geschrieben/verfasst von Fis Inf-Forum: Torason
bearbeitet von Schwarz Gerald & Pichler Thomas

Kapitel I: Media Types

Media classification

- **Perception Media**: help humans perceive their environment
- **Representation Media**: characterized by computer representation of info
- **Presentation Media**: tools and devices for I/O
- **Storage Media**: data carriers for info storage
- **Transmission Media**: info carrier for continuous data transmission (not storage!)
- **Information exchange Media**: info carrier for both

Representation

- **Representation Spaces**: where goes the output (zB Screen)
- **Representation Dimension**: number of dimensions (zB Screen=2, stereophony=3)
- evtl. additional Dimension: **Time**; two media types:
  - time-independent (**discrete**) media (zB text, graphics)
  - time-dependent (**continuous**) media (zB audio, video), processing is time-critical

Multimedia

can produce, manipulate, present, store and communicate information, which is encoded at least through 1 continuous and 1 discrete medium.

Media Types

- **Text**: presented via ascii/iso charsets, hypertext, …; operations: string operations, encryption, spell checking, pattern matching/searching, compression, editing
- **Images**: (a)chromatic light
  - Color: Depends on surface, surroundings, visual system, light sources
  - eye sees small changes of small intensities like big changes of big intensities
  - dynamic range: ration between min/max intensities
  - enough intensities: $n=\log_{1.01}(1/\text{lowest possible intensity})$
  - halftone-approximation/dithering: 
  - **color models**:
    - CIE (langgezogenes), RGB (hardware-oriented), CMY, HSV (Kegel), YUV (luminance, UV)
  - **image representations**: # channels, channel depth (bits per pixel), pixel aspect ratio, interlacing (wikipedia-dings), compression lossy or lossless, indexing (color maps, speichere Tabellenzeile, wo Farbe steht, statt Farbe)
  - **operations**: edit, color correction, filter, composite (alpha channels used), geometric transformations, (format) conversion
- **Video**: sequence of frames
  - **characteristics of analog video**: frame rate, scan lines, aspect ratio (4:3, 16:9), interlacing (not i.=progressive), signal quality (consumer, professional, broadcast), composite/component, stability, audio tracks, tape size
  - **video representations**: component video (each primary has separate signal), composite video (signals mixed), S(eparated) –Video (compromise)

- **analog video tape formats**: VHS, Betamax, Video8, S-VHS, Hi8, Umatic, Betacam, 1 inch
- **video equipment**: routing switcher (routes 1 input to several outputs), distribution amplifier (split 1 signal to more), timebase corrector (removes timing errors), sync generator (is master clock), frame (delay) buffer (synchronizes external sources), video production switcher
  - vid.prod.sw. makes HQ visual images, n inputs, 1 output, special effects, früher HW, jz meistens SW; can: keying, blending, wipes, key framing
- time codes: HH:MM:SS:FF (SMPTE time code)
- digital video:
  - raw size: 50sec=1GB, low data rate: 1GB=30min – 1h
  - digital video uses discrete numeric values, signal is sampled, frame is represented by pixel array
    - line sampling: zB 4:2:2 (4x luminance, 2x pro chroma)
- **compression**: lossy/lossless, real-time (symmetric), spatial vs temporal, scalable, type of source material
- **video quality** comes from: frame size & depth, (key) frame rate, source material, algorithm (parameters), compressed data rate & file size
- **Operations**: storage & retrieval (zB CD-Rom), editing, digital viedo effects, conversion, compression

Akustik

- Schall: wellenförmige Ausbreitung v. Druckschwankungen, Schwingung
- Rest: eh klar

Digital Audio

- Vorteile gegenüber analog: weniger Rauschen, mehr Dynamik, verlustloses Kopieren, höhere Linearität, temperaturunempfindlich, keine Gleichlaufschwankungen
- Nachteile: anfällig für Datenverlust, höhere Übertragungsbandbreiten, HW-aufwendig
- **Digitalization**: get audio/video in computer via sampling (Abtasten) & quantization
  - Nyquist-Shannon sagt: Abtastfrequenz mind. doppelte höchste Signalfrequenz
- Signal to Noise Ration (SNR): Schalldruckpegel L_{max} - L_{noise}
- MIDI (Musical Instrument Digital Interface)
  - protocol, enables synthesizer, keyboards etc to communicate
  - **synthesizer**: sound generator
  - **sequencer**: programm, stores MIDI data
  - **terminology**: track, channel (for separating info in MIDI system, 16 channels in one cable), voice (produces sound in synthesizer), key number (are 128, for notes), controller (specifies operational characteristics), patch/program (control settings for particular timbre (Klangfarbe))
  - **Important concepts**: timing clocks, MIDI synchronization, MIDI time code

Kapitel 2: Compression

- **entropy coding**: lossless, media's specific characteristics wurscht, data taken as digital sequence

- **source coding**: lossy, takes in account data semantics, degree of compression depends on data content
- **Decompression requirements**: dialog mode needs symmetric, retrieval mode asymmetrtic compressions (once compressed, decompressed often and must be fast)
- **Compression steps**: prepare picture (make digital, make 8x8 blocks), process it, quantizise it (map real numbers to integers), entropy coding
- **Run-length-encoding**: =entropy, ABCCCCCCD → ABC!9D
- **Huffman encoding**: =statistical, Wahrscheinlichkeitsbaum, Huffman table for en/decoding
- **Differential encoding**: =source coding, calculates difference between zB pixels, supresses zeros by run-length-coding

