Technologie-Zentrum Informatik

Technical Report 34

Automatische Videoanalyse und textuelle Annotation

AVAnTA Abschlußbericht Phase 3

A. Jacobs, A. Miene und O. Herzog

TZI-Bericht Nr. 34
2006

Universität Bremen

TZI-Berichte

Herausgeber:
Technologie-Zentrum Informatik
Universität Bremen
Am Fallturm 1
28359 Bremen
Telefon: +49-421-218-7272
Fax: +49-421-218-7820
E-Mail: info@tzi.de
http://www.tzi.de

ISSN 1613-3773

Inhaltsverzeichnis

1 Allgemeine Angaben

Antragsteller:	Prof. Dr. O. Herzog
Geschäftszeichen:	He 989/41
Institut / Lehrstuhl:	Technologie-Zentrum Informatik (TZI) FB 3 - Mathematik und Informatik, Universität Bremen
Adresse:	Am Fallturm 1 28359 Bremen
Berichtszeitraum:	Januar 2003 bis Dezember 2005

2 Einleitung

Ziel des Projektes AVAnTA ist die Erforschung und Entwicklung neuer Methoden und Algorithmen zur syntaktischen und semantischen Analyse von strukturierten Videodokumenten, welche der Annotation und der Strukturierung dieser digitalen Dokumente dienen. Im dritten Abschnitt des Projektes wurden Konzepte zur inhaltlichen Analyse und Strukturierung einer eingeschränkten Klasse von Videodokumenten untersucht.
Die Analyse der logischen Struktur von Videodokumenten dient nicht nur einer weiteren inhaltlichen Erschließung der Materialien sondern ermöglicht auch die Entwicklung neuer Navigationsmechanismen für digitale Videodokumente. Die Ergebnisse der automatischen Annotation sind die Basis für eine integrierte inhaltsbasierte Suche über Bilder, Videos und Texte. Mögliche Anwender sind Betreiber großer Videobibliotheken, wie z.B. Fernsehanstalten oder Bibliotheken, die Archive mit Videos, Bildern und Texten in digitaler Form unterhalten. Eine effektive Suche in solchen digitalen Bibliotheken wird unterstützt. Durch die logische, semantisch bedeutsame Strukturierung wird eine effiziente nichtlineare Navigation durch die Suchergebnisse ermöglicht. Dabei steht der Dokumentar im Mittelpunkt, eine vollautomatische Analyse von Videodokumenten, im allgemeinen Fall heutzutage immer noch eine ungelöste Aufgabe, ist nicht die Absicht dieses Projekts.
Im nächsten Abschnitt werden die Ergebnisse der dritten Projektphase beschrieben. Nach einer Beschreibung der Arbeitspakete, wie sie im Projektantrag entwickelt wurden, folgt die Präsentation der Ergebnisse der einzelnen Arbeitspakete. Nach einer Auflistung der im Zusammenhang mit dem Projekt entstandenen Diplomarbeiten schließt der Bericht mit einer Zusammenfassung.

3 Arbeits- und Ergebnisbericht

3.1 Vorherige Projektphasen

Im Laufe der ersten Phase des Projektes wurde ein Konzept für ein Videoanalyse- und Retrievalsystem mit einer Annotations- und einer Retrievalkomponente entwickelt. Aufgabe der Annotationskomponente ist die Durchführung der einzelnen Analyseschritte und die Generierung der Annotationen, welche in einer Annotationsdatenbank zusammengeführt werden. Folgende Fragestellungen wurden während der Laufzeit der ersten Projektphase untersucht:

- Entwicklung eines kombinierten Shotanalyseverfahrens, das neben harten Schnitten auch Spezialeffekte wie Ein- und Ausblendungen sowie durch Überblendungen oder Wischblenden versteckte Schnitte erkennt. Durch Verwendung adaptiver Schwellwerte arbeitet das Verfahren auf einer breiten Palette unterschiedlichen digitalen Videomaterials ohne dass eine manuelle Justierung der Parametern erforderlich wird [12].

