Innovationsforum „CineArchiv digital“ Digitalisieren, Erschließen und Nutzen audiovisueller Materialien – Metadatenerfassung – First Draft und Diskussionsgrundlage

B. Baumann, Ch. Meinel, H. Sack, Ch. Willems
Hasso-Plattner-Institut für IT Systems Engineering
Universität Potsdam

9. Dezember 2008

Inhaltsverzeichnis

0 Präambel

Arbeitsziel der Gruppe **Metadatenerfassung** im Projekt **Digitalisieren, Erschließen und Nutzen audiovisueller Materialien** ist die Feststellung und Zusammenfassung des aktuellen Forschungsstandes im Bereich der Analyse und Annotation audiovisueller Daten mit Metadaten.

Das vorliegende Dokument soll dabei einen ersten Einstieg in die Thematik darstellen und gibt daher lediglich einen vorläufigen Stand wieder, der als Arbeitsgrundlage der Projektgruppe Metadatenerfassung dienen soll. Die Struktur des Dokuments orientiert sich an Übersichtsarbeit von Lienhart et al. [79] und wurde bzgl. Vollständigkeit, Relevanz und Aktualität erweitert und überarbeitet.

Im weiteren Verlauf sollen alle Gruppenmitglieder an der Fortschreibung dieses Dokuments beteiligt werden und so zur Vollständigkeit, Aktualität und Korrektheit dieses „State-of-the-Art“ Dokuments zur Erschließung audiovisueller Materialien beitragen.

1 Einleitung

Der Bestand an audiovisuellen Daten wächst täglich. Filme und Videobeiträge werden Tag für Tag produziert, gesendet, archiviert und mittlerweile bereits auch oft direkt über das World Wide Web (WWW) weltweit für jedermann verfügbar gemacht. Doch die Produktion audiovisueller Inhalte reicht weit in die Zeit vor den Siegeszug der elektronischen Kommunikationsmedien zurück. Seit mehr als 100 Jahren werden audiovisuelle, inhaltsbasierte Objekte produziert und die heute bereits im WWW verfügbaren Objekte stellen lediglich die Spitze eines Eisbergs von Materialien dar, dessen Fundament noch immer in analoger Form in unzähligen Archiven und Bibliotheken ruht. Die Öffnung dieser audiovisuellen Archive und das Verfügbarmachen ihrer Inhalte erfordert als erstes die Digitalisierung einer Vielzahl unterschiedlichster analoger Medien in zahlreichen Qualitätsabstufungen, z.T. mit Fehlern und Defekten.

Um aber in einem nächsten Schritt gezielt auf die digitalisierten, audiovisuellen informationstragenden Objekte zuzugreifen, müssen Inhalt und Struktur mit Hilfe von **Metadaten** beschrieben werden. Diese Metadaten liegen in strukturierter oder unstrukturierter (textueller) Form vor und dienen als Grundlage für einen Information Retrieval Prozess, an dessen Ausgang ein wahlfreier, punktgenauer Zugriff auf die gesuchten audiovisuellen Daten ermöglicht wird.

Prinzipiell lassen sich bei der Auszeichnung audiovisueller Daten mit beschreibenden Metadaten folgende drei Phasen identifizieren:

- **Datenanalyse**:
  Die Datenanalyse dient der Gewinnung von Metadaten, die Inhalt, Struktur und Eigenschaften der zu untersuchenden audiovisuellen Rohdaten beschreiben.
- **Metadaten-Annotation**:
  Die aus der Datenanalyse gewonnenen Daten müssen in einer formalisierten und interoperablen Datenstruktur abgelegt werden.
- **Metadaten-Nutzung (Processing)**:
  Die aus der Datenanalyse gewonnenen strukturierten Metadaten werden zum Zweck der inhaltlichen Kategorisierung, Durchsuchung, Visualisierung und Navigation des audiovisuellen Rohmaterials eingesetzt.

2 Datenanalyse

Zur Datenanalyse müssen analoge audiovisuelle Materialien zuerst digitalisiert werden. Ausgehend von den digitalisierten audiovisuellen Rohdaten, lassen sich unterschiedliche Analyseverfahren in Abhängigkeit des Analysegegenstandes, der jeweiligen Analyseaufgabenstellung und der Granularität des Analysegegenstandes identifizieren und anwenden. Ziel ist es dabei, inhaltliche und strukturelle Informationen aus den audiovisuellen Rohdaten zu extrahieren.

Zu den Analysegegenständen zählen alle verfügbaren medialen Informationen, die in den digitalisierten audiovisuellen Rohdaten identifiziert werden können, wie z.B. Audioinformation, Videosequenzen, Einzelbilder und eventuell vorhandene textuelle Daten, die die audiovisuellen Rohdaten ergänzen. Diese werden im Folgenden einer detaillierten Betrachtung (2.1 - 2.3) unterzogen.

2.1 Audio

Bei der Analyse von Audiodaten gibt es Hilfsmittel aus verschiedenen Bereichen. Zum einen fasst man die Methoden zur direkten Signalverarbeitung als Extraktion der Low-Level Audioeigenschaften zusammen (s. Abschnitt 2.1.1). Zum anderen gibt es Methoden zur Extraktion komplexerer Informationen (High-Level) aus den Bereichen

- Segmentierung und Klassifikation (Abschnitt 2.1.2),
- Content-basiertes Retrieval (Abschnitt 2.1.3),
- Clustering (Abschnitt 2.1.4)
- und dem großen Gebiet der automatischen Spracherkennung (Abschnitt 2.1.5).

Diese Methoden lassen sich jeweils unterscheiden in parametrisierte Methoden, die auf der Annahme basieren, dass das untersuchte Audiosignal auf einem mathematischen bzw. statistischen Modell beruht, und nicht-parametrisierten Methoden, die keine Annahmen über das Ursprungssignal voraussetzen.

Die gewonnen Informationen aus der Audiodatenanalyse können teilweise als Eingabeparameter für Methoden zur Videoanalyse genutzt werden und umgekehrt. Diese Abhängigkeiten werden in Abschnitt 2.3 betrachtet.

2.1.1 Signalverarbeitung und -analyse

Bei den Methoden der Signalverarbeitung werden aus dem rohen Datenstrom Merkmale des Audiosignals bestimmt. Diese Merkmale sind entweder zeitlich oder an der

Frequenz bemessen. Ein wichtiger Parameter für alle Methoden ist die Größe des jeweiligen Analysefensters – ein Zeit- oder Frequenzintervall, für das ein Merkmal bestimmt wird. Sinnvolle Werte für ein zeitliche Segmente können je nach Methode und Weiterverwendung der extrahierten Merkmale zwischen 10 Millisekunden (kleines Subsegment) und einer Sekunde (Makrosegment) liegen [36].

Die ermittelten signalnahen Merkmale dienen einzeln oder in Kombination entweder als Deskriptoren (Low-Level Descriptors, LLD) für die Ermittlung komplexerer Merkmale (High-Level) oder direkt für die Bestimmung von Ähnlichkeitsmaßen genutzt werden. Low-Level Deskriptoren werden auch direkt für Klassifizierung von Signalverläufen, also der Unterscheidung zwischen Musik, Sprache und Umgebungsgeräusche oder der Einordnung nach Geschlecht von Sprechern bzw. Musik-Genres verwendet.

Weit verbreitete Merkmale von Audiosignalen umfassen:

Short Time Energy (STE) Die Messung der Frequenzamplitude (STE, auch als Loudness oder Volume bekannt) innerhalb kurzer Zeitintervalle [213] kann für die Erkennung von Pausen (Stille), die Unterscheidung zwischen Musik und Sprache oder Einteilung in stimmhafte und stimmlose Sprachanteile genutzt werden.

Low Short Time Energy Ratio (LSTER) Abgewandelte STE nach [108], betrachtet insbesondere Zeitintervalle mit unterdurchschnittlicher Frequenzausprägung.

Zero Crossing Rate (ZCR) Die Messung der Vorzeichenwechsel („zero crossing“) innerhalb des Analysefensters ist nützlich bei der Unterscheidung von Musik und Sprache oder bzw. stimmloser und stimmhafter Sprache. Eine Abwandlung dieses Merkmals ist die High Zero Crossing Rate Ratio (HZCRR).

Spectral Flux Die Messung der durchschnittliche genutzten Frequenzbereiche benachbarter Zeitsegmente dient zur Unterscheidung Sprache/Nicht-Sprache bzw. Musik/Umgebungsgeräusche.

Band Periodicity Diese Variante der Subband Korrelationsanalyse untersucht statistische Abhängigkeiten zwischen verschiedenen Unterfrequenzbereichen des vorliegenden Spektrums und erlaubt die Unterscheidung zwischen Musik und Umgebungsgeräuschen.

Median Frequency, Centroid Frequency Dieses Frequenzmerkmal bestimmt einen Frequenzschwerpunkt im vorliegenden Frequenzband und dient zur Erkennung von Signalbandbreiten, Unterscheidung von männlichen und weiblichen Sprechern oder zur Klassifikation von Musik nach Genre [45].

Mel Frequency Cepstral Coefficients (MFCC) In einer mehrstufigen Transformation wird aus den Frequenzwerten das Cepstrum (logarithmiertes Fourierspektrum) berechnet. Mit den Cepstralkoeffizienten kann anhand der Mel-Skala die Tonheit bestimmt werden. Damit können insbesondere in der Spracherkennung die Formanten stimmhafter Phoneme bestimmt werden [118].

Fundamental Frequency, Pitch Detection Zur Tonhöhenerkennung (Pitch Detection) existiert eine Reihe unterschiedlicher Ansätze beschrieben z.B. in [128] (und

dort aufgeführten Referenzen), [212] oder [131]. Die gewonnenen Deskriptoren können zur Harmonieanalyse für Musik oder die Unterscheidung stimmhafter und stimmloser Sprachanteile genutzt werden.

Aus den beschriebenen Methoden zur Bestimmung der Low-Level-Deskriptoren lassen sich weitere kombinierte LLDs bestimmen. Der MPEG-7-Standard beschreibt eine Reihe von LLDs, die die genannten Methoden nutzen (vgl. [117]) und für die es auch entsprechend performante Implementierungen gibt, so z.B.:

MPEG-7 LLD	Abgeleitet von
AudioWaveform	STE, ZCR, LPC, Pitch, PLP
AudioPower	STE, ZCR, LPC, Pitch, PLP
AudioSpectrumCentroid	STE, ZCR, LPC, Pitch, PLP
TemporalCentroid	SF
...	...

Umfang und genaue Bezeichnung aller MPEG-7 Audio Deskriptoren kann den Ausführungen des Standards [75] zum MPEG-7 Audio Framework entnommen werden.

Weiterhin gibt die Bibliographiestudie in [36] einen Überblick darüber, welche Low Level Deskriptoren in welchen komplexeren Analysemethoden zum Einsatz kommen.

2.1.2 Segmentierung und Klassifikation

Als Segmentierung bezeichnen wir die Unterteilung eines zusammenhängenden, zeitbezogenen Rohdatenabschnitts in kleinere, inhaltlich möglichst kohärente Abschnitte. Klassifikation meint die Einordnung der Audiodaten bzw. von Segmenten der Rohdaten in ein Klassenschema, also z.B. die Unterscheidung zwischen Sprache, Musik und Geräuschen oder die Bestimmung der Genres von Musikstücken.
Generell kann das Problem der Segmentierung bzw. Klassifikation von Audiodaten auf drei unterschiedliche Arten angegangen werden.

Regelbasierte Methoden Einfache Algorithmen zur Segmentierung berücksichtigen lediglich die Werte aus den extrahierten Low-Level Deskriptoren ohne zum eigentlichen Inhalt Kenntnisse zu haben. Zu Segmenten zusammengefasst werden (zeitliche) Bereiche, in denen sich das beobachtete Merkmal in einem "stabilen Zustand" befindet, also einen Grenzwert nicht überschreitet oder ein Schwellintervall nicht verlässt. Diese Algorithmen sind meist sehr performant und erlauben ihre Anwendung in Echtzeit. Dafür hängt die Qualität der Ergebnisse stark vom gewählten Schwellwert ab. Eine Verbesserung kann in manchen Anwendungsfällen die Einführung adaptiver Schwellwerte bringen.

Metrikbasierte Methoden Segmentierung und Klassifikation finden jeweils auf der Basis einer speziellen Metrik statt, die die Distanz einzelner Segmente zueinander bestimmen. Mögliche Metriken umfassen:

- die Kullback-Leibler-Distanz [83],
- das Ähnlichkeitsmaß (likelihood ratio) nach [58],
- den Entropieverlust [83] oder
- das Bayessche Informationskriterium [155].

Metrikbasierte Methoden treffen ebenfalls keine Annahmen über den Inhalt der betrachteten Audiodaten.

Modellbasierte Methoden Komplexere Methoden zur Klassifikation von Audiodaten basieren jeweils auf einem Modell aus akkustischen Klassen (z.B. Stille, Musik, Sprache). Diese können entweder einen einzelnen Inhaltstyp beschreiben oder als hierarchische Struktur wiederum Subklassen enthalten, wie weibliche Sprecherin/männlicher Sprecher für die Klasse der Sprachdaten. Diese Klassen können mittels verschiedener mathematischer Modelle beschrieben werden, so z.B. Gaussian Mixture Models (GMM) oder Hidden Markov Models (HMM).
Bekannte Implementierungen für Klassifizierer können auf unterschiedlichen Ansätzen basieren:

- k-NN Algorithmen (Symbole/Deskriptoren werden mit statischen, klassifizierten Trainingsdaten verglichen und nach der Nähe zu den k nächsten Nachbarn eingeordnet [158]),
- Maximum Likelihood (ML) Methode (ein modellbasiertes statistisches Schätzverfahren [82]),
- Viterbi Algorihmus (Bestimmung der wahrscheinlichsten Zustandsequenz von versteckten Zuständen zu einem HMM und einer beobachteten Sequenz von Symbolen/Deskriptoren [193]; Hilfsmittel zur Erkennung von Mustern) oder
- mehrstufige Verfahren (z.B. nach [108]: Schwellwert-regelbasierter Algorithmus nach Pre-Klassifizierer)

Weitere Quellen zu Segmentierung und Klassifizierung: [171], [100], [200], [85] oder [142]

2.1.3 Pattern- und Content-basiertes Retrieval

Pattern Retrieval Methoden nutzen die Segmentierungs- und Klassifikationsalgorithmen nach dem „Query-by-Example“-Paradigma. GMMs oder HMMs werden mit relevanten Audiomustern trainiert, beobachtete Audiodeskriptoren werden mit geeigneten Methoden nach Ähnlichkeit gematcht.

Methoden des Pattern Retrieval kommen im Bereich des Content-basierten Retrieval, z.B. beim sogenannten „Sound Spotting“ [166] zum Einsatz: auf Basis einer Menge von Audiomustern werden die Abfrage-Audiodokumente auf Stellen mit ähnlich wahrnehmbaren Mustern durchsucht. So sollen Bereiche unterschiedlicher Inhaltstypen, z.B.

Jingles, Nachrichten, Wetter- und Verkehrsberichte oder Werbung klassifiziert und segmentiert werden können.

Auch andere Arbeiten schlagen ein Retrieval nach der „akkustischen Ähnlichkeit“ von Audiomustern vor, so z.B. [51]. Dieses Prinzip wird neben Retrieval-Systemen arbeiten auch für die automatische Klassifikation von Audiodokumenten nach Musik-Genres (vgl. [125], [183], [34]) genutzt. Diese Ansätze kommen auch in aktuellen mobilen Anwendungen wie Shazam[1] [196] oder Midomi[2] („Query by Humming“, [56]).

Weitere Arbeiten im Bereich Content-based Retrieval finden sich bei [53], [115], [106], [100], [200] und [136].

2.1.4 Clustering

Eine Reihe von Methoden dient zur Gruppierung von klassifizierten Segmenten eines betrachteten Audiostreams. Generell basieren Cluster-Kriterien auf Ähnlichkeitsmaßen, abhängig von der spezifisch gewählten Klassifizierungs- bzw. Segmentierungsmethode können jedoch spezielle Methoden genutzt werden: bei Audiosignalen, die als Musik klassifiziert wurden, können nach dem Auftreten bestimmter Instrumente als Cluster-Kriterium dienen während bei Sprachsignalen Methoden aus dem „Speaker Modelling“ (siehe Abschnitt 2.1.5) zum Einsatz kommen können.

In der Regel basieren Clustering Methoden auf einem von zwei Ansätzen:

- Bottom-Up Ansatz: Segmente werden auf Basis eines Distanzmaßes zusammengefasst (z.B. k-NN oder ML).
- (hierarchischer) Top-Down Ansatz: Segmente werden Knoten in einer vorgegeben Hierarchie zugewiesen, entweder auf Basis eines Distanzmaßes (nach [77]) oder der adaptiven Hierarchiekonstruktion nach [214].

2.1.5 Automatische Spracherkennung

Die ausgefeiltesten Methoden der Audioanalyse kommen im Bereich der automatischen Spracherkennung (Automatic Speech Recognition, ASR) zum Einsatz. Hier sind Erkenntnisse aus unterschiedlichen wissenschaftlichen Bereichen wie der Akkustik, der Phonetik, der Linguistik oder der Computerlinguistik (Verarbeitung natürlicher Sprache) gefordert.

Auch die Methoden der Spracherkennung können wieder in unterschiedliche Bereiche aufgeteilt werden, wobei hier neben der Erkennung und Verarbeitung unterschiedlicher Sprecher vor allem die Transkription – also die Umwandlung gesprochener Worte in korrekte Schriftsprache – eine wichtige Rolle spielt.

Es existiert eine Vielzahl an Systemen zur Spracherkennung auf dem Markt, darunter auch freie Systeme, wie das HTK[3] (Hidden Markov Toolkit der Universität Cambridge), VoxForge[4] oder das Sphinx-Projekt[5] [3] der Carnegie Mellon University.

[1] `http://www.shazam.com`
[2] `http://www.midomi.com`
[3] `http://htk.eng.cam.ac.uk/`
[4] `http://www.voxforge.org/`
[5] `http://cmusphinx.org/`

Kommerzielle Systeme umfassen beispielsweise das CSLU[6] Toolkit (Center for Spoken Language Understanding), Dragon Naturally Speaking[7] (Engine bildet auch Grundlage für MacSpeech[8]), Philips SpeechMagic[9] (Markführer im medizinischen Bereich) oder das bekannte IBM ViaVoice (an Nuance lizensiert, Entwicklung wird vermutlich zugunsten von Dragon eingestellt).

Eine gute allgemeine Einführung in die Spracherkennung bietet [153].

Segmentierung und Klassifikation Vor der eigentlichen Spracherkennung unterscheidet ein ASR-System eingehende Audiosignale nach Sprache und Nicht-Sprache. Diese Klassifikation erlaubt es, Segmente ohne Sprachanteile fallenzulassen und so Berechnungsaufwand einzusparen. Zur Unterscheidung zwischen Sprache und Musik bzw. Geräuschen kann eine Reihe von zu extrahierenden Low-Level Deskriptoren herangezogen werden, so zum Beispiel Short Time Energy, Zero Crossing Rate oder Spectral Flux (siehe Abschnitt 2.1.1).

Dieses Klassifikation in Sprache und Nicht-Sprache kann sehr performant realisiert werden (deutlich schneller als Echtzeit, [79]).

Segmentierung und Clustering nach Sprecher Nach der groben Klassifikation werden identifizierte Sprachsegmente anhand der Median Frequency in männliche und weibliche Sprecher aufgeteilt. Die Mel Frequency Cepstral Koeffizienten (MFCC) können mittels Gaussian Mixture Models stochastisch und mit Hidden Markov Models zeitlich modelliert werden und erlauben eine Unterscheidung verschiedener Sprecher [79]. Bei [108] dienen die Linear Spectral Pairs (LSP) als Maß für die genaue Bestimmung der Grenzen zwischen zwei Sprechern. In [203] wird ein schneller Algorithmus auf Basis der Pitch-Deskriptoren vorgeschlagen.