JPEG

- **requirements**: good compression, though good quality, user can select parameters, for every image, reasonable for HW & SW, run on many platforms, support:
  - sequential (same order as scanned), progressive (multi-pass), lossless, hierarchical (at multiple resolutions) encoding
- **prepare image**: divide into data blocks (lossy: 8x8 pixel, lossless, 1 pixel), processing sequential or interleaved (there: combine data units to minimum coded units, MCU), values in range 0–255
- **process image**: shift values to -128–127, Discrete Cosine Transform
  - Lowest frequency=DC Coefficient, determines fundamental color of block, frequency in both directions=0, rest of block are AC coefficients with frequency in both directions!=0
  - humans don't perceive high frequencies good, so compress high frequencies! high=fine structures, low=outlines
  - DCT lossy, not exact value
- **quantizise image**: goal: throw out bits, divide coefficient values by N and round, truncate according to quantization tables (per default luminance table/chroma table)
- **entropy coding**: map 8x8 pixel block in 64-element-vector via zig-zag-scan, process DC & AC coefficients, do run-length on zero ACs, do Huffman to DCs and ACs (get Huffman code from SSSS=number of bits needed to encode Differences)

MPEG

- **Objectives**: good quality at compression rate betw. 1 and 1,5 Mbps, support symmetric or asymmetric compress/decompress applications, random-access playback, fast-forward, fast-reverse, normal-reverse etc available, maintain audio/video synchronization, avoid catastrophic behaviors at data errors, control (de)compression delay, be editable, flexible to play video in windows, enable low-cost-chipsets for real-time-encoding
- **Standards**: MPEG-1 (low data rates, VHS quality at 1.5 MBits/sek), MPEG-2 (higher quality, higher data rate, 15 MBits/sek), MPEG-4 (verly low bit rate, <64 kbs, with small images)
- **MPEG in general**: standard, defines audio & video coding and system data streams with synchronization, considers functionalities of other standards, uses JPEG, provides info on aspect ratio, frequencies of refreshment (in Hz)
- **MPEG-1 in general**: compression designed for CD-Rom speeds, audio compressor uses subband coder with psychoacoustic model, video compressor uses block transform coder with motion-compensated inter-frame coding; must distinguish syntax and semantics in the standard)

- **MPEG Video Standard**: each image 3 components: 1 luminance, 2 chrominance with half resolution; 8 bits for each component, divide image intro macroblocks (=16x16 px for luminance, 2x 8x8 px foc chrominance components)
    - **4 image coding types**: for efficient coding, high compression, exploit terminal redundancies of subsequent frames
        - **I-Frames** (intercoded): no reference to other frames, like still images, JPEG is used for them, low compression rate, points of random access, 8x8-blocks within a macroblock, DCT performed on them
        - **P-Frames** (predictive-coded): know previous I-Frame or P-Frame, make use of successive images with areas that do not change or are shifted, → temporal redundancy (determine last P/I-frame that is most similar), use motion estimation method (with motion vector: look for a match of that area within a search window), apply DCT to unreduced macroblocks, consists of I-Frame macroblocks and predictive macroblocks, are quantized and entropy encoded (Run-length)
        - **B-Frames** (Bi-directionally predictive coded): know previous and following I- and/or P-Frames, gets motion vector from both directions
        - **D-Frames** (DC-coded): can be used for fast-forward/fast-rewind, DC-parameters are DCT-coded, ACs wurscht, D-Frames have lowest frequencies of an image, only in MPEG-1 used
    - Decoding: image order differ from encoding
    - **Quantization**: ACs of B & P-Frames large values, I-Frames smaller ➔ adjust MPEG Quantization bitch! if data rate over threshold, larger steps in quantization, if it decreases, finer quantization
- **How to compare**: performance-mäßig (en/decoding time, size and fps, compression factor, image quality), funcionality-mäßig (multiple resolutions, constant or variable bit rate, easy to edit?)
- Decoding easy, encoding expensive
- **MPEG-2 video standard**: extensions for more applications, enhancements: efficient coding for interlaced videos (zB 16x8 blocks), 10-bit DC coefficients (statt 8), allow to divide a signal into 2 or more coded bit streams for different resolutions (like youtube?), picture quality or picture rate; more aspect ratios, 4:2:2/4:4:4 macroblocks, progressive & interlaced frame coding, more prediction modes, 4 scalable modes, better picture quality
    - **scalable modes**: spatial scalability (codes a base layer at lower sampling dimension =resolution), data partitioning (breaks the 64 quantized coefficients into 2 bit streams, first one (higher priority) has more critical, lower frequency coefficients), SNR scalability (code channels at same sample rates, bit different quality), temporal scalability (higher priority bit stream codes video at lower frame rate, 2nd stream comes from reconstructing and predicting)
- **Profiles and levels:** Profiles: Defined subset of bit stream syntax, level is defined within profile as a set of constraints on the bit stream parameters (zB resolution, max. bit rate), have hierarchical relationship
    - Profiles: simple, main, 4:2:2, SNR, spatial, high, multiview
    - Levels: low (SIF), main (CCIR 601), high-1440, high (HDTV)

MP3

- open standard (specifications available for fee, patents fair, not owned by a company, format well defined), available for en/decoders, widespread
- **basic tasks**: efficient compression, reconstructed audio should sound the same, low complexity, flexibility for applications

- **Basic perceptual audio encoder** consists of: filter bank (decompose input signal into subsampled spectral components), perceptual model (masking threshold by rules of psychoacoustics), quantization and coding (keep noise low here), encode bitstream