- Automatische Reduktion des Videos auf eine Menge repräsentativer Einzelbilder (key frames oder Mosaicbilder). Hierzu wurden basierend auf einer Analyse der dominanten Kamerabewegung Entscheidungskriterien für die Verfahrensauswahl – Mosaicing und key frames Bestimmung – entwickelt. Zur Extraktion repräsentativer key frames wurde eine Methode entwickelt, die in Abhängigkeit vom Inhalt der jeweiligen Einstellung automatisch eine geeignete Menge von key frames auswählt und extrahiert. Durch Einsatz eines Verfahrens zur automatischen Gruppierung von Kameraeinstellungen anhand ihres Bildinhaltes kann die Zahl der key frames im Falle sich wiederholender Einstellungen mit gleichem Bildinhalt zusätzlich reduziert werden. Darüber hinaus liefert die Gruppierungsanalyse Aufschluß darüber, wie die Sequenz geschnitten wurde [3, 15]. Die Einzelbilder werden hinsichtlich ihrer Farb-, Textur- und Kontureigenschaften analysiert. Hierzu wurden die im Rahmen des IRIS-Projektes [23, 24, 27, 28] entwickelten Einzelbildanalysemethoden integriert und weiterentwickelt [4, 5].

- Automatische Extraktion von Texteinblendungen. Die Trennung der Texteinblendungen von strukturiertem Bildhintergrund erfolgt über ein spezielles Farbsegmentierungsverfahren in Kombination mit Heuristiken über das Aussehen von Buchstaben und Text [13].

- Synthese der verschiedenen multimedialen Informationskanäle im Video über einen datenbankbasierten Ansatz [15]. Dies eröffnet die Möglichkeit der Anbindung des AVAnTA-Systems an andere datenbankbasierte Retrievalsysteme wie z.B. OMNIS/2 [35].

- Bewegungsanalyse anhand der Bewegungsvektoren aus dem MPEG-Format. Das entwickelte Verfahren eignet sich gut, um mit geringem Berechnungsaufwand Bewegungstendenzen und grobe Bewegungsrichtungen wie

z. B. Kameraschwenks zu identifizieren. Die Nachteile des Verfahrens liegen in der starken Abhängigkeit der Ergebnisse davon, wie der verwendete MPEG-Encoder implementiert wurde, und darin, dass es besonders bei differenzierten Bewegungen mehrerer Objekte zu ungenau arbeitet, um klare Aussagen über die Objektgrenzen und ihre Bewegungsrichtung zu liefern.

Der in Phase 1 verfolgte Ansatz behandelt das Problem der Bildanalyse im Video, indem das Video mittels einer Menge repräsentativer Einzelbildern dargestellt wird, welche mit Methoden der Einzelbildanalyse untersucht werden. Hierbei bleibt jedoch der dynamische Aspekt des Videos gegenüber dem Einzelbild, die Bewegungsinformation, unberücksichtigt. Gerade die Darstellung von bewegten Objekten, welche nicht nur in räumlichen, sondern auch in zeitlichen Beziehungen zueinander stehen, macht einen wichtigen Teil des Videos als Bewegtbildmedium aus. Deswegen wurde in Phase 2 des Projektes der Forschungsschwerpunkt auf die Untersuchung verschiedener Bewegungsanalyseverfahren zur Erkennung und Verfolgung von bewegten Objekten und auf die inhaltliche Interpretation von bewegten Szenen im Video gelegt.

Folgende Fragestellungen wurden im Rahmen der zweiten Projektphase untersucht:

- Entwicklung von Bewegungsanalyseverfahren

  Im Rahmen dieses Arbeitspaketes wurden verschiedene Bewegungsanalyseverfahren untersucht und/oder entwickelt:

  - Ein merkmalsbasiertes Zuordnungsverfahren basierend auf Farbregionen. Hierzu entstand eine Diplomarbeit [11].
  - Ein fortgeschrittenes Blockvergleichsverfahren. Das Verfahren wurde ebenfalls im Rahmen einer Diplomarbeit entwickelt [21].
  - Ein filterbasiertes Verfahren unter Verwendung von Gaborfiltern [38].