Auch die Sprechersegmentierung kann deutlich schneller als in Echtzeit berechnet werden.

Auch die Erkennung eventueller Dialekte gehört zum Bereich der Sprachsegmentierung [48]. Weitere relevante Arbeiten zu Segmentierung und Clustering nach Sprechern finden sich bei [91], [90] oder [109] und den dort aufgeführten Quellen.

Sprecheridentifikation Die Identifikation von Sprechern beruht auf dem Ähnlichkeitsvergleich zwischen den Deskriptoren eines beobachteten Audiosignal und einer Datenbank von Sprechermodellen. Diese Modelle werden zuvor in einer Trainingsphase (wenige Minuten Audiodaten je Sprecher nötig) erstellt.

Generell haben die Systeme zur Sprecheridentifikation einige wichtige Nachteile zu überwinden:

- Die zugrundeliegenden Modelle lassen keine Klassifizierung als „unbekannter Sprecher“ zu. Daher verschlechtert jeder nicht-identifizierte Sprecher die Datenbank mit „Störgeräuschen“.

[6] `http://cslu.cse.ogi.edu`, Software für nicht-kommerziellen Einsatz frei
[7] `http://www.nuance.com/dragonnaturallyspeaking`
[8] `http://www.macspeech.com`
[9] `http://www.philips.com/speechrecognition`

- Die Berechnungskosten für das Retrieval steigen überlinear.
- Die Sprecherdatenbank erfordert manuelle Wartung, um übermäßiges Anwachsen der Modellgröße zu verhindern.

Im Allgemeinen kann die Sprecheridentifikation in Echtzeit ausgeführt werden (mehr als 50 Sprechermodelle gleichzeitig).

Aktuelle Arbeiten in diesem Bereich befassen sich mit der zuverlässigen Sprecheridentifikation in lauten Umgebungen (z.B. Meetings [112], [157]), Sprecheridentifikation in Echtzeit [86] oder semantische Analyse von Konversationen [192].

Sprachtranskription Die Methoden der Sprachtranskription (auch Speech-to-Text-Transcription) dienen dazu, gesprochenen Text aus Audiosignalen möglichst automatisch in geschriebenen Sprache zu überführen.

Die verwendeten Methoden zur Transkribierung lassen sich nach dem Grad der Automatisierung differenzieren.

- **manuelle Methoden:** hier sind insbesondere effiziente und benutzerfreundliche Verfahren gefragt, die eine Transkribierung mit möglichst wenig Arbeitsaufwand (Kosten) gestatten.
- **semi-automatische Methoden:** hier sind stets manuell erstellte Transkripte der Ausgangspunkt für maschinelle Lernverfahren, mit denen weitere, automatisch erstellte Transkripte erstellt werden können.
- **automatische Methoden:** hier kommen trainierte und untrainierte Spracherkennungssysteme zum Einsatz, um automatisch Transkripte zu erzeugen.

Bei der automatischen Transkription werden die vorher ermittelten Sprechersegmente („Turns“) weiter in Subsegmente untergliedert: anhand von Pausen oder anderer Merkmale wird dann in Sätze segmentiert. Aus diesen werden mittels verschiedener Low-Level Features (hauptsächlich MFCC) Deskriptoren ermittelt, die anhand verschiedener Modelle auf entsprechende Symbole gematcht werden. In diesem Schritt kommen in der Regel zwei Modelle zum Einsatz:

- **Lexikalisches Modell:** Sammlung von Wörtern (typische Wörterbuchgröße: 65000 Einträge) samt phonetischer Repräsentation jedes Wortes.
- **Sprachmodell:** wird während der Trainingsphase anhand eines großen Textkorpus entwickelt. Ein wichtige Parameter beim Training ist die Anwendungsdomäne. Im Sprachmodell werden unter anderem Wahrscheinlichkeiten für das gemeinsame auftreten bestimmter Wörter innerhalb der jeweiligen Domäne statistisch modelliert. Auch die Warscheinlichkeiten für Wortkompositionen gehört zum Sprachmodell.

Viele Systeme zur automatischen Spracherkennung arbeiten mit einem adaptiven zweistufigen Mechanismus: die Resultate der nicht-adaptiven ersten Phase werden zur Anpassung des Modells für die adaptive Phase (“Second Pass”) verwendet.

Wichtige Einflussfaktoren für die Qualität der Sprachtranskription umfassen die Beschaffenheit der akkustische Umgebung (laute Hintergrundgeräusche), die Überlappung von Sprechern (bei Meetings oder Diskussionen oder die Sprachqualität (z.B. geplante und spontane Sprache).

Aktuelle Arbeiten zum Thema befassen sich unter anderem mit dem Problem der multilingualen Sprachtranskription [199, 154], der Steigerung von Performance und Erkennungsgenauigkeit (z.B. [169]) oder der Anwendung in spezifischen Domänen, wie Radio- und TV-Nachrichten [54] oder der Transkription von Vorlesungen und Seminaren. Letzteres entsteht im kommerziellen Umfeld aus den Arbeiten bei IBM [71] aber auch im Kontext des Forschungsprojekts tele-TASK[10] (Ziel: semantisch annotierte Vorlesungsaufzeichnungen, [141]).

2.2 Video und Videoeinzelbilder

Bei der Analyse von Videodaten können zwei Hauptzweige identifiziert werden: strukturelle und inhaltliche Analyse. Die strukturelle Analyse zielt auf die Erfassung kohärenter (zusammenhängender) Abschnitte, welche in ihrer Gesamtheit eine Szene ergeben. Videos können dabei strukturell nach zeitlichen und räumlichen Gesichtspunkten segmentiert werden. Ziel ist es in jedem Fall, das Video in logische Abschnitte zu unterteilen. Diese Unterteilung wird dabei mit der Aufnahme von Zeitmarken festgehalten. Eine inhaltliche Analyse geht dagegen im wesentlichen von der Objektidentifikation aus.

Für die Videoanalyse, gibt es ein breites Feld an Konzepten, Technologien und Methoden. [79] nennt u.a. die im folgenden vorgestellten Analyseverfahren für visuelle Inhalte.

2.2.1 Deskriptoren

Deskriptoren können – wie im Bereich der Audioanalyse – generell in Low-Level und High-Level Deskriptoren eingeteilt werden. Dieses Kapitel stellt eine Einführung in die Low-Level Deskriptoren dar. Extrahierte Low-Level Deskriptoren können z.B. als Input für die Extrahierung von Merkmalen für High-Level Deskriptoren (wie z.B. die Segmentierung oder Klassifizierung) genutzt werden. Ein Low-Level Merkmal kommt typischerweise als Eigenschaft entweder in sehr vielen oder gleich allen Pixeln des Bildes vor. Jeder Pixel eines Bildes hat zum Beispiel eine Farbe. Eine daraus kompakt abgeleitete Repräsentation eines Merkmals, kann als Deskriptor Bezeichnung finden und repräsentiert ein bestimmtes Merkmal eines Bildes oder einer Reihe von Bildern (Sequenz).

Arten von Deskriptoren Für die Interoperabilität und Vergleichbarkeit von Videodeskriptoren über Anwendungsgrenzen hinweg, sind standardisierte Deskriptorendefinitionen notwendig. Der MPEG-7 Standard definiert ein Set von Low-Level Deskripto-

[10] Tele-Teaching Anywhere Solution Kit, `http://www.tele-task.de/`

ren. Allerdings besitzen diese keinen normativen Charakter, um eine Weiterentwicklung und Ergänzung der Deskriptorenmenge zu ermöglichen. Folgende Deskriptoren für visuellen Merkmale können unterschieden werden:

- **Farbmerkmale**:
  Hauptausprägung eines Farbmerkmals ist der **Farbraum**. Daher sind Farbräume in der Fachliteratur entsprechend stark vertreten [50] [72]. Farbräume haben verschiedene Eigenschaften, wie zum Beispiel die Zuordnung verschiedener Farbbestandteile oder der Leistungsstärke für Helligkeit und Farbton.
  Klassen von Farbdeskriptoren: die große Zahl möglicher Farbdeskriptoren sind in [11] aufgeführt, weiterhin stellt [113] eine Übersicht der verfügbaren Klassen von Farbdeskriptoren in MPEG-7 auf. Beispiele für diese Klassen sind

  – Histogramm-basierende Methoden oder
  – die Methode der dominierenden Farbe.

  Räumliche Informationen in Farbdeskriptoren: für einige Verarbeitungsaufgaben ist es relevant, Informationen über die räumliche Verteilung von Farben in einem Bild zu haben. Die einfachste Herangehensweis dafür ist die Zerlegung des Bildes in einzelne Blöcke und deren separierte Betrachtung und Bildung eines Deskriptors für jeden Block. Komplexere Ansätze werden z.B. in [170] besprochen.

- **Texturmerkmale**:
  Die Textur beschreibt die Struktur eines Bildes. Periodizität, Grobkörnigkeit, Richtungsabhängigkeit und Kontrast können als Beispiele beschreibender Dimensionen des Texturdeskriptors betrachtet werden. Viele Texturdeskriptoren arbeiten nur mit Graustufenbildern (grey-level images). Die folgenden Aufzählungen beinhalten aber auch Betrachtungen auf Farbtexturen.

  Räumliche Texturmerkmale können auf unterschiedliche Weisen ermittelt und beschrieben werden:

  – Auto-Korrelation – Strukturelle Merkmale eines Bildes, abhängig von der räumlichen Ausdehnung der Graustufen-Primitive.
  – Fraktalanalyse – Mit Hilfe der Messung der Fraktalausdehnung, können Texturen ebenso beschrieben werden [46].
  – Co-Occurences – Beschreibt die Beziehung der räumlichen Nachbarschaft von Grauwerten.
  – High-Level Merkmalsbeschreibung – Aussagekräftigere Merkmale als bei der Co-Occurences, wie beispielsweise Kontrast oder Richtungsabhängigkeit, werden herangezogen.

  Textursignaturen stellen ein weiteres Einteilungskriterium dar. Zu den gebräuchlichsten zählen:

  – Grobkörnigkeit, Richtungsabhängigkeit und Kontrast

  - Wiederholungshäufigkeit, Richtungsabhängigkeit, Aufwand
  - Texture Energy – Unter Anwendung spezieller Funktionen, wird eine Reihe so genannter „Energy Images“ umberechnet (in diesem Fall gefaltet). Jeder Originalpixel wird dabei durch einen Vektor repräsentiert.

- **Form**:
  Formbeschreibung und -abgleich werden für gewöhnlich gemeinsam betrachtet. Für die folgenden Formbeschreibungen mit ihren Analysealgorithmen wurde bereits vorab eine räumliche Bildaufteilung vorgenommen. Einfach gesagt soll die räumliche Bildaufteilung Bildmasken bereitstellen, die jeweils relevante Bildregionen oder Bildobjekte repräsentieren.

  - Global Contour Based Methods (Shape Signatures, Fourier Deskriptoren, Wavelet Deskriptoren, Contour Distributions, Scale Space, Curvature Scale Space, Autoregressive Methoden)
  - Structural Contour Based Methods (Chain Codem, Polygon, Curvature and Orientation, Syntactic Analysis, Shape Invariants)
  - Global Region Based Methods (einfache geometrische Eigenschaften, Geometric Moment, Algebraic Moment Invariants, Orthogional and other Moments, Angular Radial Transfom, Grid Method, Shape Matrix)
  - Structural Region Based Methods (Convex Hull, Medial Axis)

- **Bewegungsmerkmale**:
  Eine Bewegung stellt sich als Veränderung im visuellen Inhalt einer Aufnahme dar. In Bildsequenzen gibt es zwei hauptsächliche Bewegungstypen, die unterschieden werden können: globale Bewegung, bei der jeder Pixel eines Bildes involviert ist und die regionale Bewegung, welche die Bewegung eines Objektes oder einer Region relativ zu einem anderen Objekt oder zum Hintergrund beschreibt. Methoden zur Ermittlung der Bewegungsmerkmale können in Bewegungsabschätzung und -beschreibung eingeteilt werden:

  Bewegungsabschätzung

  - Bewegungsmodelle – Ein Bewegungsmodell legt unter anderem die Anzahl der Parameter für eine Bewegung fest. Einen umfassenden Überblick zur Thematik der Bewegungsmodelle und Ihrer formalen Definition ist in [168] beschrieben.
  - Hierachical Motion Estimation – Die Hierachical Motion Estimation (auch Mehrfachbewegungsabschätzung) bezieht sich auf Methoden zur Bewegungsabschätzung, die sukzessiv auf Instanzen der Originalbilder in unterschiedlichen Auflösungen angewandt werden.
  - Optical Flow Methods – Man unterscheidet zwei Modellansätze: parametrisierte und nicht parametrisierte Modelle. Ziel ist die Definition von Bedingungen für das Bewegungsvektorfeld.

– Blockbasierte Methoden – Blockbasierte Bewegungsabschätzungsmodelle teilen das Bild in eine Anzahl von Blöcken und ermittlen in jedem Block gleichartige Bewegungen.
– Bayes'sche Methoden – Basiert auf einem Abschätzungskriterium nach Bayes [168].

Bewegungsbeschreibung

– Kamerabewegung – Die Kamerabewegung ist bei Objekten von Interesse, die das gesamte Bild füllen oder zur Ermittlung des Blickwinkels.
– Bewegungsbahn – Beschreibt den Bewegungspfad eines Objektes über einen bestimmten Zeitraum.
– Bewegungsaktivität – Manchmal ist nicht der genaue Bewegungspfad von Interesse, sondern lediglich ob eine Bewegung überhaupt stattgefunden hat.

2.2.2 Segmentierung

Die Zerlegung eines Bildes bzw. einer Bildsequenz in Teile, die einheitlich einem bestimmten Merkmal oder einer Gruppe von Merkmalen entsprechen, bezeichnet man als Segmentierung. Einige der zuvor besprochenen Low-Level Deskriptoren, können für eine Segmentierung genutzt werden.

Räumliche Segmentierung

- Thresholding – Die Segmentierungsentscheidung wird auf Grund von einzelnen Pixelinformationen getroffen.
- Edge Based Segmentation – Die Konzentration liegt hierbei auf Abgrenzungs- bzw. Kontureninformationen.
- Active Contour Model – Das Active Contour Model ist ein relativ neuer Ansatz und wird auch als so genanntes SNAKE Konzept bezeichnet [96]. Hauptansatzpunkt hierbei ist die Darstellung der Formbegrenzung als Spline-Kurven und diese Darstellung iterativ anzupassen.
- Textursegmentierung – Eine Segmentierung erfolgt bzgl. der Zuordnung gleichartiger Flächen. Lösungsansätze sind u.a. in [107] angeführt.
- Normelised Cut – Hierbei wird das segmentierte Bild als gewichteter, ungerichteter Graph repräsentiert [159]. Jeder Pixel ist dabei ein Knoten im Graph und eine Kante verbindet jedes Pixel-Paar. Die Gewichtung der Kanten, ist dabei ein Maßstab für die Gleichartigkeit der Pixel. Siehe hierzu auch Graph Partitioning Methods.
- Objekterkennung – Erkennung bestimmter Objekte in einer Szene. Neuere Ansätze basieren auf neuronalen Netzen für die Objekterfassung in Videosequenzen. [160]

Zeitliche Segmentierung Eine zeitliche Segmentierung kann mit Hilfe des MPEG-7 Scalable-Color Deskriptors durchgeführt werden. Dazu muss zunächst eine Szenenerkennung stattfinden, z.B. auf Basis von Farbhistogrammen. Hier wird die Histogrammdifferenz von zwei benachbarten Einzelbildern als Abstandsmaß verwendet. Nachfolgend können z.B. unter Einsatz der Twin-Comparison-Methode ein oberer und unterer Schwellwert ermittelt werden. Der sogenannte obere Schwellwert dient dabei der Ermittlung von direkten Schnitten, wohingegen der untere Schwellwert dem Auffinden von Übergängen dient. Beides kann nun als Grundlage für die Sequenzermittlung genutzt werden.

Farbsegmentierung

- JSEG Colour Image Segmentation – Die Segmentierung wird in zwei unabhängige Schritte eingeteilt: Farbquantisierung und Raumsegmentierung. [38]

- Mean-Shift Algorithmus für Farbsegmentierung – Die Pixel werden mit einem Farbraum abgeglichen und geclustert. Jeder Cluster repräsentiert einen bestimmten homogenen Bereich im Bild.

- Morphological Greyscale Analysis – Um ein Bild in homogene Teilbereiche zu unterteilen ist es notwendig, die Objektgrenzen ausfindig zu machen. Ein Ansatz hierbei ist das Ermitteln Grauwert-Kontrastunterschiede.[191]

- Watershed Segmentation – Die Watershed Segmentation fußt auf der Idee, dass jedes Graustufenbild als topografische Oberfläche interpretiert werden kann. Dabei stehen dunkle Pixel für den Oberflächengrund und helle Pixel für die Peaks einer Oberfläche.

Bewegungssegmentierung

- Segmentierung basierend auf Bewegungsvektorfeldern – Zunächst erfolgt eine Abschätzung des Bewegungsvektorfeldes, für das ein nicht-parametrisiertes (translatorisches) Bewegungsmodell genutzt wird, da entsprechende Informationen aus dem Untersuchungsbereich nicht vorhanden sind. Dann werden parametrisierte Bewegungsmodelle an das Bewegungsvektorenfeld angepasst („Second Pass“, adaptives Verfahren), beginnend mit einer Startsegmentierung und Abschätzung der Bewegungsparameter für jeden Bereich. Anschließend werden die Pixel mittels DPD (Displaced Pixel Difference) unter Berücksichtigung des Bewegungsmodells geclustert, um schließlich die Finalsegmentierung vornehmen zu können[41]. Ein Nachteil dabei besteht in der Abschätzung des Bewegungsvektorfeldes [79].

- Bewegungsabschätzung basierend auf Segmentierung – Statt der eigentlichen Bewegung wird ein anderes Kriterium zur Erstellung einer Initialsegmentierung benutzt. Die Bewegungsabschätzung erfolgt nun für einen bestimmten Bereich unter Ausnutzung der Initialsegmentierung und wird mit deren Hilfe verfeinert.

2.2.3 Shot Boundary Detection

Ein **Shot** ist eine von einer Kamera aufgenommene Bildfolge, wobei diese Bildfolge durch sogenannte „Shot Boundaries“ (auch: Einstellungsenden) begrenzt wird. Shot Boundaries sind Übergänge zwischen zwei Sequenzen. Man unterscheidet zwei Arten von Übergängen: abrupte Übergänge (Schnitte) und sanfte Übergänge (z.B. Blenden und Fading) [89]. Die meisten Arbeiten beschäftigen sich mit Schnitten, Blenden und Fadings, da diese 99 Prozent aller Übergänge darstellen. Die nachfolgenden Betrachtungen zum Thema, zeigen verschiedene Ansätze auf. Weiterführende Vergleiche können z.B. [89] [102] [18] entnommen werden.

- **Auf Farbe und Helligkeit basierende Ansätze**:
  Viele Ansätze nutzen Farben oder abgeleitete statistische Messungen (z.B. Histogramme), um zu bestimmen, ob sich visuelle Veränderungen ergeben haben, die einen Übergang darstellen.