Kapitel 2.5: Spektrogramm, Hören und Audiokodierung

Spektrogramm

- **Probleme Signaldarstellung**: Zeitdomäne zeigt keine Frequenz-Anteile, Frequenzdomäne keine Zeit → daher kombiniere zu Spektrogramm! Schwärzung eines Punktes: Energie der Frequenz zu der Zeit, Analysen: Auftreten von Frequenzen, Musik vs Geräusch
- **Vorgangsweise**: verarbeite Eingangssignal blockweise, verwende überlappende Segmente des Signals, Fensterfunktionen geben Ausschnitte an, auf die sich die Analyse konzentriert (kann mehrere Formen haben): multipliziere punktweise den Signalblock mit der Fensterfunktion, resultierendes Signal → Fouriertransformation, reihe die Spektralvektoren → bekomme Zeit-Frequenzdarstellung
- Spektrogramm: häufige statistische Interpretation als nicht normalisierte Dichtefunktion; erlaubt Berechnung statistischer Parameter, zB Lagemaße, Streuung

Grundlagen der MPEG Audiokodierung

- **Hörbereich**: Schmerzschwelle und Sprachbereich für MPEG wurscht, wichtig: Hörschwelle: Ton muss über Mindesschallpegel (frequenzabhängig) liegen; Hörschwellenkurve=Mindestschallpegel/Frequenz=Ruhehörschwelle?
    - **Ruhehörschwelle**: Messung durch Testperson, gibt an, wann sie einen Ton gerade noch hört

Das Ohrli

- **Innenohr**: komplexer Hohlraum im Felsenbein, gefüllt mit Flüssigkeit, darin dünnwändiges Labyrinth mit Endolymphe, 2 Öffnungen zum Mittelohr: Ovales Fenster (überträgt Schallwellen) und rundes Fenster (dämpft Schwingungen der Cochlea)
- **Hören** geht so: Schallwellen treffen Trommelfell, werden übertragen auf die Gehörknöchelchenkette im Mittelohr, das letzte Glied bewegt Fußplatte des Steigbügels und das Ovale Fenster, daher schwingt die Endolymphe in der Vorhoftreppe
    - **Hörschnecke**: 3 Gänge mit Flüssigkeit: Vorhoftreppe, Schneckengang, Paukentreppe; letztere zwei werden von Basilarmembran getrennt. Die hat 15.000 Haarzellen, die äußeren 3 Reihen verstärken die Schallwanderwellen, die innere Reihe macht aus den Schwingungen Nervenimpulse.
    - **Basilarmembran** gerät für unterschiedliche Frequenzen an unterschiedlichen Stellen in Resonanz, wegen ihrer Form. Dort werden die Schwingungen verstärkt und die inneren Haarzellen damit stärker stimuliert.
        - Die Membran ist in 24 Frequenzgruppen aufgeteilt, Impulse auf einer Gruppe werden gemeinsam ausgewertet.
        - Daher kommt der psychoakustische Effekt des Gehörs: Ähnliche Frequenzen gleichzeitig gespielter Töne können nicht unterschieden werden → **Maskierungseffekt**, man hört nur lauteren Ton.
            - Maskierungsschwellwert: min. Lautstärke, die der leise Ton bräuchte, um nicht maskiert zu werden

- Maskierung hängt von Lautstärke, Frequenz und Zeit(dauer & intervall) des maskierenden Tons ab

MPEG – Maskierung, Kodierung

- **Maskierung bei MPEG**: mehrere Frequenzanteile pro Audiosignal, jeder beeinflusst Hörschwelle, weil Audiosignal variiert, psychoakustisches Modell berechnet immer aktuelle Hörschwellkurve, damit kann so quantisiert werden, dass Quantisierungsrauschen immer gerade unter der Hörschwellenkurve ist, wo maskierung: grobe Quantisierung
- **Filterbank**: teilt Audiosignal in Frequenzbänder auf (Subbands), ideale Filterbank: Subbands wären identisch mit Frequenzgruppen des Ohres
  - **MPEG Polyphase Filterbank:** teilte Signal in 32 Subbands, alle gleiche Bandbreite → einfacher Aufbau, aber untere Subbands umfassen mehrere Frequenzgruppen, Überlappungen bringen Infoverlust
  - Abtastrate wird auf Subbands aufgeteilt, zb 48 KHz in 4*12 → **kritische Abtastung**, sorgt dafür, dass Filterbank keinen Einfluss hat auf Datenmenge
  - **Subbandanzahl/Kompression**: Untersuche für jedes Subband, wie hoch das Quantisierungsrauschen sein darf, das noch maskiert wird; schmale Subbands → hoher Rauschen-Maskierungsschwellwert, hohe Kompression, je mehr Subbands, desto schmäler, desto höhere Kompression
- **Psychoakustisches Modell**: analysiert Audiosignal auf Maskierungseffekte, simuliert menschliches Gehör, berechnet Maskierungsschwellwert und erlaubtes Quantisierungsrauschen (alles pro Subband), ist verdammt wichtig für Qualität der Kodierung, MPEG bietet 2 Beispielimplementationen, eine schnelle und eine genaue mit komplexeren Algorithmen
  - Ton und Geräusch haben verschiedene Maskierungseigenschaften, daher braucht Psycho verschiedene Algorithmen zum Unterscheiden (Modell 1: starke Spitzen im Frequenzspektrum=tonal, Rest= Geräusche, Modell 2 benützt Phaseninformationen von einer Fast Fouriertransformation zur genaueren Berechnung der Maskierungsschwellwerte)
- Problem: **Kompression bei steilem Signalanstieg**: erzeugt Pre-Echo-Rauschen, MP3 verbessert das, AAC noch mehr
- **Hohe Frequenzauflösung bei langem Zeitfenster**: gut: viele Subbands (gut an Hörschwellenkurve anzupassen), gute Kompression; schlecht: Quantisierungsrauschen länger nicht änderbar (im ganzen Fenster halt), blöd bei Signalanstiegen
- **Geringe Frequenzauflösung bei kurzen Zeitfenster**: genau umgekehrt
- **Skalierungsfaktor**: bei MPEG: multipliziere kleine Werte vor Kodierung mit Faktor, bei Dekodierung dividiere Faktor → Quantisierungsrauschen wird um den Faktor gedämpft, verbessert SNR für leise Signale
- **MPEG-Datenstrom**: Header (beschreibt Strom), Redundancy Code (CRC, erkennt Datenstromfehler), Bitzuweisung (Wortlänge der folgenden Subbandwerte), Skalierfaktor (für folgende Subbandwerte, 6-Bit), Subbandwerte, Hilfsdaten
- **Layer-3-Encoding-Algorithmus**: **basic building blocks**: filterbanks (polyphase filterbank, additional modified DCT), perceptual model (determines quality of encoder, may have own filter bank, output=values for masking threshold or allowed noise), quantization and coding (2 loops: rate loop and noise control loop, code larger values not so accurate, code quantized values with Huffman)
- **Quality**: artifacts: loss of bandwidth, pre-echoes, roughness/double speek
  - measureable only with listening tests, listener must distinguish original from encoded version, may give points on CCIR impairment scale (1.0-5.0)
  - tests must be subjective, else perceptual coding is for Hugo