- Verbesserung des Mosaicverfahrens durch Einbeziehung von Bewegungsinformationen

  Zur Verbesserung der modellbasierten Bewegungsschätzung als Grundlage des Mosaicings wurden alternative Verfahren zur Bestimmung von Stützpunkten untersucht, die zusätzlich zur Minimierung des Aperturproblems auch das bei kleinskaligen, hochfrequenten Bildbereichen (welche weniger anfällig für das Aperturproblem sind) häufig auftretende Korrespondenzproblem berücksichtigen. Bei der Betrachtung wurde zusätzlich Wert gelegt auf eine homogene Verteilung der Stützpunkte, welche sich i.A. positiv auf die globale modellbasierte Schätzung auswirkt. Das Problem der globalen Bewegungsschätzung und des Mosaicings in Szenen mit bewegten Objekten wurde in einer Diplomarbeit behandelt [9].

- Szeneninterpretation anhand von Bewegungsinformationen

Dieses Arbeitspaket beschäftigte sich mit Methoden zur qualitativen Beschreibung von Bewegungsinformationen [16, 19]. Hier kann abstrahiert werden von der konkreten Art und Weise, wie die Bewegungsinformation gewonnen wurde. Anstelle von Ergebnissen der Videoanalyse können u.A. auch Mitschnitte von simulierten Fussballspielen für die entwickelten Methoden verwendet werden [10].

- Modellierung von Bewegungssituationen

  Im Rahmen dieses Arbeitspaketes wurden Methoden zur Modellierung komplexerer Bewegungssituationen auf Basis der elementaren Beschreibungen des vorigen Arbeitspaketes entwickelt [16]. Die entwickelten Methoden wurden in verschiedenen Anwendungsdomänen eingesetzt, von der Fussballsimulation [19, 20] bis zum intelligenten Fahrzeug [18]. Es entstand eine Dissertation zu diesem Thema [17].

Folgende Publikationen sind noch nach Beendigung der zweiten Phase aus dem Projekt hervorgegangen:

- Andrea Miene und Thomas Wagner. *Static and Dynamic Qualitative Spatial Knowledge Representation for Physical Domains*. In KI, no. 2, 2006 (noch nicht erschienen).

- Andreas D. Lattner, Andrea Miene, Ubbo Visser und Otthein Herzog. *Sequential Pattern Mining for Situation and Behavior Prediction in Simulated Robotic Soccer*. In RoboCup International Symposium 2005 (noch nicht erschienen).

- Andrea Miene, Ubbo Visser und Otthein Herzog. *Recognition and Prediction of Motion Situations Based on a Qualitative Motion Description*. In D. Polani, B. Browning, A. Bonarini und K. Yoshida (Hrsg.), RoboCup 2003: Robot Soccer World Cup VII, Vol. 3020, Lecture Notes in Computer Science, S. 77-88, Springer, 2004.

  Diese Publikation hat den Scientific Challenge Award RoboCup 2003 gewonnen.

- Andrea Miene, Andreas D. Lattner, Ubbo Visser und Otthein Herzog. *Dynamic-Preserving Qualitative Motion Description for Intelligent Vehicles*. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV '04), S. 642-646, Parma, Italien, 14.-17. Juni 2004.

3.2 Kooperationen

Die folgenden Kooperationen mit Vertretern anderer Forschungsdisziplinen und Forschungseinrichtungen konnten auf Grundlage der im Projekt AVAnTA entwickelten Methoden durchgeführt werden:

3.2.1 Projekt Notebook-Universität

In diesem erfolgreich in Kooperation mit dem Fachbereich Kulturwissenschaften der Universität Bremen durchgeführten Projekt wurde der Prototyp eines Video-Content-Management-Systems entwickelt, auf Basis der in AVAnTA entwickelten Algorithmen. Das entstandene System wurde in der hier gegenständlichen dritten Phase des Projekts AVAnTA erweitert. Das System wird im Fachbereich Kulturwissenschaften zur Verwaltung umfangreichen Videomaterials zu Lehr- und Forschungszwecken eingesetzt.
Eine kommerzielle Vermarktung des Systems durch eine vor kurzem geschehene Ausgründung aus dem TZI ist geplant.

3.2.2 Projekt iMediathek

In diesem zusammen mit der Kunsthochschule Bremen durchgeführten Projekt wurde der Prototyp eines Internet-Archivs für Videokunst erstellt [33]. Die Annotation des Videomaterials wurde mit einer angepassten Version des im Projekt „Notebook-Universität“ entwickelten und in AVAnTA erweiterten Systems vorgenommen.