  - Pixelabgleich – Bei diesem einfachen Ansatz werden Pixelwerte paarweise verglichen. Das Verfahren ist effektiv bei veränderter Helligkeit oder Bewegung in einzelnen Bereichen. Sukzessive Übergänge werden allerdings nur schwer von Bewegungen unterschieden.[208]
  - Blockabgleich – Statt einen Abgleich einzelner Pixel durchzuführen, führt dieses Verfahren den Abgleich der Durchschnittswerte von Pixelblöcken durch.Dieser Ansatz ist weniger störanfällig z.B. gegen Rauschen und langsame lokal beschränkte Bewegungen . Ein Shot Detection Ansatz wird in [24] und [17] zu vorgestellt.
  - Globaler Histogrammabgleich – Histogramme sind ein statistisches Maß, das für die Beschreibung eines Bildes genutzt werden kann. Der Vorteil globaler Histogrammmethoden, gegenüber Pixel- oder Block-basiereten Vergleichen besteht darin, dass sie noch weniger anfällig gegenüber Rauschen und lokal begrenzten Bewegungen sind. Einzelne Objektbewegungen innerhalb einer Szene beeinflussen das globale Histogramm nicht. In [208] findet sich ein Ansatz, der unter Einbeziehung der globalen Bildeigenschaften (etwa Haupt- und Standardabweiung) monochrome Bilder erkennt und diese zur Übergangsbestimmung nutzt.
  - Lokaler Histogrammabgleich – Lokale Histogrammmethoden wurden für die Anwendungsfälle entwickelt, in den lokale Veränderungen starken Einfluss auf das globale Histogramm haben, wie z.B. das Auftreten von Objekten oder Texteinblendungen in einer Szene.

- **Ansätze basierend auf Kantenmerkmalen**:
  Gegenüber farb- und helligkeitsbasierten Ansätzen, haben diese Verfahren u.a. den Vorteil, dass sie weniger anfällig für globale Variationen in der Helligkeit (im moderaten Bereich) und gegenüber lokalen Helligkeitsvariationen sind. Sie können aber stets nur in Verbindung mit farb- und helligkeitsbasierten Verfahren eingesetzt werden [207].

- **Bewegungsbasierte Ansätze**:
  Dieser Ansatz basiert auf der Annahme, dass sich die Pixelintensität entlang einer Bewegungsbahn nicht ändert. Zuerst erfolgt eine Bewegungsabschätzung (Motion Estimation), darauf ein Bildvergleich. Dabei stellen abrupte oder stufenweise Wechsel der Pixelintensitäten einen Hinweis auf einen Übergang dar. Akkerdings gilt die optische Abschätzung des Bewegungsverlaufs als unzuverlässig bzgl. Helligkeitsveränderungen [102].
- **Ansätze basierend auf Merkmalsverfolgung**:
  Um die Informationen aus der Objektverfolgung zur Unterscheidung zwischen Shot Boundaries und Szenen mit hoher Bewegungsaktivität bei angemessenem Berechnungsaufwand und Vermeidung von Fehlabschätzungen zu nutzen, kommen merkmalsbasierte Ansätze zum Ensatz. Merkmalsbasierte Ansätze beruhen auf der Überlegung, dass sowohl abrupte, als auch stufenweise Veränderungen im Wegfallen oder Auftreten von Merkmalen resultieren [19].
- **Änsätze basierend auf der MPEG-Compression-Domain**:
  Dabei muss das Video vor der Verarbeitung nicht extra dekomprimiert werden. Allerdings können nur MPEG-kodierte Videos verarbeitet werden [130].

2.2.4 Optical Character Recognition

Schrifterkennung (Optical Character Recognition, OCR) lässt sich entsprechend der dabei anfallenden Aufgaben unterteilen: Erkennung, Segmentierung und ggf. Verarbeitung. OCR Teilaspekte lassen sich nach weiteren Gesichtspunkten gliedern [105]:

Unterscheidung anhand des Textauftretens: Hierbei gilt es, Szenen- und Overlaytext zu unterscheiden [104] [103]. Szenentext tritt meist in Filmen unvermittelt und unbeabsichtigt auf, etwa auf Straßenschildern. Sichtbar werden diese Texte aus unterschiedlichen Blickwinkeln, Neigungen und Beleuchtungsstuationen. Ebenso sind ihre Oberflächen von unterschiedlicher Natur.

Overlaytexte besitzen eine vordergründige, meist zur Achse der Kamera ausgerichtete Position. Anders als Szenezexte treten Overlaytexte stets beabsichtigt auf. Ihr Auftreten ist demnach durchdacht und ist mit einer definierten Absicht verbunden. Overlaytexte treten z.B. in Nachrichtensendungen oder Werbung auf, also dort, wo eingebettete Bildtexte in komprimierter Form Schlüsselinformationen zum Inhalt des Videos liefern [205].

Ebener Text im 3D Raum, Platzierung: In Abhängigkeit von der Platzierung planer Texte im dreidimensionalen Raum kann unterschieden werden nach:

- Kameraebene horizontal zum Overlaytext
- Ebene des planaren Overlaytexts parallel zur Kameraebene
- ungebundener, ebener 3D Text
- ungebundener 3D Text

Unterscheidung nach Schriftattributen: In einem wohldefinierten Bereich können bestimmte, relevante Schriftattribute ein Merkmal der Unterscheidung darstellen. Dazu zählen Schriftfarbe, Schriftgröße, Schriftschnitt und Schriftart.

Art der Mediendaten: Tritt derselbe Text mehrfach in Erscheinung, können die einzelnen Textinstanzen für eine Verbesserung der Erkennung, Segmentierung und Auswertung genutzt werden. Alternativ kann eine bildweise Segmentierung und Auswertung durchgeführt werden, bei der das Video als Abfolge von unabhängigen Bildern betrachtet wird.

Intendierte Nutzung des Ergebnisses: Zieht man die beabsichte Nutzung der Ergebnisse einer Video OCR in Betracht, unterscheidet man eine Nutzung zum Zweck der Indexierung oder der objektbasierten Videokodierung. Praktische Auswirkungen hat dies z.B. in der Anzahl tollerierbarer Pixelfehler in den Lokalisierungs- und Segmentierungsschritten. Weitere Möglichkeiten der Nutzung bestehen in der Übersetzung von erkanntem Text in eine andere Sprache oder die Entfernung von Text aus dem Video.

Intelligent Character Recognition: OCR wird meist in Verbindung mit Intelligent Character Recognition (ICR) eingesetzt. Die ICR stellt Methoden zur Kontextanalyse bereit, die die Ergebnisse der OCR verbessern oder in Frage stellen können. Eine von der OCR gelieferte Texterkennung wird von der ICR im Kontext betrachtet und ggf. bestätigt oder zur Korrektur gebracht. Dabei unterscheidet man Methoden, mit denen die Qualität des im Einzelbild dargestellten Texts verbessert werden kann u.a. nach folgenden Kriterien:

- Analyse aufeinanderfolgender Einzelbilder zur Verbesserung der Darstellungsqualität (Auflösung, Kontrast, Trennung Bildvordergrund-Hintergrund, Feststellung von Bewegung)
- Analyse von perspektivischen Verzerrungen aufgrund der dargestellten Geometrie des Bildinhalts

ICR setzt ein bestimmtes Vorwissen über die zu untersuchenden Objekte voraus.

Texterkennung

Bei der praktischen Realisierung der OCR lassen sich die zwei Hauptschritte Texterkennung und Textsegmentierung unterscheiden. Die Texterkennung ist Vorbedingung für die Textsegementierung und dient vordergründig der Lokalisierung des Textauftretens innnerhalb des Videos oder der Videoeinzelbilder. Man unterscheidet die Erkennung gesamter Textabschnitte oder die Erkennung einzelner Zeilen. Die Texterkennung kann zur Verbesserung der Lesbarkeit von relevanten Informationen in Videos mit mangelhafter Bildqualität eingesetzt werden, etwa bei Aufnahmen mit Handkameras.

Die Qualität der Texterkennung ist abhängig von den Eigenschaften des Texts selbst.

- **Texteigenschaften**: Die Texteigenschaften beeinflussen die Möglichkeiten einer erfolgreichen Texterkennung in hohem Maße. Charakteristisch für Texterkennung in Schriften mit lateinischem Ursprung sind die in [104] und [103] aufgeführten Eigenschaften. Weitere Texteigenschaften sind:
  - Textausrichtung
  - Zeichenabstand (gleich/unregelmäßig)
  - Zeichengröße (gleich/unregelmäßig)
  - Schriftstärke (gleich/unregelmäßig)
- **Textoberfläche**:
  - Graustufen unbearbeiteter Bildpunkte
  - Lokale Abweichungen
  - Lokale Kantenstärke
  - Kantendichte

Textsegmentierung

Die Aufgabe der Textsegmentierung besteht darin, die Bereiche, in denen Text erkannt wurde, unter Einsatz von verschieden Techniken (z.B. Bildkontrasterhöhung) für die eigentliche OCR vorzubereiten. Ein Problem der Bilderkennung ist unter anderem die eventuell mangelhafte Auflösung. Einzelne Buchstaben sollten mindestens eine Größe von 40 Pixeln aufweisen, gemeinhin beträgt diese im Schnitt elf Pixel. Um dieses Problem zu lösen, kann eine möglichst akkurate Neuskalierung des entsprechenden Textausschnitts erfolgen [79]. Bei [105] werden folgende Teilschritte der Segmentierung aufgeführt:

- Bildkontrasterhöhung (Auflösungserhöhung und Buchstabenreskalierung)
- Videokontrasterhöhung
- Segmentierung (Seedfilting from Border Pixels, Schwellenwertbildung)

OCR-Verarbeitung

Bei der OCR-Verarbeitung, kommen Verfahren zum Einsatz, die die aus den vorangegangenen Schritten gewonnenen visuellen Informationen in gültige Textinformationen übersetzen. Auf dem Markt findet sich eine Vielzahl kommerzieller und freier OCR-Anwendungen, darunter OCRopus[11] (frei), GOCR[12] (frei) oder Adobe Acrobat[13]. Diese Systeme bieten unterschiedliche Funktionalitäten und Methoden:

- Optional sind dabei Funktionen zur Layouterkenung (siehe Abschnitt 2.2.4)

[11] `http://code.google.com/p/ocropus/`
[12] `http://sourceforge.net/projects/jocr/`
[13] `http://www.adobe.com/acrobat/`

- Funktion der Spracherkennung
- statistischen Methoden zur Sprachmodellierung

2.2.5 Gesichtserfassung und -erkennung

Sollen in den zu analysierenden Videodaten Personen erkannt und identifiziert werden, beginnt dieser Prozess meist mit der Gesichtserfassung, d.h. mit der Identifikation all derjenigen Bildbereiche, die ein menschliches Gesicht beinhalten.

- **Gesichtserfassung** (Face Detection):
  Die Identifikation menschlicher Gesichter in Videodaten ist keine triviale Aufgabe, da Gesichter in unterschiedlichen Größenordnungen, Blickrichtungen, Drehbewegungen und Verzerrungen im Video auftreten können. Weitere Schwierigkeiten bereiten unterschiedliche Beleuchtungssituationen und komplexe Bildhintergründe. Gesichter selbst sind keine starren Objekte, sondern sind einer Vielzahl von Eigenbewegungen und Veränderungen durch die Gesichtsmimik unterworfen. Hilfreiche Zusammenfassungen der unterschiedlichen Methoden findet man in [204, 68]. Man unterscheidet folgende Verfahren der Gesichtserfassung:
  - **Wissensbasierte Methoden**: Die Grundlage dieser Verfahren bilden regelbasierte Ansätze, die das Wissen um Proportionen und andere Gesichtseigenschaften hierarchisch in einem Top-Down Ansatz in Regeln ausdrücken. Allerdings ist das Formulieren dieser Regeln nicht immer trivial und das Erkennen mehrerer Personen bzw. von Personen in unterschiedlichen Posen und Bewegungen vor komplexen Hintergründen schwierig.
  - **Eigenschaftsbasierte Methoden**: Ausgehend von Charakteristika niedriger Komplexität (Low-Level Features) wie z.B. Farbe, Umriss, Textur und Kanten sollen komplexere Charakteristika (High-Level Features) identifiziert werden (Augen, Nase, Mund, etc.). Aus diesen wiederum werden Kandidaten zu identifizierender Gesichter ermittelt, die durch weiterführende Analyse verifiziert werden [99, 206]. Die identifizierten Charakteristika sind zwar invariant bzgl. Lage und Bewegung, ihre Identifikation wird aber durch unterschiedliche Beleuchtung, Rauschen oder Abdeckung verfälscht.
  - **Template-basierte Methoden**: Ausgehend von standardisierten Gesichtsvorlagen wird mit unterschiedlicher Skalierung eine erschöpfende Suche über den gesamten Bildinhalt durchgeführt. Die Vorlagen basieren auf Kanten, Kontouren oder Regionen, die manuell erstellt wurden. Obwohl einfach zu implementieren stoßen diese Verfahren bei unterschiedlichen Blickrichtungen und Bewegungen schnell an ihre Grenzen.
  - **Maschinelles Lernen**: Im Gegensatz zu Template-basierten Methoden werden hier die Vorlagen anhand derer Gesichter identifiziert werden sollen mit Hilfe unterschiedlicher Verfahren, wie z.B. Neuronale Netze, Support Vektor Maschinen (SVM), Naive Bayes Klassifizierer, Hidden Markov Modelle, etc. erlernt. Diese Verfahren sind robust gegenüber unterschiedlicher Beleuchtung, Blickrichtungen und Bewegungen.

Erfolgt die Gesichtserfassung nicht im Einzelbild, sondern in einer Videosequenz, können zusätzliche Informationen die Erfassung erleichtern und beschleunigen. So kann z.B. die Einschränkung auf Sequenzen, die menschliche Sprache enthalten, den Suchraum erheblich verkleinern. In [52] werden die populärsten Softwarelösungen zur Gesichtserfassung zusammengefasst.

- **Gesichtserkennung** (Face Recognition):
  Die Gesichtserkennung erfuhr in den vergangenen Jahren vermehrt Aufmerksamkeit, insbesondere da sie verstärkt in der Überwachung zur Prävention gegen Verbrechen und Terrorismus zum Einsatz kommt. Allerdings ist heute eine fehlerfreie Gesichtserkennung in komplexen Bewegungs- und Beleuchtungssituationen noch nicht immer möglich. Eine Zusammenfassung der gängigen Methoden der Gesichtserkennung liefert [215].

  - **Gesichtserkennung in Einzelbildern**: Hier unterscheidet man **holistische Methoden**, bei denen das vollständige Gesicht als Eingabe dient (Principal Component Analysis, Eigenfaces, SVM, etc,), **Eigenschaftsbasierte Methoden**, die lokale Charakteristika, wie z.B. Nase, Augen und Mund, deren Lage und statistische Größen als Eingabe verwenden (Dynamic Link Architecture, Hidden Markov Modelle, Neuronale Netze, etc.), sowie **hybride Methoden**, die beide Ansätze miteinander kombinieren.
  - **Gesichtserkennung in Videosequenzen**: Werden anstelle von Einzelbildern ganze Videosequenzen analysiert, kommen Methoden auf der Basis von **Gesichtssegmentierung** zum Einsatz, wobei zuerst bewegte Objekte erkannt und segmentiert werden, auf denen dann Einzelbildverfahren zum Einsatz kommen. Aus den erkannten Gesichtsflächen werden anschließend Bewegung und Posen abgeschätzt, um zu einer virtuellen Frontalansicht des zu identifizierenden Gesichts zu gelangen. Danach wird das Gesicht und seine Charakteristika über die Dauer der Sequenz verfolgt (**Gesichtsverfolgung**, Face Tracking), um ein Gesichtsmodell zu rekonstruieren **Gesichtsmodellierung**. Die verfolgten Charakteristika können dazu verwendet werden, um Gesichtsmimik zu erkennen. Aktuelle Software zur Gesichtserkennung wurde in [32] zusammengestellt und evaluiert.

2.2.6 Defekt- und Qualitätsanalyse

Mit Hilfe der Analyse der Qualität audiovisueller informationstragender Objekte und auf darin vorhandene Defekte können verschiedene Ziele verfolgt werden. Mit Hilfe der Indexierung einer Beschreibung von Qualität und Defekten kann eine Qualitätsabschätzung von Archivinhalten getroffen werden. Zusätzlich kann diese Indexierung im Rahmen der inhaltsbasierten Suche verwendet werden. Erfolgt eine Beschreibung des zeitlichen Verlaufs von Qualität und Defekten kann diese Information direkt bei der Restauration des betroffenen Archivmaterials unterstützend eingesetzt werden.

Defekte lassen sich modal untergliedern (visuell / auditiv) bzw. nach dem jeweiligen Medientyp (Film, Video, etc.). Man unterscheidet nach [1]:

- **Filmdefekte**, wie z.B. Flickern, Unregelmäßigkeiten in der Bildgeschwidigkeit, Kratzer, Verschmutzungen, fehlende Einzelbilder und Sequenzen, altersbedingtes Verblassen und Beschädigungen des Filmmaterials, usw.

- **Videodefekte**, wie z.B. Dropouts, Verzögerungsschwankungen (Jitter), Ghost-Effekte, Head-Clogging, Cross-Chroma, etc.

- **Medienunabhängige Defekte und Qualitätskenngrößen**, wie Störungen, Rauschen, Unschärfe, etc.

- **Kodierungsbezogene Defekte und Qualitätskenngrößen**, wie z.B. Komprimierungsartefakte, etc.

Zusätzlich kann die Analyse und Messung der betreffenden Kenngrößen in Bezug auf eine gültige Referenz erfolgen (Reference Based Measurement, im Gegensatz zu Non-Reference Based Measurement).

- **Reference Based Measurement**:
  Die betreffenden Kenngrößen zur Qualitäts- und Defektbestimmung werden durch Expertengremien und Standardisierungsorganisationen aus den Bereichen Video Broadcast und Video Delivery festgelegt, wie z.B. der Video Quality Experts Group[14] oder der ANSI T1.801.03-1996[15].

- **Non-Reference Based Measurement**:
  Durch ihre Unabhängigkeit vom Vorhandensein von qualitativ fehlerfreien Referenzdaten finden diese Verfahren trotz der schwierigeren Bewertungssituation weit häufiger Anwendung. Üblicherweise wird die menschliche Wahrnehmung als Maßstab herangezogen, was aber aufwändigere statistische Berechnungen und Tests zur Folge hat [33, 47, 195, 20, 143, 201, 25, 27].

Bei der Bestimmung und Untersuchung von Defekten hervorgerufen von Film- und Videomaterial fokusierten sich die bisherigen Untersuchungen auf detaillgenaue spatiotemporale Analyse zum Zweck der späteren Restauration [15, 22, 78, 88, 152, 194]. Allerdings sind diese Verfahren bei der Bewältigung großer Mengen Archivmaterials nur schwer skalierbar.

2.3 Gemeinsame Audio- und Videoanalyse

Die Analyse von audiovisuellen, informationstragenden Objekten vereint die bereits in den vorangegangenen Kapiteln behandelte Techniken zur Audio- und Videoanalyse, um durch deren kombinierte Anwendungen Synergieeffekte ausnutzen zu können, damit qualitativ bessere Analyseergebnisse erzielt werden können.

[14] http://www.its.bldrdoc.gov/vqeg/
[15] http://www.its.bldrdoc.gov/n3/video/standards/

2.3.1 Szenen-Segmentierung

Ähnlich der Boundary Shot Detection ist das Ziel der Szenen-Segmentierung (auch Story Segmentierung) die zeitliche Zerlegung audiovisueller informationstragender Objekte in inhaltlich kohärente Untereinheiten. Dabei bestehen allerdings folgende fundamentale Unterschiede zur Shot Boundary Detection:

- Während die Shot Boundary Detection die festdefinierten Grenzen einer zusammenhängenden Kameraaufnahme eindeutig bestimmt, kann die zeitliche Zerlegung in inhaltlich zusammenhängende Einheiten auf unterschiedliche Weise, abhängig vom jeweiligen Standpunkt und Abstraktionsniveau des Betrachters erfolgen. Die Grenzen sind also nicht eindeutig festgelegt.
- Während die Shot Boundary Detection direkt von Low-level Charakteristika abhängt und daher direkt aus dem audiovisuellen Material ermitteln werden kann, hängt die Ermittlung inhaltlich koheränter Abschnitte von High-level Charakteristika ab, die sich auf einem höheren Abstraktionsniveau befinden. Um diese zu bestimmen, muss die sogenannte „semantische Lücke“ überwunden werden, da die Low-level Charakteristika gemäß ihrer Bedeutung interpretiert werden müssen.
- Während die Shot Boundary Detection lediglich von visuellen Charakteristika abgeleitet werden kann, können auditive Charakteristika über die Shot Boundary hinweg fortdauern. Eine zusammenhängende Szene hängt von multimodalen Charakteristika ab. So kann die Grenze eines inhaltlich zusammenhängenden Segments innerhalb oder auch erst außerhalb eines Shots liegen.