Kapitel 3: Multimedia Environments

Videodiscs

- **Laser Vision** (LV) = most common format: read only, technology is forerunner of CD, laser burns pits in master CD, master CD is stamped and molded to other CDs, read with laser and immediately made analog
    - **Formats:** CAV (constant angular velocity, all frames start at same angle → flexible playback, freeze…), CLV (constant linear velocity, twice the capacity)

CD

- made by Sony & Philips, true digital medium, data is encoded to be robust,
- **CD Family**: CD-DA (Digital Audio), CD+G, CD+MIDI (graphics & MIDI stored in unused bits), CD-ROM, Photo-CD (mehrere Arten), CD-R, CD-WO (write once), CD-ROM XA (Extended architecture, open form of CD-i), CD-i (interactive), Video CD, CD-V (Video, hybrid digital audio/analog video)
    - **CD-Rom**: im Gegensatz zu CD-DA: additional layer for error detection &correction, divided into data blocks (98 frames): 2 block formats
    - **CD-i:** first MM-platform early 90s, complete system needs HW & SW interfaces, **Media types**: audio (CD-DA), image (RGB, DYUV, CLUT, Run-length), video (MPEG-1), text and graphics (interpreted by jeweiliger applic.)

DVI

- **technology**: several components: formats & codecs for audio & video, enstpr. HW & SW needed, integrate the components into host HW & SW platform
- **Media types**: video (real-time, production-level), audio (zB FM (5h stereo), mid-range (20h mono), near AM mono (40h)), image (some formats: RGB, YUV, color mapping, alpha channels, lossy & lossless compression), text & graphics (how decides HW)

Quick Time

- extension of MacOS for time-based data, **media types**: video & images (based on Apples PICT-format), audio (mono or stereo, interleaved), text & graphics (can be included in PICT data, basically ignored)
- **Time:** explicit, values represented by unsigned integers, time scale (units per sek), duration (max time value), timebase (playback-rate), time coordinate system (TCS, scale and duration)
- **conceptional level**: data entity (actual storage for data), media entity (sequence measured in media time, references storage regions), track entity (reordering of media entity, time in TCS), movie entity (group of track entities); physical organization: atom (basic storage unit, consists of size in bytes, four Chars code (identifies type), content section)

Kapitel 4: Multimodal Info Retrieval

Grundlagen

IR= computergestützte, inhaltsorientierte, unscharfe Suche in unstrukturierten Datenmengen
Multimodal: auf vielfache Weise

- **Grundprinzip**: extrahiere inhaltsbasierte Merkmale aus Medienobjekt, fasse sie in Mermaksvektor zusammen, speichere diesen, mach das für alle; dann kommt eine Anfrage: Was ist ähnlich wie das hier? Vergleiche Merkmalsvektoren, Ähnlichkeitssuche, Ranking der Ergebnisse, evtl modifiziere Anfrage

Retrieval Modelle:

Legt fest: interne Dokumentdarstellung, Anfrageformulierung und – darstellung intern, Vergleichsfunktion zw. Dok. bzw. Dok und Anfrage

- **Boolesches Modell:** Konzept der Mengenlehre, Gewicht 1 oder 0 (hat/hat nicht), kombinierbar mit booleschen Junktoren, Anfrage wird normalisiert (KNF/DNF); Problem: keine Ähnlichkeitssuche, oft zu viele oder keine Ergebnisse, Anwender checken boolesche Junktoren nicht → daher: mach Disjunktionen aus Konjunktionen, führe Relevenzstufen ein, Anfrage kann verfeinert werden, all & any statt and & or
- **Fuzzy-Modell:** Erweiterung vom Boolean, bietet Grad der Zugehörigkeit zw. 0 und 1, daher nicht so scharf, Anfrage wird in DNF übergeführt, Ergebnis wird sortiert
- **Vektorraummodell:** erfasse Doks als Vektoren, bringe bei Retrieval die lineare Algebra ins Spiel, Ähnlichkeit zwischen zwei Vektoren wird berechnet, Anfrage ist auch ein Vektor, Berechne Ähnlichkeit mit Cosinusmaß oder zB Distanzfunktion (Euklidische Distanz)