3.2.3 DELOS Network of Excellence on Digital Libraries

Das DELOS Network of Excellence on Digital Libraries[1] ist ein Verbund der wichtigsten europäischen Arbeitsgruppen im Bereich der digitalen Bibliotheken, mit dem Ziel der Entwicklung von Technologien für die digitale Bibliothek der nächsten Generation. Die Aktivitäten der Arbeitsgruppe von Prof. Herzog vor allem auch in Bezug auf das AVAnTA-Projekt haben dazu geführt, dass das TZI im DELOS Network beteiligt ist. Das TZI leitet eines der momentanen Arbeitspakete im Bereich „Audio-visual and non-traditional objects“, bei dem es um die automatische Segmentierung und Strukturierung von Nachrichtensendungen geht. Es arbeitet in diesem Arbeitspaket mit Wissenschaftlern der Technischen Universität Kreta (TUC) sowie des Fraunhofer Instituts für Medienkommunikation (IMK) zusammen.

3.3 Ausgangsfragen und Zielsetzung

Der Schwerpunkt des Arbeitsprogrammes der Phase 3 des Forschungsvorhabens liegt in der *inhaltlichen Strukturierung und Klassifikation von Videodokumenten.* Die Arbeitspakete sind

- AP1: Untersuchung von Methoden zur Modellierung von Metadatenmodellen zur inhaltlichen Strukturierung von Videos
- AP2: Identifizierung und Extraktion strukturierender Merkmale
- AP3: Erforschung von Methoden zur Abbildung der automatisch erkannten strukturierenden Merkmale auf die Metadatenmodelle

[1] Online unter `http://www.delos.info/`, zuletzt geprüft am 28.12.2005

- AP4: Automatische Klassifikation der strukturellen Einheiten
- AP5: Entwicklung neuer Navigationsstrukturen für Videodokumente

In den nun folgenden Abschnitten wird der Inhalt der einzelnen Arbeitspakete konkretisiert.

AP1: Untersuchung von Methoden zur Modellierung von Metadatenmodellen zur inhaltlichen Strukturierung von Videos

Im Rahmen dieses Arbeitspaketes ist zunächst zu untersuchen, welche Möglichkeiten der Gliederung für verschiedene Typen digitaler Videos bestehen.
Bei Magazinsendungen wäre eine Gliederung der Sendung in einzelne Beiträge sinnvoll. Ein Beitrag behandelt jeweils ein in sich abgeschlossenes Thema, z.B. eine Reportage über ein Sportereignis, eine Kunstausstellung o. ä. . Ein neuer Beitrag wird in der Regel durch eine Anmoderation eingeleitet.
Ein digitales Videoarchiv kann Videos mit unterschiedlichem Aufbau enthalten (z.B. verschiedene Sendungen eines Senders). Dabei entspricht der Aufbau jeder Folge einer Sendung einem bestimmten Sendungsformat, das Vorgaben über die Struktur der Sendung enthält. Deswegen soll eine Modellierungsumgebung für Sendungsformate geschaffen werden, mittels derer ein Dokumentar den generellen Aufbau der anschließend zu analysierenden Videodokumente spezifizieren kann. Die Kenntnis über den zu erwartenden Aufbau der Sendung erleichtert die automatische Strukturierung des Videos. Es findet ein Abgleich zwischen automatisch extrahierten strukturierenden Merkmalen und der zu erwartenden Struktur statt.
Mit Hilfe der Modellierungsumgebung generiert der Dokumentar ein Metadatenmodell, das nicht nur den Rahmen sondern auch stilistische Eigenschaften der Sendung, die zur Erkennung der Struktur hilfreich sein können, beschreibt. Hierzu gehören beispielsweise verwendete Schnittechniken, Aussehen und Aufbau der Texteinblendungen, Einsatz von Animationen und Einblendung von Logos und anderen grafischen Elementen, etc., darüber hinaus Zeitpunkt und Dauer des Einsatzes dieser Elemente. Die Modellierungsumgebung sollte es daher auch erlauben, bildhafte Informationen z.B. in Form von Beispielbildern hinzuzufügen. Die dem Metadatenmodell zugrunde liegende Syntax muß ausreichend flexibel gestaltet sein, da eine Sendung nicht nach einem starren Schema abläuft. Anzahl und Länge der Beiträge variieren in der Regel.