Seit 2003 ist die Szenen-Segmentierung Bestandteil der TREC Video Retrieval Evaluation.

2.3.2 Szenen Definition

Die inhaltliche Bedeutung einer Szene zu bestimmen ist keine triviale Aufgabe und benötigt vom künstlerischen Standpunkt aus betrachtet Kontextwissen, Welt- und Erfahrungswissen, dessen Ermittlung im Nachhinein nahezu unmöglich ist. Eine logisch zusammenhängende Szenen-Einheit (Logical Story Unit, LSU) kann definiert werden als zeitlich zusammenhängende Folge von Shots, die durch einzelne Inhaltselemente miteinander verbunden sind [94]. Verfeinert wurde diese Definition durch die Einbeziehung wahrgenommener Grenzen in Zeit und Raum [190], d.h. Shots der selben LSU sind einander immer ähnlich.

Basierend auf einer länger andauernden Konsistenz von Audio und Videosignal kann alternativ ein Konzept der **berechenbaren Szenen** definiert werden, die sich eindeutig aus Low-level Charakteristika ableiten lassen [176], zusammen mit zwei für die automatische Szenen-Segmentierung notwendigen Bedingungen ein citeHan04:

- **Berechenbarkeit** (Computability), d.h. die Existenz einer Menge von Charakteristika, mit der es möglich ist, Veränderungen im inhaltlichen Zusammenhang zu ermitteln.

- **Übersetzbarkeit** (Parsability), d.h. wenn der Inhalt aus semantischen Segmenten zusammengesetzt wurde und der inhaltliche Zusammenhang berechenbar ist, dann ist der Inhalt auch übersetzbar.

2.3.3 Verschiedene Lösungsansätze

Die unterschiedlichen Lösungsansätze für das Problem der Szenen-Segmentierung lassen sich untergliedern in Verfahren, die auf einem vorhandenen Vorwissen bzgl. der Struktur und Inhalte eines Sachgebiets basieren (heuristische, Regel-basierte Verfahren oder Modell-basierte Verfahren), und Verfahren, die statistische Klassifikationsverfahren einsetzen und mit Beispieldatensätzen trainiert werden (statistische Verfahren) [70, 93]. In der Praxis lassen sich die gegenwärtigen Verfahren nach ihrem jeweiligen Einsatzgebiet klassifizieren (Nachrichten, Sport, Spielfilm, etc.).

- **Segmentieren von Nachrichten**:
  Da Nachrichtensendungen meist einem bestimmten Aufbau folgen (Nachrichtensprecher, Bildbeiträge, Wettervorhersage, etc.) fällt ihr Segmentierung leichter als die Segmentierung anderer Klassen von Videoinhalten. Die meisten Ansätze zur Segmentierung von Nachrichtensendungen basieren auf einer kombinierten Analyse von visuellen, auditiven und sprachlichen Charakteristika. Visuelle Ähnlichkeit verschiedener Shots innerhalb eines Zeitfensters in Verbindung mit dem zeitlichen Abstand zwischen den Shots [122, 67, 43, 69]. Zusätzlich werden Ähnlichkeiten in erfassten Gesichtern innerhalb der einzelnen Shots verwendet [67]. Andere Verfahren verwenden das Vorwissen um die strukturelle Ähnlichkeit der meisten Nachrichtensendungen, um z.B. Shots mit einem bestimmten Nachrichtensprecher zu identifizieren (über verschiedene Charakteristika [69, 135] oder über eine Ähnlichkeits-basierte Suche [70, 67, 16, 137, 69, 43]. Daneben werden noch Sprecherwechsel [69, 137, 26], Wechsel zwischen Musik und Sprache [137, 26, 69], Erfassung von Audio-Jingles und weitere Veränderungen in der akustischen Umgebung erkannt (z.B. Studio oder Außenaufnahme) [67].

  Eine weitere Möglichkeit besteht in der Zuhilfenahme von textuellen Daten aus manuellen oder automatisch erstellten Transkriptionen der sprachlichen Informationen. Dabei wird das Auftreten ähnlicher Wörter in mehreren verschiedenen Shots [69, 67, 135, 16] oder das Auftreten sogenannter Trigger-Phrases [43] analysisert, die einen Hinweis auf das Nachfolgen unterschiedlicher Shot-Typen liefern. Einige dieser Ansätze beruhen auf überwachten Lernverfahren. In [69] und [70] werden mögliche Kandidaten für Segmentgrenzen in inhaltliche Segmentgrenzen und bloße Shot-Boundaries mit Hilfe von Support Vektor Maschinen (SVM) klassifiziert, [26] führt eine Klassifikation mit Hilfe eines Hidden Markov Modells durch.

  Der Ansatz von [210] nutzt einen Shot Connectivity Graph (SCG) zur Klassifikation. Dabei werden einzelne Shots in einem Graphen angeordnet, dessen Kanten einen Übergang von einem Shot zu einem anderen repräsentieren. Unter der Prämisse, dass der Nachrichtensprecher immer wieder im Laufe einer Nachrichtensendung auftaucht, wird dabei nach Zyklen im Graph gesucht Spezielle

Arten von Shots können über das Auftreten bestimmter Schlüsselwörter (Key Term Spotting) identifiziert werden.

- **Segmentieren von Spielfilmen**:
  Szenen-Segmentierung in Spielfilmen wurde zuerst mit Hilfe einfacher visueller Charakteristika angegangen. Dabei wurde von der Annahme ausgegangen, dass zusammengehörige Shots einer Szene ähnlichere Eigenschaften in Bezug auf Helligkeits- und Farbeigenschaften besitzen, als nicht zusammengehörige Shots [145, 65]. Um den Suchraum zu verkleinern, werden die Shots lediglich durch Key Frames repräsentiert. Zur differenzierteren Klassifikation einzelner Segmente werden unterschiedliche Segment-Typen definiert und die Analyse wird auf Bildregionen beschränkt, die robust gegenüber Veränderungen durch Bewegung sind [178] in Verbindung mit der Analyse der Szenendynamik und verschiedener Szenenübergänge (Fading, Dissolving, etc.) [139, 181].

  Analog zu Abschnitt 2.3.2 werden Segnentgrenzen ebenfalls auf Basis eines berechenbaren Szenenmodells bestimmt [176]. Dabei werden Kandidaten für Szenenbegrenzungen zuerst getrennt aus der Audio- und der Videoanalyse ermittelt und anschließend korreliert, wobei Farb- und Belichtungsähnlichkeiten in Key Frames, Korrelationen im Audiosignal und Pausen (Stille) ausgenutzt werden.

- **Segmentieren von Sportbeiträgen**:
  [9] schlägt eine Modell-basierte Segmentierung von Sportsendungen vor, die hauptsächlich durch Sprecherwechsel im Sprachsignal bestimmt wird. Anschließend finden eine Klassifikation mit Hilfe eines Bayes-Netzwerks statt, aus der eine hierarchische Struktur entwickelt wird. Ein weiterer Modell-basierter Entwurf verwendet Farbe, Objektlayout und Kantenstruktur als bestimmende Charakteristika [216].

Insgesamt verbessert die Einbeziehung von Audio- und Videoinformationen zur Segmentierung die Ergebnisqualität, wobei zusätzlich die Anwendung von Spracherkennung die Präzision (Precision) der Ergebnisse verbessert, während dagegen weniger der tatsächlich vorhandenen Segmente erkannt werden können (Recall). Verschiedene Methoden der Evaluation sowie Problemfälle, die mit den vorhandenen Methoden nicht korrekt klassifiziert werden können, werden in [190, 92] diskutiert.

2.3.4 Shot- und Szenenklassifikation

Klassifikationsaufgaben für informationstragende audiovisuelle Objekte können in unterschuiedlicher Granularität, beginen bei Einzelbildern bis hin zu ganzen Sendefolgen und Programmen durchgeführt werden. Die Klassifikation kann nach Zugehörigkeit zu unterschiedlichen Genres (z.B. Werbung, Spielfilme, Nachrichten, etc.) erfolgen. Szenen können inhaltsbezogen nach darin gezeigten Umgebungen (z.B. Innen- oder Außenaufnahmen, Studio- oder Liveproduktionen, etc.), Ereignissen oder Objekten klassifiziert werden oder aber auch einfacher nach Aufnahmerichtung oder Blickwinkel, um bestimmte Orte identifizieren zu können.

- **Genreklassifikation**
  Eine Auflistung aller typischer Genres, nach denen eine Klassifikation informationstragender audiovisueller Inhalte erfolgen kann, liefert [164].
  - Ein frühes Verfahren nach [49] führt dabei zuerst eine syntaktische Analyse des Videos durch, gefolgt von einer Extraktion der stilbestimmenden Attribute und Eigenschaften, die schließlich verschiedenen Genres (Nachrichten, Autorennen, Tennis, Werbung, Cartoons) zugeordnet werden.
  - Ein weiteres Verfahren [182] klassifiziert Videos nach den Genres Sport, Nachrichten, Werbung, Cartoons und Musik mit Hilfe eines Entscheidungsbaum-basierten Ansatzes, wobei Charakteristika, wie z.B. Cliplänge, Fade-In und Fade-Out, Kamera Bewegungsaktivität, Änderung der Lichtverhältnisse, Länge von erkannten Objektbewegungen, Farbkohärenz und Helligkeitsverteilung zur Bewertung herangezogen werden.
  - Weitere Ansätze klassifizieren audiovisuelles Material in disjunkte Kategorien, wie z.B. Cartoon oder Realfilm [74], Werbung oder Nicht-Werbung [59], wobei letzteres Verfahren auch erfolgreich zur Erkennung von Spielhöhepunkten in Fußballspielen zum Einsatz kam.
  - Ein alternativer Anstz basiert auf der Unterscheidung grundlegender Szenentypen innerhalb eines Videos, wie z.B. Dialog mit Aktion, Dialog ohne Aktion und Aktion ohne Dialog, basierend auf den Charakteristika Bewegungsintensität, Shotlänge, Audiointensität und Audioklassifikation [173].
  - In [39] wird ein Klassifikationssystem vorgestellt, das auf der Erkennung von eingeblendetem Text (in Abhängigkeit der erkannten Zeilenanzahl) und Gesichtern (Größe und Position) beruht, und aus den erkannten Charakteristika das Video in eines der Genres (Nachrichten, Werbung, Sitcom, Soap) über ein Hidden Markov Modell zuweist.
  - In [138] wird ein Verfahren vorgestellt, das Spielfim Previews anhand der jeweiligen Shotlänge und Bewegungsintensität den Kategorien Action und Nicht-Action zuweist. Erkannte Action-Szenen werden auditiv analysiert und dabei Feuer- und Explosionsszenen erkannt, Nicht-Action-Szenen werden anhand ihrer Beleuchtungsverhältnisse in Kommödien, Horror und Drama unterteilt.
  - In [76] werden Audio-Charakteristika, die im MPEG-7 als Deskriptoren festgelegt sind, zusammen mit visuellen Charakteristika, wie z.B. Kamerabewegung, Bewegungsaktivitäten und HSV Histogramme, genutzt, um die Genres Cartoon, Drama, Musikvideo, Nachrichten und Sport zu unterscheiden.
  - In [57] wird ein 8-dimensionaler Eigenschaftsvektor vorgeschlagen, der zeitliche Variationen innerhalb der einzelnen Shots (Activity Power Flow) und Bewegungsintensität quantifiziert, wobei der Activity Power Flow von den bei der MPEG-Komprimierung entstandenen Blöcken (mit zuverlässigem bzw. unzuverlässigem Bewegungsvektor) abgeleitet wird. Die Klassifikation in die Genres Sport, Drama, Landschaft und Nachrichten erfolgt mit Hilfe eines Radial Base Function Network (RBF).

- **Konzepterkennung**
  Die Konzepterkennung umfasst die automatische Erkennung bestimmter inhaltlicher Konzepte (Objekte, Orte, Ereignisse) in informationstragenden audiovisuellen Objekten.

  - Der Ansatz von [121] basiert auf der Einführung des Konzepts eines probabilistischen Multimediaobjekts (Multijekt). Multijekte werden anhand charakteristischer Eigenschaften, wie z.B. Farbe, Textur, Umriss mit Hilfe von Support Vektor Maschinen kategorisiert und abgeleitet. Aus den erkannten Multijekten wird ein Netzwerk erstellt, das die Multijekte miteinander über positive oder negative Assoziationen in Bezug setzt (eine positive Assoziation wäre „Straße“ und „Außenaufnahme“, während „Straße“ in Verbindung mit „Studioaufnahme“ eine negative Assoziation darstellt).
  - In [140] wird der Einsatz selbstorganisierender Landkarten (Self-organizing Maps) [87] zur Erkennung semantischer Konzepte vorgschlagen.
  - In [165] werden aus Key Frames des Videos visuelle Charakteristika extrahiert und mit der Bewegungsaktivität und einer Text-Transkription korreliert. Die Klassifikation erfolgt eines Trainingsdatensatzes in Verbindung mit einem k-NN-Klassifizierer, Support Vektor Maschinen und einem speziellen Klassifizierer für Schlüsselwörter.
  - In [6] liegt der Klassifikation und Ereigniserkennung ein lernendes Bayes-Netzwerk zu Grunde. Dabei werden sowohl Farbe und Textur als auch Audio FFT (Fast Fourier Transformation) und MFCC (Mel Frequency Cepstrum Coefficients) Koeffizienten, Bewegungsenergie, Gesichtserkennung und automatische Texterkennung (OCR) verwendet, um Low-level Charakteristika zu bestimmen, die auf 168 Konzepte abgebildet werden und mit Hilfe einer SVM klassifiziert werden. In einem Folgeschritt werden Kombinationen aus Konzepten genutzt, um Klassen und Klassenzugehörigkeit daraus logisch abzuleiten [31].
  - In [163] wird das Verfahren „Semantic Value Chain“ zur Konzepterkennung vorgestellt, eine aus drei Gliedern bestehende Prozesskette, dem Content Link mit Low-Level Audio-, Video- und Text-Charakteristika, Style Link zur Erkennung der Faktoren Layout (basierend auf Shot-Länge, Text-Overlay oder Sprache), Inhalt (Gesichter und deren Position, Autos, Bewegungserkennung, benannte Entitäten aus Text und Sprachinformation), Capture (Kameradistanz- und Bewegung) und Concept (Namen von Personen Links zum Inhalt), und Semantic Content Link, das ein Lexikon mit 32 vordefinierten semantischen Konzepten verwendet. Die Klassifikationen der ersten beiden Glieder beruhen auf der Anwendung von Support Vektor Maschinen.

- **Kategorisierung**
  Innerhalb automatisch erkannter oder manuell zugewiesener Genres wird in diesem Aufgabenschritt eine differenziertere Klassifizierung durchgeführt, deren Methoden jeweils von dem ausgewählten Genre bestimmt werden.

- In [197, 174] werden Klassifikationsverfahren für **Fußballspiele**, basierend auf visuellen Charakteristika (Bewegungsaktivität, Farbähnlichkeiten in Segmenten mit ähnlicher Bewegung) und auditiven Charakteristika getrennt durchgeführt, indem aus den Charakteristike Schlüsselwörter abgeleitet werden, die mit Hilfe von Support Vektor Maschinen aggregiert und fusioniert werden. Ein weiteres Verfahren basiert auf einem Szenenalphabet, das sich an dem in jedem Frame gezeigten Ausschnitt des Spielfeldes orientiert und Hidden Markov Modelle zur Klassifikation verwendet [127]
- In [40] wird ein auf Support Vektor Maschinen beruhendes Klassifikationssystem für Szenen vorgestellt, das Relevanzfeedback-Information des Benutzers einbezieht. Ähnlich arbeiten binäre Klassifikatoren zur Diskriminierung zwischen **Außenaufnahmen und Studioaufnahmen** (anhand von Farb- und Texturinformationen in Einzelbildern) [156]. Eine analoge binäre Klassifikation wird in [177] vorgestellt, die einen k-nearest-neighbor Klassifizierer nutzt.
- In [187] wird ein Verfahren zur Klassifikation von Szenen in **Landschafts- und Stadtaufnahmen** vorgeschlagen, die weiter in die Kategorien Sonenaufgang, Sonnenuntergang, Wald und Gebirge aufgeteilt werden mit Hilfe der Charakteristika Farbhistogramm, Farb-Kohärenzvektor, DCT Koeffizienten (Diskrete Cosinus Transformation), Kantenrichtungs-Histogramm und -Kohärenzvektor. In einer Folgearbeit werden zusätzlich noch Vegetation und Himmel in den Szenen identifiziert [186].
- Der in [151] beschriebene Ansatz versucht Szenen zu klassifieren, die am **selben Ort** aufgenommen wurden bzw. die **identische Objekte** beinhalten.
- In [63] wird zuvor eine manuelle Klassifikation vorgenommen, indem Probanden Bilder entsprechend ihrer Ähnlichkeit zu vorgegebenen Beispielbildern zuordnen und dabei den Grad der Ähnlichkeit festlegen sollen. Daraus wird eine globale Distanzmatrix ermittelt und in einen zweidimensionalen Raum projiziert (Curvilinear Component Analysis, CCA) um daraus ein Modell entsprechend dem menschlichen Klassifikationsvermögens zu erstellen.

- **Affektive Inhaltsanalyse**
  Die bisher beschriebenen Faktoren der inhaltlichen Analyse, wie z.B. die Identifikation verschiedener Genres oder Objekten, liegen ausschließlich auf einer kognitiven Ebene. Daneben kann der Inhalt auch auf einer affektiven Ebene beschrieben werden, die der Intention des Inhaltsproduzenten bzgl. des damit erreichten (als subjektiv zu beurteilenden) Affekts des Inhaltskonsumenten entspricht. Zur Affektiven Inhaltsanalyse zählt z.B. die Identifikation der lustigsten Szenen innerhalb eines Cartoons oder einer Kommödie bzw. die Spannendsten oder aufregendsten Szenen innerhalb eines Horror-Films. Affektive Inhaltsanalyse versucht so die beabsichtigte affektive Reaktion des Betrachters zu ermitteln.

In [64] werden drei Dimensionen der affektiven Reaktion spezifiziert, die auf der Basis von auditiven und visuellen Low-Level Charakteristika bestimmt werden: Valenz (angenehm oder unangenehm), Erregung (Ruhe oder Aufregung) und Dominanz (kontrolliert oder unkontrolliert).

2.3.5 Erkennung von Ereignissen

Bei der Erkennung von Ereignissen (Event Detection) geht es um die Lokalisierung einzelner Segmente in audiovisuellen informationstragenden Objekten, die jeweils eine relevante Aktion enthalten, wie z.B. einen Dialog oder ein Tor in einem Fußballspiel. Allgemein und objektiv lässt sich dabei nur schwer festlegen, was genau man unter einem Ereignis versteht. Das Ereignis ist immer an den Kontext eines bestimmten Inhalts gebunden, der dessen Relevanz bestimmt.

Erkennung von Ereignissen lässt sich klar mit den bereits beschriebenen Klassifikationsaufgaben in Bezug bringen. So beruht die Erkennung von Ereignissen oft auf der zuvor durch eine Kategorisierung getroffene Festlegung von Genre oder anderen inhaltsbestimmenden Elementen.

Ein großer Teil der hier angesiedelten Verfahren bezieht sich auf Videos mit Sportinhalten, da diese meist einer festen Struktur folgen [2] und nur eine begrenzte Anzahl von Ereignissen darin relevant ist. Dialoge dagegen sind wichtige Ereignisse in Nachrichten oder Spielfilmen.