Relevance Feedback:

Anfragemodifikation notwendig, weil oft: vage Vorstellung über Suchergebnis, schlechte Anfrageformulierung, unbekannte Datenkollektion, keine relevanten Dokumente verfügbar, also reagieren User so: Browsing, manuelle Anfragemodifikation, **Relevance Feedback** (bewertet Ergebnisse)

- **Bewertung:** bewerte <10 Doks nach relevant & keine Bewertung, irrelevant & keine Bewertung, gestufte Relevanz & gestufte Bewertung; Bewertung löst aus: Modifikation von Anfrage, Nutzerprofilen, Dokumentbeschreibung, Suchalgorithmus oder Anfragetermgewichten
    - **Verfahren von Rocchio:** Modifiziert Termgewichte im Vektorraummodell, relevante werden stärker, irrelevante schwächer, verschiebt Anfragepunkte im Vektorraum zu den relevanten hin
- Es gibt: false/correct alarms (gefundene), false/correct dismissals (weggeschmissene)

Content-based image retrieval:

How to retrieve? → retrieval by browsing (via thumbnails), by objective attributes (query with meta and logical attributes, like database queries, perfect match), by spatial constraints (relative spatial relationships of objects, zB adjacency, overlap; can be relaxed (must satisfy many, ranking) and strict (must satisfy all), by semantic attributes, by feature similarity (select features of example images, content-based)

- **Motivation**: inhaltliches Erfassen aufwändig und manchmal unmöglich (zB Bilder im Internet), daher Content-based retrieval: suche Bilder aufgrund von dargestelltem Inhalt, extrahiere Merkmale, Ähnlichkeitsmessung
- **Anwendungsgebiete**: Medienagenturen, Markenzeichen, Produktkataloge (Kunde erinnert sich eher an Aussehen), Kriminalistik, Medizin
- **Anfragearten**: Browsing, Schlagwortsuche, visuelle Beschreibung (query by example: Beispielbild; by sketch, by template (Auswahl Farbe u. Textur)
- **Merkmale**: Primitive (sieht man gleich, Farbe, Textur,…), semantische (Objekte erkennen, System braucht Wissen darüber, Rollen und Szenen)
    - **Farbe**: je nach Farbmodell, Histogramm (hierbei Auflösung, Zoomen egal, Beleuchtung nicht, keine Info über räumliche Verteilung der Farbe, Distanzmaße: (Weighted) Euklidische, Histogram Intersection)
    - **Texturmerkmale**: schwer zu beschreiben, auch Graustufenbilder aussagekräftig, Eigenschaften: Körnigkeit, Periodizität, räumliche Ausrichtung, **Verfahren**: Strukturell nach Lage u. Ausrichtung, Statistisch nach Verteilung der Helligkeitswerte, Markov Random Fields (modelliere

Pixel je nach Nachbarpixel), Fraktale Modelle (für unregelmäßige Texturen), Analyse im Frequenzbereich (für räumliche Verteilung der Helligkeit, mit Fourier, liefert Frequenzbild, tiefe Fr. im Mittelpunkt, Tamura Modell (features Contrast, Körnigkeit, Directionality), RISAR (Rotation invariant Simultaneous Autoregressive model), MRSAR (Multiresolution SAR), world features (describe texture by periodicity, direcionality, randomness); Evaluate texture algorithms with Brodatz Database

- **Formmerkmale**: Formen sind umrissbasiert (Kanten) oder bereichsbasiert (Regions), MPEG-7 Shape Descriptors beschreiben: Region Shape (Pixel in einem Bereich, für einfach und komplexe Formen, mehrere Berechnungsverfahren), Contour Shape (Umrisse), Shape 3D Spectrum (# Konvexitäten von 3D-Obeflächen), Shape 2D/3D (beschreibt 3D-Form durch 2D-Deskriptoren)
- **Weitere Merkmale**: Räumliche Anordnung von Objekten im Bild, Kantenbilder (Suche mit händischer Skizze), Beziehungen der Objekte zueinander (Richtungsbezogen: rechts, links, Entfernung, Winkel; Topologie: gleich wie, innerhalb von)
- **User Interface**: Query by visual templates, drawing a Query (sketching), Query by visual examples (formulate query, choose examples then)

Video Retrieval:

must do video segmentation (find boundaries between shots=camera breaks) via frame difference techniques (pixel comparison, frame histogram comparison), camera operation, techniques on compressed video