AP2: Identifizierung und Extraktion strukturierender Merkmale

Strukturierende Merkmale sind Merkmale im Video, an denen die Struktur bzw. der Übergang zwischen den aufeinander folgenden logischen Einheiten des Videos erkannt werden kann. Bei einer gut konzipierten Sendung muß auch für den Zuschauer der Übergang von einem Thema zu nächsten klar erkennbar sein, damit er dem Inhalt der Sendung folgen kann. Ein sehr einfaches Beispiel für die Gestaltung eines Beitragsüberganges wäre eine Sendung, bei der jeder Beitrag

mit einer sich wiederholenden, für diese Sendung charakteristischen Sequenz beendet wird, gefolgt von einem Standbild mit eingeblendeten Produktionsdaten zu dem Beitrag. Oftmals erkennt man den Beginn eines neuen Beitrages aber auch nur daran, dass eine Studioszene mit der Anmoderation des nächsten Beitrages folgt. Oftmals unterscheidet sich auch die Schnittechnik innerhalb eines Beitrages von der zwischen zwei verschiedenen Beiträgen, z.B. Wischblenden beim Themenwechsel, harte Schnitte innerhalb eines Filmbeitrages.
Bei der Identifizierung strukturierender Merkmale kann auf den bereits geleisteten Vorarbeiten in den vorangegangenen Projektphasen (vgl. Abschnitt 3.1) aufgebaut werden. Auch auf die Erfahrungen aus der Zusammenarbeit mit internationalen Partnern im Rahmen des Projekts ADViSOR kann zurückgegriffen werden. Diese beziehen sich vor allem auf die Detektion menschlicher Gesichter und auf die Erkennung gesprochener Sprache im Audiosignal der Videodokumente. Eine breite Palette an Merkmalen kann bereits automatisch extrahiert werden. Im Rahmen dieses Arbeitspaketes soll untersucht werden, welche dieser Merkmale zur automatischen Erkennung der logischen Struktur des Videos verwendet werden können. Einige der bereits untersuchten Methoden können erweitert werden, um weitere für eine Strukturierung des Videos wesentliche Merkmale zu liefern:

- Erweiterung der automatischen Erkennung von Kameraeinstellungen um die Klassifikation der verwendeten Schnittechnik.
- Erweiterung der automatischen Erkennung von Texteinblendungen zur Extraktion von Metadaten bezüglich des Textes wie z.B. Position, Größe, Stil, Farbe des Textes, Erkennung strukturierender grafischer Elemente wie z.B. Trennlinie usw.
- Abgleich zwischen Schnitten und Pausen im Audiosignal: Ein Schnitt mit einer gleichzeitigen Audiopause stellt i.A. einen stärkeren Einschnitt dar als ein Schnitt, bei dem das Audiosignal ohne Unterbrechung weiterläuft.

Gegebenenfalls müssen aber auch weitere Methoden untersucht und integriert werden, wie z.B.

- Automatische Erkennung wiederkehrender Bildelemente wie z.B. Logos und Schriftzüge etc.
- Detektion menschlicher Gesichter: Falls zeitgleich eine Texteinblendung vorhanden ist, kann der Text mit der Person in Bezug gebracht werden. Dies erleichtert die Erkennung von Interviewsituationen und ermöglicht die Übernahme der Personennamen in das Inhaltsverzeichnis.
- Automatische Unterscheidung verschiedener Einstellungsgrößen wie Weiteinstellung, Totale, Halbtotale, Naheinstellung oder Detaileinstellung. Der Wechsel zwischen den Einstellungsgrößen findet durch Schnitt oder Zoomen statt. Daher ist eine automatische Erkennung von Zooms ebenfalls eine gute Ergänzung des Analyse-Instrumentariums.

AP3: Erforschung von Methoden zur Abbildung der automatisch erkannten strukturierenden Merkmale auf Metadatenmodelle

Um eine aktuelle Instanz einer Sendung zu strukturieren, deren Sendungsformat zuvor als Metadatenmodell modelliert wurde, muß eine Zuordnung zwischen den extrahierten strukturierenden Merkmalen und dem Modell hergestellt werden. In diesem Arbeitspaket sollen Methoden untersucht werden, die basierend auf Korrelationsverfahren Hypothesen über eine mögliche Sendungsstruktur aufbauen, diese verifizieren, auf ihre Plausibilität prüfen und so schließlich eine Strukturierung der aktuellen Sendung entlang des Modells vorschlagen. Dieses Ergebnis bildet die Grundlage für die Erzeugung eines Inhaltsverzeichnisses und damit für einen Zugriff auf Videos entlang ihrer logischen Struktur.