Im Bereich der Sicherheitsüberwachung (Surveillance) ist die Erkennung von oft sicherheitsrelevanten Ereignissen, wie z.B. kriminellen Handlungen, besonders wichtig. Allerdings unterliegen Sicherheitsüberwachungen gegenüber allgemeinen audiovisuellen Objekten anderen, z.T. sehr eingeschränkten Rahmenparametern, wie z.B. fixer Kamerastandpunkt, etc.

- **Ereigniserkennung in Sportvideos**
  Eine Übersicht über die verschiedenen Verfahren der Event Detection in Sportvideos wird in [2] engegeben, wobei die meisten Verfahren vom jeweiligen Domänenwissen abhängen, d.h. Wissen, das bereits vorab über audiovisuelle Charakteristika relevanter Ereignisse bekannt ist (wie z.B. Jubeln und Torrufe nach einem Tor beim Fußball).

  - Basierend auf Bewegungseigenschaften, der Tatsache, das Torsituationen meist zweimal aus unterschiedlichen Blickwinkeln gezeigt werden und der Amplitude des Audiosignals wird in [98] ein Verfahren zur automatischen Identifikation von Torsituationen in Fußballspielen präsentiert. Dabei konnte festgestellt werden, dass für diese spezielle Aufgabe auch die Analyse der Videoinformation alleine schon ausreicht und nicht notwendigerweise mit der Audioinformation korreliert werden muss. Die Audioinformation alleine hingegen ist nicht für diesen Zweck ausreichend.
  - Bei Fußballspielen folgt die Kamerabewegung der Bewegung des Balls im Spiel. In Verbindung mit der Erkennung der Spielfeldlinien können damit Aktionen im Strafraum als Ereignisse identifiziert werden [14].

- In [202] wird eine reine audiobasierte Ereigniserkennung vorgestellt, die sich auch auf andere Sportarten anwenden lässt. Sie basiert auf der Voraussetzung, dass relevante Ereignisse über Applaus und Beifall zu erkennen sind, die über Charakteristika wie z.B. Hintergrundgeräuschpegel, Audiosegmentklassifikation, etc. mit Hilfe eines Hidden Markov Modells erkannt werden.
- Speziell für alle Feldsportarten eignet sich das Verfahren von [148], das zur Entdeckung relevanter Ereignisse die Bedeckung von Grasflächen, den Winkel der Spielfeldbegrenzungen, Nahaufnahmen, Bewegungsaktivität und Aktivität des Sprachsignals heranzieht und mit Hilfe von Support Vektor Maschinen klassifiziert.
- In [81, 197] werden eine Ereigniserkennung auf Basis der Klassifikation einzelner Segmente vorgeschlagen, bei der den Segmenten in Abhängigkeit der Klassifikation Schlüsselwörter zugewiesen werden, die anschließend aggregiert werden und aus denen relevante Ereignisse abgeleitet werden.
- Weitere Ansätze legen der Ereignisentdeckung in Sportvideos die Prämisse zu Grunde, dass dort relevante Ereignisse üblicherweise in Zeitlupe wiederholt werden. In [126] weden mit Hilfe eines Hidden Markov Modells Kandidation für Zeitlupensegmente klassifiziert und über weitere Charakteristika, wie z.B. Farbähnlichkeiten von Werbebeiträgen unterschieden.
- Basierend auf einem semantischen Modell für relevante Ereignisse beim Fußballspiel werden auch eine regelbasierte Ansätze verfolgt. Dabei werden Spieler und Ball verfolgt, um Ereignisse als Folgen von Interaktionen zwischen Spielern und Ball zu modellieren [180].

- **Ereigniserkennung in Spielfilmen**
  Relevante Ereignisse in Spielfilmen sind meist Dialoge, daher ist die Erkennung von Dialogszenen mit zwei oder mehreren Personen Gegenstand verschiedener Methoden und Verfahren, die sich in drei verschiedene Gruppen klassifizieren lassen:

  - **Szenen-basierte Klassifikation**:
    Bei diesen Verfahren wird zunächst eine Segmentierung der audiovisuellen Daten durchgeführt, bevor eine Klassifikation bzgl. Dialogszene oder Nicht-Dialogszene stattfindet. In [93] wird ein Verfahren vorgestellt, das eine Klassifikation in 2-Sprecher Dialoge, Multi-Sprecher Dialoge und hybride Ereignisse durchführt. Zuerst wird dabei eine Shot-Erkennung, gefolgt von einer Gruppierung der einzelnen Shots in ganze Szenen vorgenommen auf Bases der visuellen und akustischen Ähnlichkeit. Danach findet eine regelbasierte Klassifikation auf Basis heuristisch gewonnener Regeln statt.
  - **Direkte Dialog-Szenenerkennung**:
    Bei diesen Verfahren werden in einem ersten Schritt Charakteristika der einzelnen Shots bestimmt, aus denen danach direkt über ein Hidden Markov Modell Dialogszenen identifiziert werden. In [5] und [4] wird ein ganzes Framework zur Erkennung von Dialogszenen vorgestellt, das auf Basis

akustischer Information (Sprache, Stille, Musik) und visueller Information (Gesichts- und Ortsveränderungen) basiert. Einer Shot-Erkennung folgt eine Extraktion audiovisueller Charakteristika in Form eines Feature-Vektors, der über ein Hidden Markov Modell klassifiziert wird.

Das in [30] vorgeschlagene Verfahren erkennt Dialog- und Actionszenen und basiert auf einem Top-Down Ansatz über Video-Editier-Regeln. Dabei wird ein Audio-Klassifizierer auf Basis einer Vektor Support Maschine als Eingabe für einen anschließenden endlichen Automaten (Finite State Machine) zur Identifikation verwendet.

Ein weiteres Verfahren dient der Segmentierung eines Videos in vordefinierte Szenentypen (Dialoge, Geschichten, Aktionen und Rest), indem zuerst Audio- und Video-Shots unabhängig voneinander erkannt bestimmt werden und anschließend werden aufeinanderfolgende Video-Shots zusammengefasst zu Gruppen mit ähnlichen auditiven und visuellen Charakteristika [150].

Als alternative Kriterien für die Segmentierung werden in [133] folgende Varianten vorgeschlagen: Szenen mit ähnlicher Audiocharakteristik, Szenen mit ähnlichem Schauplatz und Dialoge. Dieses Verfahren bestimmt als erstes Video-Shot-Grenzen, danach werden die Audio- und Farbcharakteristika zusammen mit Orientierung und Gesichtern ermittelt. Danach wird eine Distanztabelle konstruiert, die die Distanzzweier Video-Shots bzgl. der ermittelten Charakteristika angibt, mit deren Hilfe Video-Szenen aus den Video-Shots zusammengesetzt werden.. Audio-Shots werden über Vektoren aus Audio-Charakteristika bestimmt, die mit den Video-Szenen korreliert werden. Dialoge werden anschließend durch eine Gesichtserkennung in den einzelnen Video-Szenen erkannt, wenn in einer Szene abwechselnd verschiedene Gesichter erkannt werden.

– **Shot-basierte Klassifikation**:
In diesen Verfahren findet als erstes eine Klassifizierung der Shots statt, ob diese Teil einer Dialogszene sind oder nicht. Danach findet eine Gruppierung in ganze Dialogszenen statt. Ein dieser Gruppe zugehöriges Verfahren [149] basiert auf einem Multi-Experten System, dass von den folgenden Prämissen bei der Klassifikation ausgeht:

  * Eine Szene besteht aus einer Gruppe semantisch korrelierter Shots.
  * Nahezu alle Shots, die zu einer Dialogszene gehören können als Dialog-Shot charakterisiert werden.
  * Shots, die zu derselben Dialogszene gehören sind zeitlich benachbart.

  Dabei startet das Verfahren mit einer Shot-Boundary Erkennung, bevor drei Expertensysteme (Gesichtserkennung, Kamera Bewegungsvorhersage, Audio Klassifikation) die Shots klassifizieren. Anschließend erfolgt eine Regelbasierte Auswertung, die über eine Zuordnung zu Dialog oder Nicht-Dialog mit Hilfe eines endlichen Automaten entscheidet.

Ein alternatives Verfahren[209] definiert Ereignisse als temporale Objekte, die mit Hilfe von Eigenschaften über verschiedene zeitliche Skalen hinweg charakterisiert werden können. Dabei wird eine „temporale Textur“ einer Bildsequenz bestimmt und die Distanz zwischen zwei Segmenten als normalisierte Differenz der Eigenschaften ihrer temporalen Texturen bestimmt, ohne dabei Vorwissen mit einbeziehen zu müssen.

[144] schlägt ein Verfahren zur Bestimmung von Kandidaten für Ereignisbegrenzungen vor, die auf der Erkennung von Unregelmäßigkeiten in Bewegungsmustern basiert. Eine Sequenz optischer Flussfelder wird näherungsweise bestimmt und die Hintergrundbewegung herausgerechnet. Für jedes Feld werden dominante Bewegungsvektoren ermittelt (Singular Value Decomposition, SVD) und die Trajektorien der SVD-Koeffizienten werden auf Unregelmäßigkeiten untersucht. Die dabei entstandene Menge von Unregelmäßigkeiten ist eine Übermenge zur Menge der tatsächlichen Ereignisgrenzen.

2.3.6 Inhaltliche Abstraktion

Informationstragende audiovisuelle Objekte, wie z.B. Spielfilme oder die Aufzeichnung Sportereignisse, dauern oft mehrere Stunden. Problematisch dabei ist die Tatsache, dass ein Mensch diese Zeit investieren muss, um den gesamten Inhalt des Objekts erkennen und beurteilen zu können. Daher ist die inhaltliche Zusammenfassung informationstragender audiovisueller Objekte von besonderer Bedeutung, da sie eine beschleunigte Wahrnehmung der wichtigsten Inhalte erlaubt. Eine Übersicht über verschiedene Verfahren der inhaltlichen Abstraktion findet sich in [101, 93].

Prinzipiell sind zwei Arten der inhaltlichen Abstraktion denkbar.

- **textuelle Abstraktion**, d.h. textuelle Zusammenfassung des Videoinhalts in natürlicher Sprache. Diese Variante birgt zahlreiche Schwierigkeiten und ist bislang nur in einzelnen, thematisch eng fokussierten Bereichen realisiert [55].
- **visuelle Abstraktion**, d.h. Wiedergabe einer visuellen Zusammenfassung des Videoinhalts. Auf diese Variante soll im Folgenden detaillierter eingegangen werden.

Visuelle Abstraktionen können aus einer Folge von Einzelbildern bestehen (*Still-Image Abstract, Video Summary*), die aus dem zusammenzufassenden Video stammen, oder aber auch aus einem einzelnen Video-Clip (*Moving-Image Abstract, Video Skimming*), der die wichtigsten Ereignisse des Videos wiedergibt und dabei von signifikant kürzerer Dauer ist.

- **Video Summary**:
  Eine Video-Summary lässt sich mit sehr schnell erzeugen, da üblicherweise nur visuelle Informationen dazu verwendet werden und weder Audio- noch Textinformation benötigt wird. Zudem wird bei der Wiedergabe auch keine zusätzliche Synchronisation nötig. Generell dreht sich die Video Summary um die Ermittlung von inhaltlich relevanten Einzelbildern, den Key Frames.

- **Sampling-basierte Key Frame Extraktion** basiert auf zufällig oder gleichförmig gewonnenen Stichproben aus allen aus allen Einzelbildern [116]. Dabei läuft man aber Gefahr, dass kürzere, aber bedeutende Szenen nicht repräsentiert werden, während längere, aber unbedeutende Szenen als Einzelbild in die Video Summary aufgenommen werden. Zur korrekten inhaltlichen Zusammenfassung ist diese Methode nicht geeignet.
- **Shot-basierte Key Frame Extraktion**, orientiert sich dynamisch an der inhaltlichen Zusammensetzung des Videos. Die einfachsten Verfahren extrahieren einfach jeweils den ersten Frame eines Shots als Key Frame [185, 7, 44]. Dynamischer visueller Inhalt lässt sich aber mit einem Einzelbild pro Shot nur unzureichend wiedergeben. Daher werden mehrere Key Frames pro Shot bestimmt, wobei einfache visuelle Charakteristika wie Farbverteilung [211] und Bewegungsaktivität die Wahl der Key Frames festlegen. Bewegungsaktivitäts-basierte Ansätze eignen sich besser zur Bestimmung der relevanten Key Frames, wenn die zeitliche Dynamik eines Videos zusammengefasst werden soll. Zu diesen zählen Berechnung des optischen Fluss [29] und Pixel-basierte EInzelbildvergleiche, aber auch Domänenspezifische, ausgefeilte Verfahren zur Ermittlung globaler Bewegungen und Gesten [80].

  Dennoch sind diese Verfahren nicht dazu geeignet, Key Frames zu ermitteln, die den Inhalt des vollständigen Videos repräsentieren. Mosaik-basierte Verfahren können Panoramabilder aus Einzelbildern erstellen, in denen Kamerabewegungen gut festgehalten werden können [189]. Dazu muss zuerst ein Bewegungsmodell (Translation, Rotation, Skalierung, usw.) über einen Shot hinweg erstellt werden, mit dem die Zusammensetzung des Panoramabildes aus den Einzelbildern berechnet werden kann.

  In [167] wird ein Verfahren zur Zusammenfassung von Video-Datensätzen vorgestellt, das die Trajektorien der dargestellten Objekte mit Hilfe Selbstorganisierender karten (Self-Organizing Maps) analysiert. Dabei werden kritische Punkte der Trajektorie identifiziert, die das Verhalten des betreffenden Objekts innerhalb einer Szene am besten beschreiben, und als Key Frame extrahiert.

  Ein alternatives Verfahren verwendet das im MPEG-7 Standard FDIS festgelegte Attribut „fidelity“ um eine skalierbare hierarchische Zusammenfassung und Suche zu ermöglichen. Als Fidelity wird dabei die Güte bezeichnet, wie gut ein Key Frame innerhalb der Baumstruktur einer Key Frame Hierarchie seine Nachfolgerknoten repräsentiert[84].
- **Segment-basierte Key Frame Extraktion**: Ein entscheidender Nachteil dieser Video Summaries durch das Zusammenfassen eines Shots mit einem oder mehreren Key Frames besteht darin, dass das Verfahren nur schlecht skaliert, d.h. es ist ineffizient und nicht benutzerfreundlich, ein längeres Video durch Hunderte von Key Frames zu repräsentieren. Daher wird versucht, auf einem höheren Abstraktionsniveau – dem Segment als Folge zusammengehöriger Shots, einer Szene oder eines Ereignisses – Key Frames zu gewinnen.

Das in [184] dazu vorgeschlagene Verfahren weist die Key Frames zunächst vordefinierten Clustern zu, die als Entscheidungsgrundlage für eine erneute Segmentierung des Gesamtvideos entsprechend ihrer Relevanz dienen. Segmente, die unter einem festgelegten Relevanz-Schwellwert liegen, werden entfernt, und das Einzelbild, das möglichst nahe der Mitte eines qualifizierten Segments liegt, wird als Key Frame ausgewählt. Die ausgewählten Key Frames werden anschließend noch in einer Bildzusammenfassung aggregiert.

Weitere Ansätze zur Segment-basierten Key Frame Extraktion werden in [175, 66, 217, 37] dargestellt.

- **Video Skimming**:
  Dem gegenüber erscheint Video Skimming als die dem ursprünglichen Medium angemessenere, d.h. ähnlichere Form der Abstraktion, da der Video Summary jegliche Bewegungsinformation fehlt. Üblicherweise unterscheidet man zwei verschiedene Varianten des Video Skimmings [65]: **Summary Sequence**, die dem Betrachter einen Eindruck über das gesamte Video vermitteln soll, und **Highlighting**, das lediglich die interessantesten Ausschnitte des Videos zusammenfasst. Die Beurteilung der Highlights ist ein sehr subjektiver Wahrnehmungsprozess, der nur schwer maschinell abzubilden ist, daher beziehen sich die meisten Ansätze zum Video Skimming auf den Summary Sequence Ansatz..

  Ein sehr einfacher Ansatz komprimiert einfach die zeitliche Dimension und gibt das Video mit schnellerer Geschwindigkeit wieder[124], wobei die zeitliche Komprimierung den Faktor 2.5 auf Grund der Verständlichkeit nicht überschreiten sollte [97].

  Im Informedia Projekt [162] werden Kurzzusammenfassungen informationstragender audiovisueller Objekte durch Extraktion relevanter Audio- und Videoinhalte erzeugt. Dabei werden von einem manuell erzeugten Transkript textuelle Schlüsselwörter ausgewählt und nach dem TF/IDF (Term Frequency/Inverse Document Frequency) Verfahren entsprechend ihrer Relevanz bewertet. Zu den so gewonnenen Schlüsselworten werden die entsprechenden Audio-Segmente ermittelt und mit benachbarten Audiosegmenten zusammen aufgrund besserer Verständlichkeit ausgegeben. Video Skimming wählt anschließend Einzelbilder aus, die

  – Gesichter oder Text beinhalten,
  – nach einer Kamerabewegung ein statisches Bild beinhalten,
  – in einer Kamerabewegung Gesichter oder Text beinhalten,
  – am Anfang einer Video-Szene stehen.

  Einzelbilder werden anschließend mit dem Audio-Extrakt synchronisiert.

  In [179] wird ein multimodales Verfahren vorgestellt, das zuerst erkannte Shots in sogenannte „Story Units“ gruppiert, entsprechend einem erkannten Sprecherwechsel oder Objektwechsel , die innerhalb der Shots oftmals von Texteinblendungen begleitet werden. Zu diesen Story Units werden Audio-Extraktionen er-

stellt und synchronisiert und abschließend über Benutzer-Feedback falls nötig reorganisiert.

Weitere Video Skimming verfahren sind in [146, 111, 110] beschrieben.

2.4 Textuelle Metadaten

Audiovisuelle informationstragende Objekte können per se bereits textuelle Daten beinhalten, die mit Hilfe von Audio- oder Videoanalysetechniken aus diesen extrahiert werden können. Dazu zählen sowohl die Transkribierung sprachlicher Inhalte mit Hilfe von NLP-Technologien als auch die optische Erkennung textueller Information in Videobildern vermittels OCR. Dazu kommen auch autoritative und nichtautoritative, jeweils vom Menschen explizit generierte Metadaten, die ebenfalls in textueller Form vorliegen.

Diese gesammelte textuelle Information kann als Ausgangsdaten für ein Information Retrieval System dienen, um den zielgenauen Zugriff auf die audiovisuellen Daten zu ermöglichen. Generell unterscheidet man verschiedene Analysekategorien bei der Auswertung dieser textuellen Informationen:

- Strukturelle Analyse der textuellen Metadaten
- Inhaltsanalyse der textuellen Metadaten
- Korrelation der textuellen Metadaten

3 Metadaten

3.1 Grundlagen

3.1.1 Begriff und Klassifikation

Metadaten sind strukturierte, kodierte Daten, die Charakteristika informationstragender Entitäten beschreiben, zum Zweck der Identifikation, Recherche, Beurteilung und der Verwaltung der damit beschriebenen Entitäten [42]. Insbesondere steht heute der Aspekt der maschinellen Verarbeitung von Metadaten zum Zweck der Identifikation und der Recherche im Vordergrund. Metadaten können grob klassifiziert werden nach dem **Grad ihrer Strukturierung**.

Strukturierte Metadaten können z.B. durch einfache, vorgegebene Schemata (z.B. Autor, Titel, Verlag, usw.) strukturiert werden. Zusätzlich lassen sich Kategorien und einfache Ober- und Unterklassenbeziehungen zwischen den Kategorien über Taxonomien ausdrücken. Komplexere Strukturen, die etwa auch unterschiedliche Beziehungen, Abhängigkeiten und Regeln zwischen den mit Metadaten ausgezeichneten Objekten zulassen, können mit Hilfe von Ontologien formuliert werden. Gänzlich ohne jede Struktur dagegen ist eine Menge von zugeordneten Schlagworten, die selbst keinerlei Regeln unterworfen sind [35].