- **Frame difference**: pairwise-pixel-comparison, segment boundary (gehört dazu, wenn ein gewisser % an pixel anders sind), Likelihood ratio (schau intensity values of regions an, more tolerance for small object motion than pairwise-pixel), Histogram differences (problem: dissolve sequence, so use another threshold for that
- **Motion continuity**: compute motion vectors via block matching, if differs too much → camera break
- **Twin Comparison Approach** (for **Gradual Transitions**): 1 threshold for camera break detection, 1 for special effects (fade, wipe, slide, Übergänge halt), when one shows difference → mark frame, transition is over when: below Effect Threshold, but difference between this frame and the marked frame is over break Treshold
- **Multi-pass Approach**: reduces processing time, first low resolutions (zB alle 5 frames), then better resolution to be more accurate
- **Motion analysis**: Detect image motion induced by camera movements: **Motion Vector analysis**: optical flow (vector field) detects panning, tilting, zooming, particular motion patterns; also used for object motion (waling, jumping, moving cars), **Video-X-Ray**: spatiotemporal image (mach Quader aus Film, schau Seiten oder Oben an: Schräge Linien oben=Panning, seitlich=Tilting, bei Zooming nähern oder entfernen sie sich)
- **Video Representation**: we must abstract video content, extract salient features, typical style, major subjects. we do that with:
  - **Video Icon construction**: statical icon representing a shot, pseudodepht shows arrows and signs, synthezises visual contents (salient stills [Turmspringen], videospace icon [Frau am PC], videomap)
    - **Tools**: movie icon (micon, 3D-Volume), interactive micon, paper-video (chart-based video browser), video panorama (shows the video space), videoscope (content analyzer), sound browser (detects music)

    - **Key frame extraction**: take frame, when there are significant changes, not very representative
- **Content Indication**: present a part of content for better comprehension, interfaces should offer smooth stage transition between observation modes, sense of overview and partitioning, effective presentation (highly intuitive), attractiveness
- **Segmentation of compressed video**: examine videos in compressed form! SW decompression not so efficient as HW decompression, if no HW decompr. available, use metrics on the compressed representations:
    - **DCT coeffs**, MPEG Video has only I-Frames encoded with DCT, compare DCT coeffs of consecutive frames, get the difference of a block that has changed and the persentage of all blocks, determine break with thresholds
    - **MPEG Motion Vectors**: MPEG stream has 1 set of motion vectors for P-Frames, 2 for B-Frames, within a shot the vectors change relatively continuous, a break disrupts the continuity
    - **Hybrid Approach to Partitioning**: multiple passes and comparisons: first: DCT comparison with high skip factor, second: smaller skip factor, further: motion-based
    - **Evaluation**: motion vector and hybrid best at breaks, motion vector bad at gradual transitions, hybrid is the best

Audio Retrieval:

Audio-Klassifizierung: Sprache (männlich/weiblich), Musik (Arten), Umgebungsgeräusche (zB Tierlaute)

- **Sprache**: geringe Bandbreite, 100–7.000 Hz, Zentroid niedriger als bei Musik, häufige Pausen, viel Stille, Silben mit kurzen Konsonanten (hier hoher Nulldurchlauf) und langen Vokalen
- **Musik**: hohe Bandbreite, 16–20.000 Hz, Zentroid höher, wenig Stille (Außer Solos und A-Capella), Nulldurchlauf variiert nicht so stark, regular beat
- **Klassifizierung** also: schrittweise, hoher Zentroid? → Musik, dann wenig Stille? → Musik, dann niedrige ZC-Variabilität (Nulldurchlauf) → Solo-Musik, sonst → Sprache; Reihenfolge ist wichtig, oft reicht ein Merkmal für korrekte Klassifizierung

Distanz und Ähnlichkeit

Distanzfunktionen

- Vergleichen Merkmale zweier Meidenobjekte, Invarianz drückt aus, was nicht verglichen werden soll; haben Eigenschaften: Selbstidentität, Positivität, Symmetrie, Dreiecksungleichung; alles bzgl einer Metrik
- zB: (**gewichtete**) **Minkowski-Distanzfkt** (setzt bei euklidischer ein “m” ein), quadratische Distanz (erweitert gewichtete euklidische durch A → Einheitsmatrix, Diagonalmatrix, orthonormale oder symmetrische Matrix), Mahalanobis Distanzfkt (quadratische, basiert auf Kovarianzmatrix), Quadratische Pseudodistanz (Abstand 0 auch für nichtidente Punkte), Bottleneck-Distanz (für Mengen gleicher Kardinalität, sucht Minimum der maximalen Distanzen aller Bijektionen)

Ähnlichkeitsmaße

Objekte sind ähnlich, wenn sie beim Menschen ähnliche Reize auslösen, nicht allgemein definiert, es gibt Ähnlichkeitsmodelle in Mathe, Statistik, BV und Mustererkennung, **Ähnlichkeitsmaß** gibt einem Objektpaar eine reelle Zahl von 0–1

- **Probleme mit Distanzfkt**: oft verwendet, aber Psychologie anders, nicht so restriktiv, Selbstidentität/Positivität/Symmetrie/Dreiechsungleichung subjektiv anders wahrgenommen

- **Ähnlichkeitsabstand:** muss haben: Dominanz, Konsitenz, Transitivität (diese Eigenschaften sind allgemeiner als Distanzeigenschaften)
- **Ähnlichkeitsmaße**: Weltwissen wichtig, man nimmt Inhalt in 3 Ebenen wahr: syntaktisch (ohne Bedeutung d. Objekte), semantisch (ähnlichkeitsvergleichend) und pragmatisch (Interpretation, Thematik) → **pre-attentive** Wahrnehmung ist in den ersetn 250 ms, ohne Interpretation/Weltwissen
    - **Modelle**: Kosinusmaß (Vektorraummodell, Skalarprodukt von Vektoren, Winkel für Ähnlichkeit?)
    - **Aggregation von Werten**: Aggregatfunktion muss haben: Ähnlichkeitswerte, Monotonie, strikte Monotonie, Stetigkeit, Idempotenz (mit sich selbst verknüpft ergibt es selbst), Unabhängigkeit von der Reihenfolge
    - **Umwandlungsfunktionen von Distanzen zu Ähnlichkeitswerten**: muss haben: Granzbedingung max. Ähnlichkeit=1, min.=0, streng monoton fallend, stetig; können parametrisierbar und modifizierbar sein

