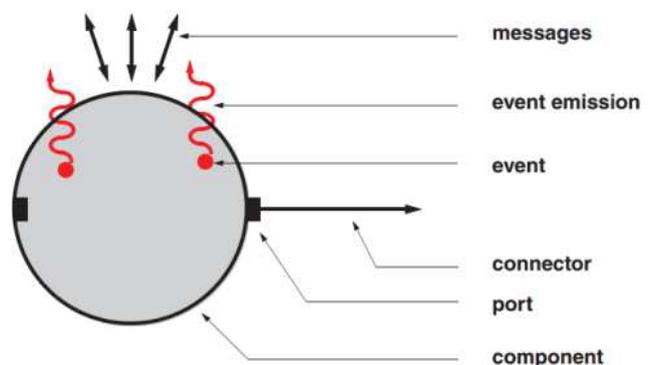


- Spatial Composition:
  - Spezifikation des Layout, Transformationen, ...



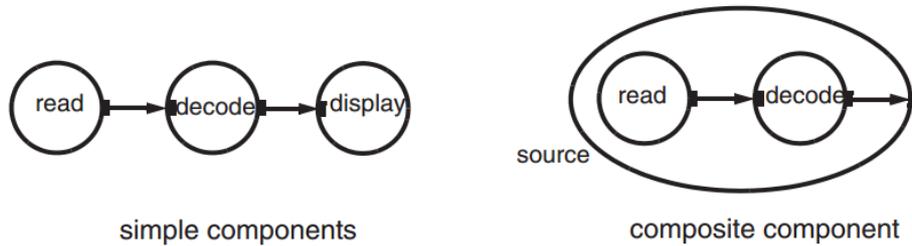
**Components (MPEs)**

- Produzieren, konsumieren und/oder transformieren Timed Media Streams
- Interfaces: synchron, asynchron oder isochron
- Beispiele:
  - Kamera: produziert Video
  - Video Codec: transformiert Video
  - Audio-Recorder: konsumiert Audio
- Port Connection:
  - 1 Input + 1 Output-Port
  - Müssen „plug compatible“ sein



## Configuration

- Zusammenführung von MPEs, Streams und Medien in eine Applikation
- Sollte dynamische Änderungen und eine beliebige Anzahl an Komponenten unterstützen



## Quality of Service (QoS)

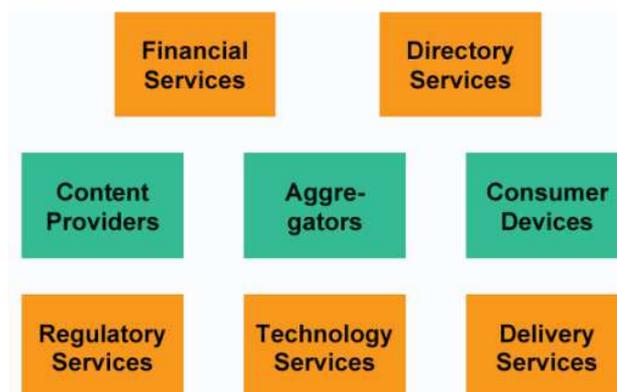
- Beschreibt die Anforderungen der jeweilige Ressource mittels „Fuzzy“-Parameter
- Bezieht menschliche Wahrnehmung mit ein
- Ermöglicht Ressourcen-Management
- QoS-Parameter: z.B. maximale Auflösung, erlaubte Fehlerrate, Jitter- und Delay-Grenzen

## MPEG-21 – MM-Framework

### Kontext

- Neue internationale Kommunikationsnetzwerke (z.B. Internet) → Umdenken bei traditionellen Business-Modellen notwendig
- Standards für Infrastruktur und das Handeln von digitalen Ressourcen notwendig

### Komponenten eines MM-Frameworks:



- Zwischen diesen Komponenten: Austausch von Content, Content-Informationen, Verwendungsrechten, Geld, Authentifikationen, ...

### Current Practice

- Heutzutage: Medien enthalten implizite oder explizite Rechte, z.B.:
  - Buch: darf gelesen und wiederverkauft werden
  - CD: darf abgespielt werden, aber nicht kopiert
  - Öffentlicher Broadcast: darf angesehen/angehört werden, da Lizenzgebühren bereits bezahlt wurden

## Future Practice

- Multimedia-Framework ermöglicht unbegrenzte Flexibilität
- MPEG-21-Item darf nach dem Kauf z.B. ...
  - ... einmalig auf ein Mobilgerät kopiert werden
  - ... für 24 Stunden ausgeliehen werden
  - ... nur 10 mal abgespielt werden

## MPEG-21 Vision Statement

- Integration von verschiedenen Standards und Schnittstellen
- Füllen von eventuell vorhandenen Lücken in der Infrastruktur
- Transparente und erweiterte Verwendung von MM-Ressourcen in verschiedensten Netzwerken und Geräten
- Use Cases: siehe Folien 817, 818

## MPEG-21 Spezifikationen

- Digital Item (DI): Ein strukturiertes digitales Objekt innerhalb des MPEG-21 Frameworks
- User:
  - interagiert mit dem MPEG-21 Framework und verwendet Digital Items
  - z.B. Einzelpersonen, Firmen, Organisationen, ...
  - Rollen: Erzeuger, Konsument, Rechteinhaber, ...
  - Jede Rolle erhält spezielle Rechte und Pflichten

## Teile des Standards

- Vision, technologies, strategy
- Digital Item declaration (DID)
- Digital Item identification (DII)
- Intellectual property management and protection (IPMP)
- Rights expression language (REL)
- Rights data dictionary (RDD)
- Digital Item adaptation (DIA)
- Reference software
- File format
- Digital Item Processing (DIP)