Um das gezielte Auffinden audiovisueller, informationstragende Objekte zu ermöglichen, müssen deren Inhalte auf der Basis standardisierter Metadaten beschrieben werden.

3.1.2 Quelle der Metadaten

Das Auszeichnen von informationstragenden Objekten mit Metadaten ist eine nichttriviale Aufgabe, die in der Regel über den Urheber (Autor) des informationstragenden Objektes selbst bzw. durch einen ausgewiesenen Experten zentral erfolgt (**autoritative Metadaten**), aber auch durch den Konsumenten der informationstragenden Objekte, d.h. dem Nutzer erfolgen kann (**nicht-authoritative Metadata**, siehe auch Kap.3.2.3).

Daneben können Metadaten auch mit Hilfe analytischer und statistischer Verfahren direkt aus den informationstragenden Objekten automatisch gewonnen werden. Zwischen automatisch gewonnen Metadaten und Nutzergenerierten Metadaten besteht oft ein qualitativer Unterschied bzgl. ihrer semantischen Ausdruckstärke und Präzision. Diese Lücke zwischen atomatischen Analyseergebnissen und von einem Menschen interpretierten Informationsinhalten wird auch als **semantische Lücke** bezeichnet, die mit Hilfe semantischer Technologien auf dem Bereich der Wissensverarbeitung (Knowledge Engineering) geschlossen werden soll.

3.1.3 Granularität der Metadaten

Die Auszeichnung informationstragender Objekte mit zusätzlichen Metadaten wird als **Annotation** bezeichnet. Die Annotation eines informationstragenden Objekts kann sich auf das Objekt als Ganzes, aber auch auf Teile des Objekts gemäß dessen räumlicher und zeitlicher Differenzierung beziehen. Zu diesem Zweck müssen geeignete Adressierungsmechanismen verwendet werden, um Teilbereiche informationstragender Objekte gezielt mit Metadaten zu annotieren. Eine aktuelle Arbeitsgruppe des W3C[1]beschäftigt sich mit dieser Problematik [132].

3.2 Annotation von audiovisuellen Daten

Im Gegensatz zu textbasierten informationstragenden Objekten spielt bei audiovisuellen Medien insbesondere die zeitliche Dimension, d.h. die zeitliche Dauer bei der Wiedergabe eine besondere Rolle. Üblicherweise müssen sich Aufnahmegeschwindigkeit und Wiedergabegeschwindigkeit audiovisueller Medien entsprechen, da sonst ihr Informationsinhalt vom Nutzer nicht korrekt interpretiert werden kann.

Zerlegt man das informationstragende Objekt vor der Annotation in inhaltlich kohärente Segmente, können diese Segmente jeweils wieder als Ganzes mit Metadaten annotiert werden. Andererseits können Metadaten auch mit einer zeitlichen Koordinate versehen werden, die einen Bezug zu einem bestimmten Segment des informationstragenden Objekts herstellt.

3.2.1 Annotation informationstragender Objekte als Ganzes

Metadaten, die sich auf ein informationstragendes Objekt als Ganzes beziehen erlauben keine weitere zeitliche Differenzierung und Zuordnung zu einem bestimmten Teilsegment des informationstragenden Objekts. Zu dieser Art von Metadaten zählen bei audiovisuellen informationstragenden Objekten unter anderem

- technische Parameter (z.B. Dimensionsangaben, Aufzeichnungs- und Wiedergabegeschwindigkeit, Angaben über technische Parameter des Aufnahmegeräts)
- rechtliche Parameter (z.B. Urheberschaft, Einschränkungen bzgl. Wiedergabe, Verbreitung oder Weiterverarbeitung)
- aggregierte inhaltliche Informationen (z.B. inhaltliche Zusammenfassung, repräsentierte Personen, Objekte, Orte, etc.)

3.2.2 Isochrone Annotation

Als isochrone Annotation audiovisueller informationstragender Objekte bezeichnet man die Verknüpfung von Metadaten mit einer zeitlichen Positionsangabe, um so eine exakte zeitliche Zuordnung der Metadaten zu ermöglichen. Isochrone Annotation kann zu verschiedenen Zwecken in unterschiedlichem Kontext erfolgen:

[1]World Wide Web Consortium, http://w3c.org

- isochrone Transkribierung gesprochener Texte und Dialoge,
- inhaltliche Strukturierung audiovisueller informationstragender Objekte,
- Identifikation und Annotation von im zeitlichen Verlauf auftauchender Objekte, Umgebungen, Kontext und Pragmatik.

3.2.3 Nicht-autoritative Annotation

Unter nicht-autoritativer Annotation versteht man die Auszeichnung informationstragender Objekte durch deren Benutzer. Nicht-autoritative Annotation dient einerseits der Personalisierung von Metadaten und ermöglicht damit auch einen personalisierten Zugang zu einem großen Datenbestand informationstragender Objekte.

Fasst man die nicht-autoritativen Annotationen verschiedener Benutzer zusammen ergibt sich eine Form der **kollaborativen Annotation** (Folksonomie) [114]. Diese kommt derzeit in zahlreichen Web 2.0 Anwendungen zum Einsatz: Benutzer können über ein Webportal informationstragende Objekte mit eigenen Metadaten – üblicherweise einfache Schlagworte (Tags) – auszeichnen. Das Webportal aggregiert diese Metadaten und ermöglicht damit personalisierte Suche und Zugang zum erfassten Gesamtdatenbestend. Mehr noch können über das soziale Beziehungsgeflecht der Benutzer untereinander (Social Networks) ähnlichkeitsbasierte Suchen oder auch Empfehlungen (Recommendations) ermöglicht werden [60]. Bei audiovisuellen Daten stellt insbesondere die kollaborative Annotation eine wichtige Informationsquelle zur Beschaffung von Metadaten dar.

Aktuelle Videoportale und Suchmaschinen erlauben die kollaborative Annotation von audiovisuellen Daten. Zu diesen zählen u.a.:

- **youTube**[2]: Aktuell das populärste Webportal zur kollaborativen Annotation von Videodaten. Üblicherweise erfolgt eine Annotation der Daten als Ganzes, da die Mehrzahl der dort registrierten Videodaten nur von kurzer Dauer sind. Benutzer haben zudem die Möglichkeit, einzelne Videos zu bewerten. Experimentell wird auch die isochrone Annotation bzw. die Annotation räumlicher Teilbereiche (spatiale Annotation) ermöglicht.
- **Google Video**[3]: Ähnliche Organisation und Arbeitsabläufe wie bei youtube.com. Google experimentiert ebenfalls mit der Möglichkeit einer isochronen Annotation.
- **yovisto**[4]: Spezialisiertes Webportal mit dem Schwerpunkt audiovisueller Lehr- und Lernmaterialien. Ermöglicht isochrone autoritative und nicht-autoritative Annotation. Im Gegensatz zu youTube und Google Video werden die audiovisuellen Inhalte nicht auf einem zentralen Server zur Verfügung gestellt, sondern verbleiben bei deren Urhebern [147].

[2] http://www.youtube.com
[3] http://video.google.com
[4] http://www.yovisto.com

3.3 Metadaten-Standards für audiovisuelle Daten

3.3.1 Metadaten-Standards

Eine der wichtigsten Forderungen bei der Definition von Metadaten besteht in deren **Interoperabilität** [161]. Darunter versteht man die Austauschbarkeit von Metadaten, die aus unterschiedlichen Quellen stammen, und deren gemeinsame Verarbeitung. Zu diesem Zweck wurden vielfältige Metadaten-Standards definiert, die eine interoperable Verarbeitung von Metadaten ermöglichen sollen. Hinter einem Metadaten-Standard steht ein allgemein akzeptiertes Datenmodell, das die Syntax der verwendeten Datenstrukturen und deren Bedeutung (Semantik) beschreibt.

Für audiovisuelle Daten besitzenvor allem die Standards *Dublin Core*, *MPEG-7* und *MPEG-21* eine besondere Bedeutung.

3.3.2 Dublin Core und Erweiterungen

Der Dublin Core Metadaten-Standard[5] wurde speziell zur Beschreibung textueller, informationstragender Objekte geschaffen. Die 15 Dublin Core Kernelemente[6] dienen der Erfassung bibliographischer Daten [198]. Die DCMI Metadata Terms empfehlen zusätzliche Felder sowie detaillierende Felder (Element Refinements), die eine auf speziellere Bedürfnisse zugeschnittene Beschreibung bzw. Kategorisierung erlauben. Dabei werden Metadaten zur technischen und inhaltlichen Beschreibung, bzgl. beteiligter Personen und Urheberrechte, sowie der Vernetzung der Ressourcen und des Lebenszyklus definiert. Der Dublin Core Metadatenstandard zielt nicht speziell auf audiovisuelle Objekte ab, kann aber für solche verwendet werden [73]. Allerdings werden damit lediglich kontextuelle Metadaten festgelegt, wie sie im Bibliotheksumfeld üblich sind.

3.3.3 MPEG-7

Der MPEG-7 Metadaten-Standard (ISO/IEC 15938) definiert eine umfangreiche Sammlung von Datenstrukturen zur inhaltlichen und strukturellen Beschreibung von audiovisuellen, informationstragenden Objekten [28]. Dazu zählen:

- Deskriptoren zur Beschreibung von auditiven und visuellen Charakteristika auf unterschiedlichen Abstraktionsebenen,
- Beschreibungsschemata für Multimediadaten,
- Spezifikationen und Kodierungsvorschriften für den Transport von MPEG-7 Metadaten

MPEG-7 basiert auf der semistrukturierten Beschreibungssprache XML und definiert mehr als 450 Metadaten Datentypen. Damit stellt MPEG-7 den inhaltlich reichhaltigsten Metadatenstandard für multimediale Daten dar. Es existieren Schnittstellendefinitionen und Mappings zu zahlreichen anderen Metadaten-Standards, wie z.B. Dublin

[5] Dublin Core Metadata Initiative (DCMI), `http://dublincore.org/`
[6] Dublin Core Metadata Element Set, Version 1.1 (ISO 15836)

Core, TV-Anytime, P/Meta oder Descriptive Metadata Scheme-1 (DMS-1). Problematisch jedoch bleibt die einheitliche Interpretation und Nutzung vieler Deskriptoren des MPEG-7 Standards.

DMS-1

DMS-1 (auch SMPTE S380M)[7] dient der Definition von Metadaten für den Austausch von audiovisuellen Objekten in sogenannten Material Exchange Format (MXF) Dateien. DMS-1 Metadaten enthalten textuelle Informationen über Personen und Schauspieler, Urheberrechte und weitere Produktionsinformationen.

P/Meta

P/Meta definiert semantische Metadaten speziell aus dem Blickwinkel des Medien-Distributors heraus, stellt aber kein eigenständiges Metadatenschema zur Beschreibung von multimedialen Inhalten dar[8].

TV-Anytime

Der Metadatenstandard TV-Anytime[9] zielt auf Anwendungen im Bereich des elektronischen Endverbraucher-Massenmarktes ab, wie z.B. personalisierte Videorecorder (PVR). Dabei werden Informationen aus elektronischen Programmzeitschriften (Electronic Programing Guide, EPG) über ein einheitliches Metadatenschema modelliert und standardisiert ausgetauscht. Bestandteile sind Urheberrechtsinformationen, Nutzungsinformationen, Benutzerpräferenzen und andere demographische Informationen [134].

IPTC

Der IPTC Metadatenstandard (International Press Telecommunications Council)[10] zielt auf die Annotation genrespezifischer Anwendungsbereiche, wie z.B. Nachrichten (NesML), Sportberichte (SportsML) oder Radio- bzw. TV-Programminformationen (ProgramGuideML) ab. Das IPTC entwickelte und verwaltet diesen Metadatenstandard für den vereinfachten Austausch von Nachrichten.

3.3.4 MPEG 21

Der MPEG-21 Metadaten-Standard (ISO/IEC 21000) definiert ein Rahmenwerk zur Beschreibung des Austauschs digitaler Objekte [23]. Dazu zählen sowohl die Beschreibung von urheberrechtlichen Bestimmungen, Bestimmungen über Vertrieb und Weiterverarbeitung von digitalen Objekten, als auch die Anpassung an unterschiedliche Nutzungs-Infrastrukturen.

[7] Society of Motion Picture Television Engineers
[8] EBU Project Group P/Meta - Metadata Exchange Schema, `http://www.ebu.ch/en/technical/trev/trev_290-hopper.html`
[9] `http://www.tv-anytime.org/`
[10] `http://www.iptc.org/`

3.3.5 Bewertung

Die Spannbreite aller denkbaren Anwendungen, die mit der Produktion, Distribution und Nutzung informationstragender, audiovisueller Objekte verbunden sind, erschwert die Beschränkung auf einen einzigen Metadatenstandard für alle Anwendungszwecke [73, 188, 120]. Jeder der aufgelisteten Standards zielt auf eines oder mehrere bestimmte Anwendungsgebiete ab:

- industrielle Anwendungen (Musik, Spielfilme, Fernsehen und Unternehmen)
- Prozesse und Arbeitsabläufe (Erzeugung, Produktion, Management und Distribution)
- Inhaltstypen (Bilder, Video, Audio, Grafik, Text)
- Genres (Nachrichtem, Sport, Unterhaltung)

So zielen TV-Anytime oder DMS-1 hauptsächlich auf die Verwendung im Rahmen der Fernseh- und Unterhaltungsindustrie. MPEG-7 und MPEG-21 sind generische Standards, d.h. nicht anwendungsspezifisch und bieten damit noch die größte Chance einer übergreifenden Anwendung, wobei der Bereich der Arbeitsabläufe nur unzureichend repräsentiert wird. P/Meta und DMS-1 zielen speziell auf Produktionsabläufe ab, während P/Meta zusätzlich noch den Bereich Management abdeckt. TV-Anytime und MPEG-21 zielen auf die Bereiche Distribution, Verbreitung und Nutzung.

3.4 Semantische Annotation

Soll eine inhaltliche Annotation informationstragender audiovisueller Objekte auf einer höheren Abstraktionsebene erfolgen, geschieht diese oft mit Hilfe von textueller Information in natürlicher Sprache. Die Interpretation dieser sprachlichen Annotation gelingt in der Regel nur dem Menschen und ist nicht allgemein maschinell weiterzuverarbeiten und zu verstehen. Um eine maschinenlesbare, inhaltsbezogene Annotation zu realisieren, müssen formal definierte Wissensrepräsentationsformen eingesetzt werden.

Konzepte, Beziehungen zwischen Konzepten, Individuen und Regeln können mit Hilfe von Ontologien definiert werden, die in formalen Beschreibungssprachen von unterschiedlicher semantischer Expressivität formuliert und ausgetauscht werden können.

Im Zuge der Semantic Web Initiative des W3C[11] wurden verschiedenartige semantische Beschreibungssprachen definiert, die maschinell weiterverarbeitet werden können.

3.4.1 Wissensrepräsentationen und Ontologien

Ein Problem der Interpretation von Low-Level Deskriptoren audiovisueller Daten besteht darin, aus diesen Deskriptoren eine inhaltliche Beschreibung auf einer höheren Abstraktionsebene herzuleiten. Diese Lücke zwischen Low-Level Deskriptoren und ihrer inhaltlichen (und kontextbezogenen) Interpretation wird auch als „Semantische Lücke“ (Semantic Gap) bezeichnet.

[11] W3C Semantic Web Activity Group, `http://www.w3.org/2001/sw/`

Inhaltliche Konzepte können in einer maschinenlesbaren, standardisierten Form als Ontologie definiert werden. Ontologien sind nach Gruber [62] eine formale, allgemein akzeptierte, Spezifikation einer Konzeption, d.h. ein abstraktes Modell, in dem alle Begriffe explizit definiert werden, über das Konsens besteht und das maschinell verarbeitet und verstanden werden kann.

Das W3C hat bereits die dazu notwendigen Grundlagen in Form von Ontologiebeschreibungssprachen, wie z.B. RDF, RDFS oder OWL geschaffen. Semantisch annotierte Objekte ermöglichen es autonom agierenden Agenten zielgerichtet Informationen zu sammeln, um selbstständig Entscheidungen im Sinne ihres Auftraggebers zu treffen und Transaktionen zu initiieren. Dieses semantische Netzwerk (**Semantic Web**) stellt die nächste Evolutionsstufe des WWW dar [13].

3.4.2 Semantische Annotation für audiovisuelle Daten

Das MPEG-7 Metadatenschema gibt eine fest definierte Datenstruktur vor, die mit individuellen Ausprägungen der darin vordefinierten Konzepte gefüllt werden kann. MPEG-7 basiert auf der allgemeinen Markupdefinitionssprache XML (Extensible Markup Language) und definiert damit ein allgemein akzeptiertes, statisches Austauschdatenformat für Metadaten.

Um allerdings den Inhalt von High-Level Deskriptoren formal und in maschinenverstehbarer Form beschreiben zu können, reicht die in MPEG-7 inhärente Semantik nicht aus. Zur Definition einzelner Konzepte und deren Beziehungen untereinander wird die vom W3C standardisierte Resource Description Language (RDF) verwendet.

RDF und RDF Schema

RDF (Resource Description Framework) und RDFS definieren einfache Wissensrepräsentationen, in denen Individuen und deren Beziehungen untereinander (RDF) bzw. neue Konzepte und deren Beziehungsstrukturen definiert werden können [95] . Individuen und Konzepte werden dabei über URIs (Uniform Resource Identifier), d.h. eine Adresse über die sie zugreifbar sind, eindeutig identifiziert.

Eine RDF Datei besteht aus einzelnen Tripeln (a,b,c), wobei a für ein Individuum steht, b für eine Eigenschaft dieses Individuums und c für einen bestimmten Wert, der dem Individuum a für die Eigenschaft b zugewiesen wird. Auf diese Weise können komplexe Beziehungsgeflechte zwischen Individuen ausgedrückt werden.

Indiviuduen sind konkrete Realisierungen (Instanzen) von Konzepten. Konzepte können von übergeordneten Konzepten mit Hilfe der RDF Schema (RDFS) Beschreibungssprache hergeleitet werden über Generalisierung, Spezialisierung oder Erweiterung [21]. Auf diese Weise können Beziehungen zwischen grundlegenden Konzepten modelliert werden, auch wenn diese nicht von Anfang an Bestandteil eines vorgegebenen Metadatenschemas sind.

Allerdings ist die semantische Ausdrucksstäre von RDF und RDFS beschränkt, so dass für die Verallgemeinerung von Aussagen für eine Menge von Individuen bzw. für die Definition logischer Attribute und Beziehungen komplexere Wissensrepräsentationen in Form von Beschreibungslogiken [8] herangezogen werden, die über die Beschrei-

bungssprachen OWL (Web Ontology Language) [129] oder RIF (Rule Interchange Format)[12] implementiert werden.

3.5 Metadatenablage und effizienter Zugriff

Besondere Berücksichtigung muss die Archivierung und Verwaltung der gewonnenen Metadaten finden, um einen möglichst effizienten, zielgenauen und schnellen Zugriff für die damit beabsichtigte Verarbeitung zu gewährleisten. Unterschiedliche Varianten kommen dabei zum Einsatz:

- **strukturierte Dateien**: ineffiziente Speichermöglichkeit, die nur bei relativ kleinen Datenmengen Sinn macht, anfällig für Inkonsistenzen, nicht skalierbar.
- **relationale Datenbanken**: effizient und skalierbar, problematisch bei Erweiterungen der verwendeten Datenschemata, keine Möglichkeit der Vererbung.
- **objektorientierte** und **objektrelationale Datenbanken**: effizient und skalierbar, problematisch bei Erweiterungen der verwendeten Datenschemata.
- **semistrukturierte Datenbanken**: effizient, aber nur bedingt skalierbar.