Kapitel 5: Content Description

- **Motivation of MPEG-7**: more and more digital audiovisual info, new ways to produce, offer, filter, search & manage it, better quality & access speed, info value depends on how easy you find & get it, user need something for accurate access, content must be identified and managed
- **Solution: MPEG-7**: offers description tools for quality access to content, active people are broadcasters, content creators/managers, publisher, intellectual property rights manager, telecommunications service provider, …
- **Main Goals**: describe multimedia content, flexible data management, globalization and interoperability of data resources; **standardizes**: Description Schemes and Descriptors, Description Definition Language DDL (to specify Description Schemes), scheme to code Description
- **Terminology**:
    - **Feature**: characteristic of the data, shows something to somebody, needs a good feature representation (Descriptor) and its instantiation (Descriptor Value), zB color, frequency of speech segment, music genre,…
    - **Descriptor**: represents a feature, has a syntax & semantic, zB Color: string; several Descpritors may represent a single feature (zB enumerated lists)
    - **Descriptor Value**: instantiaton of Descriptor for a given data set, Descriptor Values are combined via a Description scheme to form a Description
    - **Description Scheme**: specifies structure and semantics of the relationships between its components (Descriptors of Description Schemes)
    - **Description**: has structure of DS and set of Description Values (instantiations), describes data, contains a fully or partial instantiated DS
    - **Coded description**: encoded for compression, random access, …
    - **Description Definition Language DDL:** allows to create DS and Descriptors and to modify/extend existing DS**. DDL Requirements:**
        - **Compositional capabilities**: can compose new DS and Descriptors, DS may be composed from multiple other DS
        - **Transformational capabilities**: allows reuse, extend, inherit existing DS and Descriptors
        - **Unique identification**: provides mechanisms to uniquely identify DS and Descriptors to let them be unambiguous
        - **Data types**: provides a set of primitive data types (text, integer, version) to describe composite data types

        - **Relationships within a DS and between DSs**: express semantics of these relations (spatial, temporal, structural, conceptual)
        - **Relationship between Description and data**: supplies a rich model for links and/or references between Descriptions and their data
- **MPEG-7 Systems**: define the terminal architecture and the interfaces
- **Terminal architecture**: Terminal: entity that uses coded representation of the content, stand-alone application or part of application system: application, compression layer, delivery layer, transmission/storage medium
    - **transmission/storage medium**: lower layer of delivery infrastructure, delivers multiplexed Streams to the delivery layer
    - **delivery layer**: allows synchronization, framing and multiplexing of MPEG-7 content, that may be delivered alone or with the content it describes, provides elementary streams (consisting of Access Units) for Compression layer
        - **Access Unit**: smallest data entity with timing info
    - **compression layer**: parses (zB with BiM parser) the flow of Access Units, reconstructs content description
- **Motion descriptos**: camera: boom up/down, track right/left, Dolly forward/backward, tilt up/down, pan right/left, roll; **motion trajectory of an object**: descriptor is list of keypoints and set of optional interpolating functions (=path, 2D or 3D), **parametric motion** (used for motion estimation), **activity descriptor** (erkennt "intensity of action")
- **Localization descriptors**: Region locator (localizes regions by specifying them with brief representation of a box or polygon), Spatio temporal locator (describes spatio-temporal regions like moving object regions
- **Multimedia Description Scheme tools**: Basic elements, schema tools, content description/Management/Organisation, Navigation & Access, User Interaction
    - **Content Management tools**: allow description of content's life cycle (creation---consumption), content can have different modalities (zB audio & audiovisual)
    - **Content description**: describes structural aspects, core is: **Segment DS** (describes physical & logical aspects, can form segment trees), (you can use a Graph DS for segment relationships), a segment represents a section of content
        - **Segment DS:** abstract class, 5 subclasses: Audio-Visual, Audio, Still Region, Moving Region & Video Segment DS, **segments** can be composed, then decomposed into sub-segments, segments can described by: Creation info, usage info, media info, textual annotation, specific features (Time, shape, color,…)
- **Navigation and access:** Summaries (for efficient browsing, navigation; Hierarchical (tree, gets more detailed) & sequential summaries (allows fast navigation and access)), Views and Partitions (multiresolution views, progressive access), Variation of Content (specifies relations between av-material, Variation DS specifies the variations of content, zB summaries, low resolution version, Variation fidelity value: quality of the variation compared to original)
- **Collection structure DS**: groups content, segments and event into clusters and specifies what they have in common.
- **User Preference DS**: MPEG-7 content descriptions can be matched to user preferences for personalized, efficient access, presentation & consumption, allows: specification of preferences for different types of content and browsing modes, weighting of preference importance, specify privacy characteristics