AP4: Automatische Klassifikation der strukturellen Einheiten

Ziel dieses Arbeitspaketes ist die Entwicklung einer Methode zur thematischen Klassifikation der erkannten strukturellen Einheiten. Dazu ist zunächst eine Analyse typischer Themengebiete erforderlich. Hieraus kann eine Taxonomie von Themengebieten abgeleitet werden. Die Zuordnung von Beiträgen zu Themengebieten erfolgt über die aufgrund der Analyse verfügbaren textuellen Informationen aus dem Beitrag. Diese entstammen der Spracherkennung und der Video-OCR. Es existiert eine Liste der bei der (manuellen) Indexierung am häufigsten verwendeten Begriffe. Diese basiert auf dem Inhalt des FESAD Dokumentationssystems für Videos[2] In Zusammenarbeit mit Dokumentaren oder durch Abgleich mit den textuellen Informationen zu bereits klassifizierten Beiträgen können zu den zu unterscheidenden Themengebieten Wortlisten erstellt werden. Jedem Wort in der Liste kann mittels statistischer Methoden ein Relevanzwert zugewiesen werden, der angibt, mit welcher Wahrscheinlichkeit der Beitrag zu einem Themengebiet gehört, wenn das Wort gesprochen wurde oder in den Texteinblendungen enthalten ist. Durch Vergleich der Wortlisten mit den textuellen Informationen und Verrechnung der Relevanzwerte kommt man zu einer Klassifikation des Beitrages zu einem oder mehreren, nach Wahrscheinlichkeit geordneten, Themengebieten.
Die Wissensbasis zur Verwaltung der Themengebiete und Wortlisten soll so konzipiert werden, dass sie durch den Benutzer erweiterbar ist. Das heißt, dass der Benutzer für beliebige Themengebiete die Wortlisten und die Relevanzwerte festlegen kann. Darüber hinaus ist es denkbar, Wortlisten und Relevanzwerte aufgrund von durch den Benutzer spezifizierten Beispielen lernen zu lassen.

AP5: Entwicklung neuer Navigationsstrukturen für Videodokumente

Die automatisch extrahierte logische Struktur des Videos nebst der Klassifikation der einzelnen logischen Einheiten kann zur Navigation im Video genutzt

[2]Das FESAD Dokumentationssystem ist ein beim SWR und einer ständig wachsenden weiteren Zahl von Landesrundfunkanstalten der ARD eingesetztes System für die Dokumentation von eigenproduzierten Fernsehbeiträgen. In der FESAD-Datenbank werden Formaldaten, Sach- und Bildinhaltsbeschreibungen abgelegt.

werden. Dabei soll der Bildinhalt auch in geeigneter Weise präsentiert werden. Das Ziel ist es, aus den verfügbaren Informationen ein Inhaltsverzeichnis aufzubauen, das mit den entsprechenden Passagen im Video verlinkt ist. Beispielsweise kann zu jedem Beitrag das Themengebiet zusammen mit den relevantesten Schlüsselwörtern, die zu der Klassifikation geführt haben, aufgenommen werden. Spezielle Einheiten wie Nachrichtenüberblick oder Wettervorhersage können besonders gekennzeichnet werden. Die Namen der beteiligten Personen können auf ihre Wortbeiträge verlinkt werden. Beim Klick auf einen Beitrag kann zunächst die Anmoderation präsentiert werden, um dem Betrachter einen Überblick zu geben.
Der während der Anmoderation gesprochene Text kann als textueller Abstract des Beitrages genutzt werden.