[12] `http://www.w3.org/2005/rules/wiki/RIF_Working_Group`

4 Processing

Den Bereich des Processing betrifft die geordnete Verarbeitung und damit verbundene Auswertungsmöglichkeiten, der in den vorangegangenen Verarbeitungsschritten gewonnenen Metadaten. Audiovisuelle Informationstragende Objekte werden erst durch ausreichende Beschreibung und Annotation in Form von Metadaten für den wahlfreien Zugriff geöffnet. Insbesondere in Bibliotheken und Archiven, die sehr große Datenmengen bevorraten, ist ein gezielter Zugriff bzw. eine möglichst **inhaltsbasierte Suche** von größter Bedeutung.

4.1 Suche in audiovisuellen Daten

Informationsbeschaffung (Information Retrieval), d.h. die Suche in audiovisuellen Daten basiert auf den aus diesen gewonnenen analytischen Metadaten und manuell vom Autor oder Benutzer bereitgestellte Metadaten in textueller Form [10]. Neben der traditionellen Suche nach **Schlüsselwörtern** (Keywords), die sich am Text Information Retrieval orientiert, sind speziell im Bereich audiovisueller Daten auch folgende Suchvarianten möglich:

- **Ähnlichkeitsbasierte Suche**, zu einem vorgegebenen Beispiel (Example) sollen möglichst ähnliche Objekte gefunden werden (Query by Example). Die Ähnlichkeitsbestimmung kann dabei sowohl nach Low-Level Deskriptoren (Farb- oder Helligkeitsverteilung in Einzelbildern, Bewegungsaktivität in einzelnen Segmenten, Intensitätsverlauf und Pausen bei Audiodaten, etc.) als auch nach High-Level Deskriptoren auf höheren Abstraktionsebenen (Genrezugehörigkeit, affektiver Inhalt, etc.) erfolgen.

- **Data Mining**, die systematische Anwendung von statistischen Methoden auf einen Datenbestand von audiovisuellen Objekten mit dem Ziel der Mustererkennung, d.h. das Aufspüren von Regeln und statistischen Auffälligkeiten. Ziel ist dabei das Erkennen von vorliegenden Sachverhalten, die sich aus der Einzelbetrachtung nicht erschließen.

- **Assoziative Suche**, hier orientiert sich die Vernetzung der vorliegenden Datenobjekte an Beziehungen zwischen den einzelnen Datenobjekten, die entsprechend dem vorliegenden Umfeld (Kontext) und der intendierten Absicht (Pragmatik) eine Relevanzbewertung erfahren und es so dem Benutzer ermöglichen, neue Querverweise (Assoziationen) zwischen zuvor nicht zusammenhängenden Datenobjekten ziehen zu können.

4.2 Visualisierung audiovisueller Daten

Traditionelle Information Retrieval Systeme präsentierem dem Benutzer ihre Ergebnisse oft in Form von Antwortlisten. Liegen dies in Textform vor (z.B. Titel und Adresse eines Dokuments, Textausschnitt, Snippet), kann der Benutzer schnell auf den Inhalt des jeweiligen Suchergebnisses schließen, ohne dieses explizit zugreifen zu müssen. Dies erleichtert die Auswahl des interessantesten Ergebnisses aus einer Ergebnisliste.

Der Inhalt audiovisueller informationstragender Objekte lässt sich nur schwer auf automatische Weise in Textform zusammenfassen. Grundlage für eine textuelle Zusammenfassung und der Präsentation einer textuellen Suchergebnisliste sind die vorliegenden Metadaten. Diese geben aber oft nur einen unzureichenden Eindruck von deren Inhalt wieder.

Video Summaries und Video Skimming können hier zum Einsatz kommen und stellvertretend für die Originale lediglich eine aggregierte Zusammenfassung (z.B. eine Folge relevanter Key Frames) darstellen. Insbesondere, wenn eine große Menge audiovisueller Objekte vorliegt, aus der ein Benutzer visuell das tatsächlich passende aussuchen soll, müssen geeignete Visualisierungen gefunden werden, die

- eine schnelle visuelle Auffassung des Inhalts ermöglichen und
- eine direkte Vergleichbarkeit verschiedener audiovisueller Objekte und damit schnelle Entscheidung über deren Relevanz ermöglichen.

4.3 Navigation in großen Datenmengen

Eine effiziente Visualisierung informationstragender audiovisueller Objekte eröffnet Möglichkeiten der Navigation durch einen großen Datenbestand. Basierend auf assoziativen Verknüpfungen oder der Ähnlichkeit zweier Objekte können Nachbarschaftsrelationen ermittelt werden und eine Navigation entlang der dadurch entstehenden Baumstruktur ermöglichen.

Inhaltlich ähnliche Objekte können über Clustering-Verfahren in unterschiedliche Gruppen eingeordnet werden, deren Thema durch einen repräsentativen Stellvertreter visualisiert werden kann.

BITTE KAPITEL ERGÄNZEN .....

5 Anhang

5.1 Glossar und Akronyme

BITTE ERGÄNZEN .....

Literaturverzeichnis

[1] The brava broadcast archive programme impairments dictionary, available at `http://brava.ina.fr/brava_public_impairments_list.en.html`, 2007.

[2] Nicola Adami, Riccardo Leonardi, and Pierangelo Migliorati. Overview of multimodal techniques for the characterization of sport programs. In *Proc. Conf. Visual Communications and Image Processing, Lugano, Switzerland*, pages 1296–1306, 2003.

[3] Kartik K. Agaram, Stephen W. Keckler, and Doug Burger. Characterizing the sphinx speech recognition system. Technical report, University of Texas at Austin, Austin, TX, USA, 2001.

[4] A. Aydin Alatan. Automatic multi-modal dialogue scene indexing. In *Proc. of IEEE Int. Conf. on Image Processing, Thessaloniki, Greece*, pages 374–377, 2002.

[5] A. Aydin Alatan, Ali N. Akansu, and Wayne Wolf. Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools Appl.*, 14(2):137–151, 2001.

[6] Arnon Amir, Janne O. Argill, Marco Berg, Shih Fu Chang, Martin Franz, Winston Hsu, Giridharan Iyengar, John R. Kender, Lyndon Kennedy, Ching Yung Lin, Milind Naphade, Apostol Natsev, John R. Smith, Jelena Tesic, Gang Wu, Donqing Zhang, and IBM T. J. Watson. Ibm research trecvid-2004 video retrieval system. In *In Proc. of TREC Video Retrieval Evaluation*, 2004.

[7] Farshid Arman, Remi Depommier, Arding Hsu, and Ming-Yee Chiu. Content-based browsing of video sequences. In *ACM Multimedia*, pages 97–103, 1994.

[8] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, 2003.

[9] N. Babaguchi, N. Nitta, and T. Kitahashi. Story based representation for broadcasted sports video and automatic story segmentation. In *Proc. IEEE International Conference on Multimedia and Expo(ICME 2002)*, pages pp. 813–816, 2002.

[10] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[11] M. Barlaud. State of the art in content-based analysis, indexing and retrieval. Technical report, IST-2001-32795 D2.1, 2002.

[12] Tim Berners-Lee. Semantic web roadmap. On-line draft `www.w3.org/DesignIssues/Semantic.html`, 1998.

[13] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[14] Marco Bertini, Alberto Del Bimbo, Rita Cucchiara, and Andrea Prati. Semantic video adaptation based on automatic annotation of sport videos. In *ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2004, New York, NY, USA*, pages 291–298, 2004.

[15] Bernard Besserer and Cedric Thire. Detection and tracking scheme for line scratch removal in an image sequence. In Tomas Pajdla and Jiri Matas, editors, *8th European Conference on Computer Vision (ECCV)*, volume 3023 of *Lecture Notes in Computer Science*, pages 264–275. Springer, 2004.

[16] Bodo Billerbeck, Adam Cannane, Abhijit Chattaraj, Nicholas Lester, William Webber, Hugh E. Williams, John Yiannis, and Justin Zobel. Rmit university at trec 2004. In *Proceedings Text Retrieval Conference (TREC), Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication*, 2004.

[17] L. Boch. Tv programmes acquisition system for rai multimedia catalogue. *Elettronica e Telecomunicazioni*, pages 38–47, 2000.

[18] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *Storage and Retrieval for Still Image and Video Databases IV*, pages 170–179, 1996.

[19] A. Whitehead P. Bose and R. Laganiere. Feature based cut detection with automatic threshold selection. *Proc Intl Conf Image and Video Retrieval*, pages 411–418, 2004.

[20] A. C. Bovik and Shizhong Liu. Dct-domain blind measurement of blocking artifacts in dct-coded images. In *ICASSP '01: Proceedings of the Acoustics, Speech, and Signal Processing, 2001. on IEEE International Conference*, pages 1725–1728, Washington, DC, USA, 2001. IEEE Computer Society.

[21] Dan Brickley and R.V. Guha. Resource description framework (RDF) schema specification, February 2004.

[22] O. Buisson, S. Boukir, and B. Besserer. Motion compensated film restoration. *Machine Vision and Applications*, 13(4):201–212, 2003.

[23] Ian Burnett, Fernando Pereira, Rik Van de Walle, and Rob Koenen. *The MPEG-21 Book*. John Wiley & Sons, April 2006.

[24] G. Campra. Digital processing of television signals for multimedia cataloguing. *Tesi di Laurea*, 1998.

[25] Jorge Caviedes, Antoine Drouot, Arnaud Gesnot, and Laurent Rouvellou. Impairment metrics for digital video and their role in objective quality assessment. *Visual Communications and Image Processing 2000*, 4067(1):791–800, 2000.

[26] Lekha Chaisorn and Tat-Seng Chua. The segmentation and classification of story boundaries in news video. In *Proceedings of the IFIP TC2/WG2.6 Sixth Working Conference on Visual Database Systems*, pages 95–109, Deventer, The Netherlands, The Netherlands, 2002. Kluwer, B.V.

[27] M. Chambah, C. Saint-Jean, and F. Helt. Image quality evaluation in the field of digital film restoration. In R. Rasmussen and Y. Miyake, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5668 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 220–231, October 2004.

[28] S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 Standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

[29] Wing-San Chau, Oscar C. Au, and Tak-Song Chong. Key frame selection by macroblock type and motion vector analysis. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, Taipei, Taiwan*, pages 575–578, 2004.

[30] Lei Chen, Shariq J. Rizvi, and M. Tamer. Incorporating audio cues into dialog and action scene extraction. In *In Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, pages 252–264, 2003.

[31] M. Chen and J. Yang. Feature Extraction Techniques. CMU at TRECVID 2004, 2004.

[32] Colorado State University. Evaluation of face recognition alogrithms, available at http://www.cs.colostate.edu/evalfacerec/, 2008.

[33] Francois-Xavier Coudoux, Marc Georges Gazalet, Christian Derviaux, and Patrick Corlay. Picture quality measurement based on block visibility in discrete cosine transform coded video sequences. *Journal of Electronic Imaging*, 10(2):498–510, 2001.

[34] R. B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *In Proc. Int. Computer Music Conf. (ICMC)*, pages pages 344–347, Thessaloniki, Greece, 1997.

[35] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation. *AI Magazine*, 14(1):17–33, 1993.

[36] M. Davy and S.J. Godsill. Audio information retrieval: A bibliographical study. Technical report, Camebridge University Engineering Department, 2002.

[37] Daniel DeMenthon, Vikrant Kobla, and David S. Doermann. Video summarization by curve simplification. In *ACM Multimedia*, pages 211–218, 1998.

[38] Y. Deng, B.S. Manjunath, and H. Shin. Colour image segmentation. proc. ieee conf. on computer vision and pattern recognition. *IEEE Transactions on Data and Knowledge Engineering*, pages 446–510, June 1999.

[39] Nevenka Dimitrova. Multimedia content analysis: The next wave. In *Proc. of International Conference on Image and Video Retrieval, Urbana-Champaign, IL, USA*, pages 9–18, 2003.

[40] Andres Dorado, Divna Djordjevic, Ebroul Izquierdo, and Witold Pedrycz. Supervised semantic scene classification based on low-level clustering and relevance feedback. In *Proceedings of the European Workshop for the Integration of Knwoledge, Semantics and Digital Media Technology,London, UK*, 2004.

[41] F. Dufaux and Fabrice Moscheni. Segmentation-based motion estimation for second generation video coding techniques. *Video Coding: The Second Generation Approach*, pages 219–263, 1996.

[42] William R. Durrell. *Data Administration: A Practical Guide to Data Administration.* McGraw-Hill, 1985.

[43] D. Eichmann and D. J. Park. Boundary and feature extraction at the university of iowa. In *Proc. TRECVID workshop 2004*, 2004.

[44] Paul England, Robert B. Allen, Mark Sullivan, Andrew Heybey, Mike Bianchi, and Apostolos Dailianas. I/browse: The bellcore video library toolkit. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 254–264, 1996.

[45] S. Esmaili, S. Krishnan, and K. Raahemifar. Content based audio classification and retrieval using joint time-frequency analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, 2004. IEEE.

[46] S. Ezekiel and J. A. Cross. Fractal-based texture analysis. *ACM SIGCSE*, 2002.

[47] M. Q. Farras, J. M. Foley, and S. K. Mitra. Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing synthetic artefacts. In *In: Human Vision and Electronic Imaging X, Proc. SPIE 5666*, 2005.

[48] Emmanuel Ferragne and François Pellegrino. Automatic dialect identification: A study of british english. In *Speaker Classification (2)*, pages 243–257, 2007.

[49] Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. In *ACM Multimedia*, pages 295–304, 1995.

[50] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics: Principles and Practice. Systems Programming Series*. Addison-Wesley, 1990.

[51] Jonathan T. Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, pages 138–147, 1997.

[52] Robert Frischholz. The face detection homepage, `http://www.facedetection.com/`, 2007.

[53] K. Fushikida, Y. Hiwatari, and H. Waki. A content-based video retrieval method using a visualized sound pattern. In *Visual Database Systems 4 (VDB4)*, pages 208–213, 1998.

[54] M. J. F. Gales, Do Yeong Kim, Philip C. Woodland, Ho Yin Chan, D. Mrva, R. Sinha, and S. E. Tranter. Progress in the cu-htk broadcast news transcription system. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1513–1525, 2006.

[55] Ralf Gerber, Hans-Hellmut Nagel, and Heiko Schreiber. Deriving textual descriptions of road traffic queues from video sequences. In *Proceedings of the 15th Eureopean Conference on Artificial Intelligence, ECAI'2002, Lyon, France*, pages 736–740, 2002.

[56] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming: Musical information retrieval in an audio database. In *ACM Multimedia*, pages 231–236, 1995.

[57] Wayn J. Gillespie and D. T. Nguyen. Video classification using a tree-based rbf network. In *Int. Conf. on Image Processing*, volume 3, pages 465–468, 2005.

[58] H. Gish, M. Siu, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, Toronto, Canada, 1991.

[59] King Shy Goh, Koji Miyahara, Regunathan Radhakrishan, Ziyou Xiong, and Ajay Divakaran. Audio-visual event detection based on mining of semantic audio-visual labels. In *in Proc. Conf. on Storage and Retrieval for Multimedia*, 2004.

[60] Scott Golder and Bernardo A. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208, 2006.

[61] J. Große. Speicherverfahren und Werkzeuge für RDF/S. *XML Clearinghouse Report*, 6, 2002.

[62] T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.

[63] N. Guyader, L. B. Herve, J. Herault, and A. Guerin. Towards the introduction of human perception in a natural scene classification system. In *Proc. of IEEE Workshop on Neural Networks for Signal Processing*, 2002.

[64] A. Hanjalic. *Content-based analysis of digital video.* Kluwer Academic Publishers, 2004.

[65] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(4):580–588, 1999.

[66] Alan Hanjalic, Marco Ceccarelli, Reginald L. Lagendijk, and Jan Biemond. Automation of systems enabling search on stored video data. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 427–438, 1997.

[67] Alexander G. Hauptmann and Michael J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *Proceedings of Advances in Digital Libraries Conference*, pages 168–179, 1998.

[68] Erik Hjelmås and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[69] K. Hoashi, M. Sugano, M. Naito, K. Matsumoto, F. Sugaya, and Y. Nakajima. Shot boundary determination on mpeg compressed domain and story segmentation. In *Proc. TRECVID workshop 2004*, 2004.

[70] Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar. Discovery and fusion of salient multi-modal features towards news story segmentation. In *Proc. Storage and Retrieval Methods and Applications for Multimedia*, page pp. 244 258, 2004.

[71] Jing Huang, Etienne Marcheret, Karthik Visweswariah, Vit Libal, and Gerasimos Potamianos. The ibm rich transcription 2007 speech-to-text systems for lecture meetings. In *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 429–441, 2007.

[72] R. W. G. Hunt. *The Reproduction of Colour in Photography, Printing and Television.* Fountain Press, 1987.

[73] Jane Hunter and Liz Armstrong. A comparison of schemas for video metadata representation. *Comput. Netw.*, 31(11-16):1431–1451, 1999.

[74] T. I. Ianeva, A. P. de Vries, and H. Röhrig. Detecting cartoons: A case study in video-genre classification. In *in Proceedings ICME Multimedia and Expo*, volume 1, pages pp. 449–452, 2003.

[75] International Organistation for Standardisation. MPEG-7 Overview, 2004.

[76] Sung Ho Jin, Tae Meon Bae, and Yong Man Ro. Automatic video genre detection for content-based authoring. In *Proc. Storage and Retrieval for Multimedia, San Jose, CA, USA*, pages 335–343, 2004.

[77] S. E. Johnson and P. C. Woodland. Speaker clustering using direct maximisation of the mllr-adapted likelihood. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[78] Laurent Joyeux, Samia Boukir, and Bernard Besserer. Film line scratch removal using kalman filtering and bayesian restoration. In *5th IEEE Workshop on Applications of Computer Vision*, page pp. 8, 2000.

[79] Werner Bailer (JRS), Franz Höller (JRS), Alberto Messina (RAI), Daniele Airola (RAI), Peter Schallauer (JRS), and Michael Hausenblas (JRS). State of the art of content analysis tools for video, audio and speech. Technical report, FP6-IST-507336 PrestoSpace Deliverable, Johanneum Research, Graz (Austria), 2005.

[80] Shanon X. Ju, Michael J. Black, Scott L. Minneman, and Don Kimber. Analysis of gesture and action in technical talks for video indexing. In *Conference on Computer Vision and Pattern Recognition (CVPR '97), San Juan, Puerto Rico*, pages 595–, 1997.

[81] Yu-Lin Kang, Joo-Hwee Lim, Qi Tian, Mohan S. Kankanhalli, and Changsheng Xu. Visual keywords labeling in soccer video. In *Proc. of International Conference on Pattern Recognition, Surrey, UK*, pages 850–853, 2004.

[82] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice Hall, 1997.

[83] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:1423–1426, 2000.

[84] Jung-Rim Kim, Seong Soo Chun, Seok Jin Oh, and Sanghoon Sull. Scalable hierarchical summarization of news using fidelity in mpeg-7 description scheme. In *Recent Advances in Visual Information Systems, 5th International Conference, VISUAL 2002 Hsin Chu, Taiwan*, pages 239–246, 2002.

[85] Don Kimber and Lynn Wilcox. Acoustic segmentation for audio browsers. In *In Proc. Interface Conference*, pages 9–16, 1996.

[86] Tomi Kinnunen, Evgeny Karpov, and Pasi Fränti. Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech & Language Processing*, 14(1):277–288, 2006.

[87] T. Kohonen. *Self-Organizating Maps.* Springer-Verlag, NY, 1997.

[88] Anil C. Kokaram. *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video.* Springer-Verlag, London, UK, 1998.

[89] I. Koprinska and S. Carrato. Temporal video segmentation: A survey captions on mpeg compressed video. *DSignal Processing Image Communication*, pages 477–500, 2001.

[90] Margarita Kotti, Luis P. M. Martins, Emmanouil Benetos, Jaime S. Cardoso, and Constantine Kotropoulos. Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006*, pages 1101–1104, 2006.

[91] Margarita Kotti, Vassiliki Moschou, and Constantine Kotropoulos. Speaker segmentation and clustering. *Signal Processing*, 88(5):1091–1124, 2008.