Kapitel 6: MPEG-4

- **Motivation**: blurring border between communications, interactivity, broadcasting → so standardize algorithms for audiovisual coding, allow interactivity, high compression, scalability; support natural and synthetic audio video → MPEG-4: defines a MM system for interoperable communication of complex scenes with (synthetic) audio, video, graphics, combines features of other MPEG-standards, should satisfy needs of authors (more reusability & flexibility of content), network service providers (transparent info in appropriate signaling messages), end users (more interaction with content)
- **Principles**: compose a/v objects together according to a scene description: allows interaction with elements in the scene, coding scheme can differ for each object, easy reuse of content;
    - **a/v content can be**: audio(stereo/mono) or video, natural or synthetic, 2D or 3D, streamed or downloaded;
    - the **scene description provides**: spatial/temporal relationship between the objects, behavior and interactivity of objects and scenes, protocols to modify and animate the scene in time. All this info is compressed.
    - **MPEG-4 provides**: coding (representing media objects), composition (of the objects), Multiplex (and synchronize for network transport), interaction (at the receiver's end)
- **Standard Structure**: Systems, visual, audio, conformance testing, reference SW, Delivery Multimedia Integration Framework (DMIF)
- **MPEG-4 Systems (subgroup)**: defines the framework for integrating the natural & synthetic components of scenes, integrates elementary decoders, specifies composition & multiplex
    - **Composition**: describes composition, info consists of the representation of the hierarchical structure of the system, elementary composition inspired by Virtual Reality Modeling Language, deals with 2D-only, does interfacing with streaming media and synchronization, authors can generate this description in text format, description is binary encoded (Binary Format for Scene Description BIFS), MM scenes are conceived as hierarchical structure in a scene graph: each leaf is a media object
        - **Spatial composition**: composition stream provides info required by the terminal to set up the scene structure and map elementary streams to their media objects, each leaf has an own coordinate system
        - **Temporal composition**: composition stream (BIFS) has its own time base, elementary streams get time stamps, specify the time when the access unit should be ready at the decoder input (Decoding Time Stamp DTS), and when the composition unit should be ready at the compositor input (Composition Time Stamp CTS), time stamps on the composition stream specify when the access units for composition must be ready at the input of the composition information decoder; fields in the scene description also carry a time value (duration in time or instant in time)
    - **Multiplex**: turns elementary streams into packages, adds headers with timing info & synchronization data, 3 layer
        - **synchronization layer:** adds MPEG-4 specific info for timing and synchronization, maintains correct time base for the elementary encoders, header contains: Sequence number, instantaneous bit rate, object clock reference, decoding time stamp, composition time stamp

    - **Flexible multiplex layer**: multiplexes streams with very different characteristics, flexible for wide ranges of bit rates, groups together some low-bit-rate streams, with conventional scenes like audio plus video, this layer can be skipped.
    - **Transport multiplex layer**: adapts the multiplexed stream to the particular network characteristics → interface to different network environments
- **MPEG-4 Video**: supports: Content-based interactivity (access-tools, hybrid natural & synthetic data coding, manipulation & bit stream editing, better temporal random access), Compression (efficient coding, coding of multiple data streams), Universal access (robustness in error-prone (fehleranfällig) environments, content-based scalability)
    - **Codec structure**: syntax allow coding of rectangular and arbitrarily (beliebig) objects, scalable and nonscalable, **Coding**: each frame is segmented into arbitrary shaped regions (video object planes VOP), successive VOPs belong to a video object (VO), shape, motion and texture of the VOPs within a VO is encoded into a separate video object layer (VOL), info to identify the VOLs is also encoded
    - **Interactivity**: important: encoder & decoder can function in different frame rates, interactivity between user and en/decoder: at the coding level (zB coding control), at the decoder (zB change video object position)
- **Media Integration of Text & Graphics**: most common way: Bitmap nodes, Sound2D nodes, text nodes, BIFS 3D nodes for scene graph
- **Face animation**: related to text-to-speech, parameters for face animation (facial animation parameters, FAP, describe atomic movements of face or expressions; visemes define mouth position, expressions mimic human emotions, encoded arithmetic or via DCT) & definition (facial definition parameters, FDP, calibrates default face model of the receiver terminal or transmits new face model geometry and texture; the texture is scalable, many resolutions)
- **MPEG-4-Text-to-Speech**: defines binary representation of Text-to-Speech stream and interfaces, contains infos about the synthetic voice apart from text: gender, age, speech rate, language code, prosody (Satzrhythmus), lip shape; allows fast forwarding, pause, play, rewind; handed to the face animation engine → speech driven face animation, low bandwidth
- **Allgemeine Konzepte**: 4 Modi für B-Frame-Kodierung (3 Modi von MPEG-2: Vorwärts Modus, Rückwärts, Interpolationsmodus: übertrage beide Vektoren, interpoliere; Diskreter Modus: zusätzlich bei MPEG-4, nimmt auch Rück/Vorwärts-Werte, leite sie aber nur von einem Vektor ab, skaliere linear und kodiere nur einen Delta-Vektor (Fehlerkorrektur)), Adaptive Quantisierung (jeder Makroblock kriegt eigene Quantisierung, zB bei Xvid: quantisiere sehr helle/dunkle Blöcke stärker, man sieht dort Artefakte nicht so gut → psychovisuelle Aspekte!), Bewegungsvektoren mit bis zu Viertel-Pixel Auflösung (Bewegungskompensation: kodiert wird Bewegungsvektor und Fehler-Bild, je genauer Vektor, desto kleiner Fehler, daher virtuelle Viertel-Pixel durch Interpolation berechnen → realitätsnahe Beschreibung von Bewegungen), Globaler Bewegungsausgleich (4 globale Bewegungsvektoren pro VOP, gute Kompression bei Kamerabewegung, zusätzlich für jeden Makroblock eiN Bewegungsvektor, Encoder entscheidet, welcher Makroblock welchen Bewegungsvektor kodiert kriegt), Wahl der Quantisierungsmethode (darf User selbst wählen), Variable Blockgröße