Zeitplan

Für die zeitliche Abfolge der im letzten Abschnitt vorgestellten Arbeitspakete (AP) war ursprünglich der folgende zeitliche Rahmen vorgesehen:

AP	Quartal							
	I/02	II/02	III/02	IV/02	I/03	II/03	III/03	IV/03
AP1	X	X						
AP2			X	X				
AP3					X			
AP4						X	X	
AP5								X

3.4 Probleme und Abweichungen

Aufgrund personeller Engpässe und eines Wechsels in der Besetzung des Projektes mußte das Projekt zeitweise ausgesetzt werden. Der ursprüngliche Zeitplan konnte so nicht eingehalten werden. Im Dezember 2003 verließ Herr Rachid Fathi die Arbeitsgruppe. Die Stelle wurde mit Herrn Arne Jacobs besetzt. Das Projekt wurde insgesamt bis zum Dezember 2005 kostenneutral durch Übertragung von Kapazitäten aus der zweiten Phase verlängert.
Das Arbeitspaket vier wurde zugunsten der anderen Arbeitspakete zurückgestellt. Zur Begründung siehe die Beschreibung der Ergebnisse in 3.5.

3.5 Entwicklungsstand und Ergebnisse

AP1: Untersuchung von Methoden zur Modellierung von Metadatenmodellen zur inhaltlichen Strukturierung von Videos

Im Rahmen dieses Arbeitspaketes wurden eine Sprache zur Modellierung von Sendestrukturen sowie eine Umgebung zum Erstellen von Sendemodellen in dieser Sprache entwickelt [1]. Die Sprache unterscheidet sich von anderen Beschrei-

bungssprachen für Videos wie MPEG-7[3], SMIL[4] oder NEWS-ML[5] darin, dass sie die Beschreibung von Klassen von Sendungen erlaubt. Während NEWS-ML den Schwerpunkt auf die Beschreibung einzelner Nachrichtenbeiträge legt, SMIL eher auf die Erstellung von zeitabhängigen Multimediadokumenten (u.A. Videos) zugeschnitten ist, und MPEG-7 die inhaltliche Beschreibung konkreter Videoinstanzen erlaubt, ermöglicht die hier entwickelte Sprache die formale Beschreibung eines Sendeformats. Dies beinhaltet den allgemeinen Ablauf sowie die syntaktischen Strukturelemente, die für alle Sendungen dieses Formats gleich sind. Um die Verwendbarkeit der Sprache zu erhöhen, baut sie auf MPEG-7 auf. Mithilfe eines Java-basierten Werkzeugs zur Unterstützung der Erstellung von Sendemodellen wird dem Dokumentar eine einfache Möglichkeit an die Hand gegeben, ein von ihm betreutes Sendeformat zu beschreiben.
Aus diesem Arbeitspaket ging eine Diplomarbeit hervor [2] (siehe auch 3.6).

AP2: Identifizierung und Extraktion strukturierender Merkmale

Wie im Arbeitsplan vorgeschlagen wurden einige bereits entwickelte Ansätze zur Merkmalsextraktion weiterentwickelt. So wurde weiter an Algorithmen zur Shotgrenzenerkennung gearbeitet, deren Ergebnisse durch Teilnahme an den TRECVID Workshops evaluiert wurde [14, 8]. Bei der Bearbeitung des ersten Arbeitspaketes wurde zudem festgestellt, dass sich Position und Größe von Texteinblendungen gut dazu eignen, bestimmte charakteristische Elemente eines Sendeformats zu unterscheiden. Zu diesem Thema wurde eine Diplomarbeit verfaßt [22] (siehe auch 3.6). Sie behandelt die Erkennung von Textstellen in Videos in Echtzeit.
Zur automatischen Erkennung von Moderationen und anderen wiederkehrenden visuellen Strukturelementen einer Sendung wurde ein Ansatz entwickelt, der die Selbstähnlichkeit innerhalb von strukturierten Sendungen ausnutzt [6]. So können u.A. die am häufigsten auftretenden Sprecher und/oder Moderatoren in Magazin- und Nachrichtensendungen automatisch und nahezu parameterfrei identifiziert und zeitlich lokalisiert werden, ohne vorheriges Training. Einzig eine Angabe über die Anzahl der erwarteten Präsentationsklassen ist erforderlich. Die Arbeit des Dokumentar bei der Erstellung von Sendemodellen wird so erleichtert. Bei der Anwendung auf Nachrichtensendungen des Formats "CNN Headline News" erreicht der Ansatz eine Genauigkeit von durchschnittlich $91,6\%$. Bei der Anwendung auf die Sendungen des Magazins "ZDF Auslandsjournal" werden alle Moderationen zwischen den Berichten automatisch erkannt.
Andere Ansätze zur Identifikation von Moderationen benötigen eine Trainingsphase [25] und müssen darüberhinaus berücksichtigen, dass Kleidung und Hintergrund der Moderationen zwischen den Sendungen in gewissen Grenzen wechseln können, oder Modelle für die Identifikation von Moderationen müssen ma-