[92] W. Kraaij and J. Arlandis. Trecvid-2004 story segmentation task: Overview. In *in Proceedings Text Retrieval Conference (TREC), Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication*, 2004.

[93] C.-C. Jay Kuo. *Video Content Analysis Using Multimodal Information: For Movie Content Extraction, Indexing and Representation.* Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[94] G. C. Langelaar, P. M. Van Roosmalen, J. Biemond, and R. L. Langendijk. *Image and Video Databases: Restoration, Watermarking and Retrieval.* Elsevier Science Inc., New York, NY, USA, 2000.

[95] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. W3c recommendation, W3C, February 1999.

[96] B. R. Lee, A. Ben Hamza, and Hamid Krim. An active contour model for image segmentation: A variational perspective. *IEEE Internat. Conf. on Acoustics Speech and Signal Processing*, may 2002.

[97] G. Leighbody, G. W. Heiman, and K. Bowler. Word intelligibility decrements and the comprehension of time-compressed speech. *Perception and psychophysics*, vol. 40, no. 6:pp. 407–411, 1986.

[98] Riccardo Leonardi, Pierangelo Migliorati, and Maria Prandini. Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled markov chains. *IEEE Trans. Circuits Syst. Video Techn.*, 14(5):634–643, 2004.

[99] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Fifth International Conference on Computer Vision, ICCV'95*, page 637, 1995.

[100] S. Li. Content-based classification and retrieval of audio using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5):619–625, 2000.

[101] Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. Technical report, Report HPL-2001-191, HP Laboratories, 2001.

[102] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *International Journal of Image and Graphics (IJIG)*, pages 469–486, 2001.

[103] Rainer Lienhart. Automatic text recognition for video indexing. proc. *ACM Multimedia 96*, pages 11–20, 1996.

[104] Rainer Lienhart. Automatic text recognition in digital videos. *SPIE 2666: Image and Video Processing IV*, pages 180–188, 1996.

[105] Rainer Lienhart. Video ocr: A survey and practitioner's guide. *In Video Mining, Kluwer Academic Publisher*, pages pp. 155–184, 2003.

[106] Chih-Chin Liu, Jia-Lien Hsu, and Arbee L. P. Chen. An approximate string matching algorithm for content-based music data retrieval. In *ICMCS, Vol. 1*, pages 451–456, 1999.

[107] C.-S. Lu and P.-C. Chung. Wold features for unsupervised texture segmentation. *IAPR Inter. Conf. on Pattern Recognition*, pages 1689–1693, 1998.

[108] Lie Lu and Hong Zhang. Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing*, 10, 2002.

[109] Lie Lu and Hong Jiang Zhang. Unsupervised speaker segmentation and tracking in real-time audio content analysis. *Multimedia Systems*, 10(4):332–343, 2005.

[110] Shi Lu, Irwin King, and Michael R. Lyu. Video summarization by video structure analysis and graph optimization. In *roceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, Taipei, Taiwan*, pages 1959–1962, 2004.

[111] Shi Lu, Michael R. Lyu, and Irwin King. Video summarization by spatial-temporal graph optimization. In *2004 IEEE International Symposium on Cirsuits and Systems (ISCAS) 2004, Vancouver, Canada*, pages 197–200, 2004.

[112] Jordi Luque and Javier Hernando. Robust speaker identification for meetings: Upc clear'07 meeting room evaluation system. In *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 266–275, 2007.

[113] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Colour and texture descriptors. *IEEE Trans. Circuits and Systems for Video Tech.*, 2001.

[114] Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, 2004.

[115] Kathy Melih and Ruben Gonzalez. Structured audio coding for content based manipulation and retrieval. In *IMSA*, pages 58–62, 1999.

[116] Michael Mills, Jonathan Cohen, and Yin Yin Wong. A magnifier tool for video data. In *roceedings of the ACM CHI 92 Human Factors in Computing Systems Conference, Monterey, California, USA*, pages 93–98, 1992.

[117] Dalibor Mitrovic, Matthias Zeppelzauer, and Horst Eidenberger. Analysis of the data quality of audio descriptions of environmental sounds. Technical report, Technical University of Vienna, 2006.

[118] S. Molau and M. Pitz. Computing mel-frequency cepstral coefficient on the power spectrum. Aachen, University of Technology, 2001.

[119] Hans Moravec. Towards automatic visual obstacle avoidance. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, page 584, August 1977.

[120] Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman. That obscure object of desire: Multimedia metadata on the web, part 2. *IEEE MultiMedia*, 12(1):54–63, 2005.

[121] Milind R. Naphade and John R. Smith. A hybrid framework for detecting the semantics of concepts and context. In *2nd Int. Conf. on Image and Video Retrieval, CIVR 2003, Urbana-Champaign, IL, USA*, pages 196–205, 2003.

[122] Noel E. O'Connor, Csaba Czirjek, S. Deasy, Sean Marlow, Noel Murphy, and Alan F. Smeaton. News story segmentation in the fischlar video indexing system. *DCU Eprints Archive (Ireland)*, 2001.

[123] A. H. Omar. Audio segmentation and classification. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Kgs. Lyngby, Denmark, 2005. Supervised by Associate Professor Jan Larsen.

[124] Nosa Omoigui, Liwei He, Anoop Gupta, Jonathan Grudin, and Elizabeth Sanocki. Time-compression: Systems concerns, usage, and benefits. In *Proceeding of the CHI 99 Conference on Human Factors in Computing Systems, Pittsburg, PA, USA*, pages 136–143, 1999.

[125] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *ACM Multimedia*, pages 570–579, 2002.

[126] H. Pan, P. van Beek, and M. I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *ICASSP '01: Proceedings of the Acoustics, Speech, and Signal Processing, 2001. on IEEE International Conference*, pages 1649–1652, Washington, DC, USA, 2001. IEEE Computer Society.

[127] D. M. Papworth, E. Izquierdo, and Alan Pearmain. Using hmms to classify video sequences. In *Proceedings of the European Workshop for the Integration of Knwoledge, Semantics and Digital Media Technology,London, UK*, 2004.

[128] N. Patel and L. Sethi. Audio characterization for video indexing. In *Proceedings of the SPIE on Storage and Retrieval for Still Image and Video Databases*, volume 2670, pages 373–384, San Jose, USA, 1996.

[129] Peter F. Patel-Schneider, Patrick Hayes, and Ian Horrocks. OWL Web Ontology Language - Semantics and Abstract Syntax. W3c recommentation, W3C, February 2004.

[130] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. *Proc. TRECVID workshop 2004*, 2004.

[131] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. *ACM Multimedia*, pages 21–30, 1996.

[132] Silvia Pfeiffer. Hyperlinking to time offsets: The temporal uri specification, 12 2007.

[133] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Effelsberg. Scene determination based on video and audio features. *Multimedia Tools Appl.*, 15(1):59–81, 2001.

[134] Silvia Pfeiffer and Uma Srinivasan. Tv anytime as an application scenario for mpeg-7. In *ACM Multimedia Workshops*, pages 89–92, 2000.

[135] Marcus J. Pickering, Lawrence W. C. Wong, and Stefan M. Rüger. Anses: Summarisation of news video. In *CIVR*, pages 425–434, 2003.

[136] D. Pye. Content-based methods for the management of digital music. In *ICASSP '00: Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference*, pages 2437–2440, Washington, DC, USA, 2000. IEEE Computer Society.

[137] G. Quenot, D. Mararu, S. Ayache, M. Charhad, L. Besacier, M. Guironnet, D. Pellerin, J. Gensel, and L. Carminati. Clips-lis-lsr-labri experiments in trecvid 2004. In *Proceedings Text Retrieval Conference (TREC), Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication*, 2004.

[138] Zeeshan Rasheed and Mubarak Shah. Movie genre classification by exploiting audio-visual features of previews. In *in Proc. of International Conference on Pattern Recognition*, pages 1086–1089, 2002.

[139] Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *CVPR (2)*, pages 343–350, 2003.

[140] Mika Rautiainen, Tapio Seppänen, Jani Penttilä, and Johannes Peltola. Detecting semantic concepts from video using temporal gradients and audio classification. In *2nd Int. Conf. on Image and Video Retrieval, CIVR 2003, Urbana-Champaign, IL, USA*, pages 260–270, 2003.

[141] Stephan Repp, Jörg Waitelonis, Harald Sack, and Christoph Meinel. Segmentation and annotation of audiovisual recordings based on automated speech recognition. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pages 620–629, 2007.

[142] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle. Feature extraction and temporal segmentation of audio signals. *ICMC*, 1998.

[143] S. Rovetta, P. Gastaldo, and R. Zunino. Objective assessment of mpeg-video quality: a neural-network approach. In *in Proc. Int. Joint Conf. on Neural Networks, Washington, DC, USA*, volume vol. 2, pages pp. 1432–1437, 2001.

[144] Yong Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 1111–1118, 2000.

[145] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Constructing table-of-content for videos. *ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries*, 7:359–368, 1999.

[146] Daniel M. Russell. A design pattern-based video summarization technique: Moving from low-level signals to high-level structure. In *3rd Annual Hawaii International Conference on System Sciences HICSS 2000, Maui, Hawaii, USA*, 2000.

[147] H. Sack and J. Waitelonis. Automated annotations of synchronized multimedia presentations. In *In Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, CEUR Workshop Proceedings*, june 2006.

[148] David A. Sadlier and Noel E. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. Circuits Syst. Video Techn.*, 15(10):1225–1233, 2005.

[149] Massimo De Santo, Gennaro Percannella, Carlo Sansone, and Mario Vento. A multi-expert system for shot change detection in mpeg movies. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18(5):933–956, 2004.

[150] Caterina Saraceno and Riccardo Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. In *Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP-98), Chicago, Illinois, USA*, pages 363–367, 1998.

[151] Frederik Schaffalitzky and Andrew Zisserman. Automated scene matching in movies. In *Proceedings of the European Workshop for the Integration of Knwoledge, Semantics and Digital Media Technology,London, UK*, pages 186–197, 2002.

[152] Peter Schallauer, Axel Pinz, and Werner Haas. Automatic restoration algorithms for 35mm film. *Journal of Computer Vision Research*, 1(3):pp. 60–85, 1999.

[153] Manfred R. Schroeder. *Computer speech: recognition, compression, synthesis*. Springer-Verlag New York, Inc., New York, NY, USA, 1999.

[154] Tanja Schultz and Alex Waibel. Language independent and language adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51, 2001.

[155] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[156] Navid Serrano, Andreas E. Savakis, and Jiebo Luo. A computationally efficient approach to indoor/outdoor scene classification. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 146–149, 2002.

[157] Ismail Shahin. Speaker identification in the shouted environment using suprasegmental hidden markov models. *Signal Processing*, 88(11):2700–2708, 2008.

[158] Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision*. MIT Press, 2006.

[159] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000.

[160] S. Singh and M. Markou. An approach to novelty detection applied to the classification of image regions. *IEEE Transactions on Data and Knowledge Engineering*, pages 396–407, April 2004.

[161] John R. Smith and Peter Schirling. Metadata standards roundup. *IEEE MultiMedia*, 13(2):84–88, 2006.

[162] Michael A. Smith and Takeo Kanade. Video skimming and characterization through the combination of image and language understanding. In *International Workshop on Content-Based Access of Image and Video Databases, CAIV'98, Bombai, India*, pages 61–70, 1998.

[163] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Kolma, and F. J. Seinstra. The mediamill trecvid 2004 semantic video search engine. In *In Proc. TRECVID workshop 2004*, 2004.

[164] Cees Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl.*, 25(1):5–35, 2005.

[165] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Improved video content indexing by multiple latent semantic analysis. In *Proc. of 3rd Int. Conf. on Image and Video Retrieval, CIVR 2004, Dublin, Ireland*, pages 483–490, 2004.

[166] Christian Spevak. Sound spotting - a frame-based approach. In *Proc. of the Second Annual International Symposium on Music Information Retrieval: ISMIR*, pages 35–36. Online]. Available: http://ismir2001.indiana.edu/posters/spevak.pdf, 2001.

[167] Anthony Stefanidis, Panos Partsinevelos, Peggy Agouris, and Peter Doucette. Summarizing video datasets in the spatiotemporal domain. In *1th International Workshop on Database and Expert Systems Applications (DEXA'00), Greenwich, London, UK*, pages 906–912, 2000.

[168] C. Stiller and J. Konrad. Estimating motion in image sequences. a tutorial on modelling and computation of 2d motion. *IEEE Signal Processing Magazine*, pages 70–91, 1999.

[169] Andreas Stolcke, Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lei, Tim Ng, Mari Ostendorf, K. Sonmez, Anand Venkataraman, Dimitra Vergyri, Wen Wang, Jing Zheng, and Qifeng Zhu. Recent innovations in speech-to-text transcription at sri-icsi-uw. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1729–1744, 2006.

[170] M. Stricker and A. Dimai. Spectral covariance and fuzzy regions for image indexing. *Machine Vision and Applications*, pages 66–73, 1997.

[171] S. R. Subramanya, Abdou Youssef, Bhagirath Narahari, and Rahul Simha. Automated classification of audio data and retrieval based on audio classes. In *Computers and Their Applications, Proceedings of the ISCA 14th International Conference*, pages 141–145, 1999.

[172] C.Y. Suen. Future challenges in handwriting and computer applications. *3rd International Symposium on Handwriting and Computer Applications*, 1987.

[173] Masaru Sugano, R. Isaksson, Yasuyuki Nakajima, and Hiromasa Yanagihara. Shot genre classification using compressed audio-visual features. In *Proc. on International Conference on Image*, pages 17–20, 2003.

[174] Haiping Sun, Joo Hwee Lim, Qi Tian, and Mohan S. Kankanhalli. Semantic labeling of soccer video. In *In Proceedings of IEEE Pacific-Rim Conference on Multimedia (ICICS-PCM)*, pages 1787–1791, 2003.

[175] Xinding Sun and Mohan S. Kankanhalli. Video summarization using r-sequences. *Real-Time Imaging*, 6(6):449–459, 2000.

[176] H. Sundaram and S. F. Chang. Computable scenes and structures in films. *IEEE Transactions on Multimedia*, 4(4):pp. 482–491, Dec. 2002.

[177] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. In *International Workshop on Content-based Access of Image and Video Databases, Bombai, India*, pages 42–51, 1998.

[178] Wallapak Tavanapong and Junyu Zhou. Shot clustering techniques for story browsing. *IEEE Transactions on Multimedia*, 6(4):517–527, 2004.

[179] Candemir Toklu, Shih-Ping Liou, and Madirakshi Das. Video abstract: A hybrid approach to generate semantically meaningful video summaries. In *IEEE International Conference on Multimedia and Expo 2000, New York, NY, USA*, pages 1333–1336, 2000.

[180] Vasanth Tovinkere and Richard J. Qian. Detecting semantic events in soccer games: Towards a complete solution. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo, ICME 2001„ Tokyo, Japan*, 2001.

[181] B. T. Truong, S. Venkatesh, and C. Dorai. Film grammar based refinements to extracting scenes in motion pictures. In *IEEE International Conference on Multimedia & Expo (ICME'02)*, volume 1, pages 281–284, Lausanne, Switzerland, August 2002.

[182] Ba Tu Truong, Svetha Venkatesh, and Chitra Dorai. Automatic genre identification for content-based video categorization. In *Proc. International Conference on Pattern Recognition*, pages 4230–4233, 2000.

[183] George Tzanetakis, Georg Essl, and Perry Cook. Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, pages 293–302, 2002.

[184] Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John S. Boreczky. Video manga: generating semantically meaningful video summaries. In *ACM Multimedia Proceedings of the 7th ACM International Conference on Multimedia '99, Orlande, Florida, USA*, pages 383–392, 1999.

[185] Hirotada Ueda, Takafumi Miyatake, Shigeo Sumino, and Akio Nagasaka. Automatic structure visualization for video editing. In *Human-Computer Interaction, INTERACT '93, IFIP TC13 International Conference on Human-Computer Interaction, Amsterdam, Netherlands*, pages 137–141, 1993.

[186] Aditya Vailaya and Anil K. Jain. Detecting sky and vegetation in outdoor images. *Storage and Retrieval for Media Databases 2000*, 3972(1):411–420, 1999.

[187] Aditya Vailaya, Anil K. Jain, and Hong Jiang Zhang. On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935, 1998.

[188] Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That obscure object of desire: Multimedia metadata on the web, part 1. *IEEE MultiMedia*, 11(4):38–48, 2004.

[189] Nuno Vasconcelos and Andrew Lippman. A spatiotemporal motion model for video summarization. In *Conference on Computer Vision and Pattern Recognition (CVPR 1998), Santa Barbara, CA, USA*, pages 361–366, 1998.

[190] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Trans. on Multimedia*, 4(4):pp. 492–498, Dec. 2002.

[191] L. Vincent. Morphological greyscale reconstruction in image analysis: efficient algorithms and applications. *IEEE Trans. on Image Processing*, pages 176–201, 1993.

[192] Oriol Vinyals and Gerald Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008)*, pages 426–431, 2008.

[193] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[194] T. Vlachos. Flicker correction for archived film sequences using a nonlinear model. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(4):508–516, April 2004.

[195] A. C. Bovik Wang and B. L. Evans. Blind measurement of blocking artifacts in images, proc. In *IEEE Int. Conf. Image Proc*, volume vol. 3, pages pp. 981–984, 2000.

[196] Avery Wang. An industrial strength audio search algorithm. In *In Proc. 4th International Conference on Music Information Retrieval*, 2003.

[197] Jinjun Wang, Changsheng Xu, Chng Eng Siong, and Qi Tian. Sports highlight detection from keyword sequences using hmm. In *Proc. International Conference on Multimedia and Expo, Teipei, Taiwan*, pages 599–602, 2004.

[198] Stuart Weibel. The dublin core: A simple content description model for electronic resources, 1997.

[199] Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke. A study of multilingual speech recognition. In *In Proc. European Conf. on Speech Communication and Technology*, pages 359–362, 1997.

[200] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, 3(3):27–36, 1996.

[201] H. R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, vol. 4:pp. 317–320, 1997.

[202] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, Amsterdam, Netherlands*, pages 29–32, 2005.

[203] Min Yang, Yingchun Yang, and Zhaohui Wu. A pitch-based rapid speech segmentation for speaker indexing. In *Seventh IEEE International Symposium on Multimedia (ISM 2005)*, pages 571–576, 2005.

[204] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, 2002.

[205] Boon-Lock Yeo and Bede Liu. Visual content highlighting via automatic extraction of embedded captions on mpeg compressed video. *Digital Video Compression: Algorithms and Technologies, Proc. SPIE 2668-07*, 1996.

[206] K. C. Yow and R. Cipolla. Scale and orientation invariance in human face detection. In *British Machine Vision Conference*, pages pp. 745–754, 1996.

[207] H. Yu, G. Bozdagi, and S. Harrington. Feature based hierarchical video segmentation. *ICIP*, 1997.

[208] J. Yuan. Tsinghua university at trecvid 2004: Shot boundary detection and high-level feature extraction. *Proc. TRECVID workshop 2004*, 2004.

[209] Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kuai, HI, USA*, pages 123–130, 2001.

[210] Yun Zhai, Zeeshan Rasheed, and Mubarak Shah. University of central Florida at TRECVID 2004. In *in Proceedings Text Retrieval Conference (TREC), Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication*, 2004.

[211] Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

[212] T. Zhang and C.-C. J. Kuo. Heuristic approach for audio data segmentation and annotation. *ACM Multimedia*, pages 67–76, 1999.

[213] Tong Zhang and C.-C. Jay Kuo. *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer Academic Publishing, 1998.

[214] Tong Zhang and C.-C. Jay Kuo. Hierarchical system for content-based audio classification and retrieval. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 3001–3004, 1998.

[215] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[216] Di Zhong and Shih-Fu Chang. Structure analysis of sports video using domain models. In *ICME*, 2001.

[217] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu, and Ann Christine Catlin. Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7(4):648–666, 2005.