[3] Online unter `http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm`, zuletzt geprüft am 28.12.2005
[4] Online unter `http://www.w3.org/AudioVideo/`, zuletzt geprüft am 28.12.2005
[5] Online unter `http://www.newsml.org`, zuletzt geprüft am 28.12.2005

nuell erstellt werden [37]. Im Vergleich zu anderen unüberwachten Ansätzen wie [31] müssen die verwendete Merkmale nicht in einem Vektorraum liegen, es genügt ein Maß zum binären Vergleich.
Durch Vergleich mehrerer Sendungen eines Formats kann der Ansatz erweitert werden, um außerdem in jeder Sendung wiederkehrende strukturierende Sequenzen automatisch zu erkennen.
Es wurden zudem weitere Merkmale betrachtet: Für die Detektion von Gesichtern wurde der Ansatz von Lienhart et al. [32] verwendet. Für die Bewegungserkennung wurde ein eigener Ansatz entwickelt [7], in diesem Rahmen ist außerdem eine Diplomarbeit entstanden [9] (siehe auch 3.6).
Aus dem ebenfalls an der Universität Bremen am TZI entwickelten System PictureFinder [29] wurden Algorithmen zum visuellen Vergleich verwendet, um die Erkennung charakteristischer Sendestrukturen durch vom Dokumentar ausgewählte Beispielbilder zu ermöglichen.

AP3: Erforschung von Methoden zur Abbildung der automatisch erkannten strukturierenden Merkmale auf Metadatenmodelle

Aus Basis der extrahierten Merkmale wurde in diesem Arbeitspaket eine Abbildung auf das Sendemodell vorgenommen. Diese Abbildung läßt sich in zwei Schritte teilen:

- Klassifikation einzelner Videosequenzen auf Basis von im Sendemodell vorgegebenen Strukturprimitiva (z.B. Moderation, Vorspann, etc.);
- Reduktion von Klassifikationsfehlern durch Abgleich mit im Sendemodell vorgegebenen zeitlichen Beziehungen und Einschränkungen.

Auf dem V3D2-Abschlußworkshop in Berlin 2004 konnten für den ersten Schritt bereits Klassifikationsergebnisse auf Basis von Support Vector Machines mit einer Genauigkeit von ca. 95% präsentiert werden[6]. Diese Klassifikation wurde auf Shotebene durchgeführt.
Es wurde außerdem ein Bayes-Klassifikator auf Basis von Ähnlichkeiten zu manuell ausgesuchten Beispielbildern einiger charakteristischer Sendestrukturelemente sowie auf Basis von Eigenface-Merkmalen auf detektierten Gesichtsregionen getestet. Dieser Klassifikator arbeitet auf Einzelbildern. Wir beschränken uns hierbei auf die Betrachtung jedes zehnten Bildes im Video, um die Berechnungszeit zu verkürzen. Der Klassifikator liefert im Vergleich zu dem shotbasierten SVM-Klassifikator ein leicht schlechteres Ergebnis, er ist jedoch robust gegenüber Fehldetektionen der Shotgrenzenerkennung.
Der Klassifikator wurde getestet auf Sendungen des Magazins "ZDF Auslandsjournal" von 2002. Als Sendestrukturprimitiva wurden fünf Klassen identifiziert:

- Vorspann
- Moderation

[6] siehe auch online unter `http://pengo.cg.cs.tu-bs.de/V3D2/Workshops/V3D2Workshop2004Berlin/folien/avanta.pdf`, zuletzt geprüft am 28.12.2005

- Landkarte
- Bericht
- Abspanntafel

Der Bayes-Klassifikator für die Klassen “Vorspann”, “Landkarte” und “Abspanntafel” wurde durch manuell segmentierte Beispielsequenzen trainiert. Die Klasse Moderation wurde mit durch die im zweiten Arbeitspakete entwickelte Moderationserkennung [6] vollautomatisch identifizierten Moderationssequenzen trainiert. Die Klasse “Report” wurde mit zufälligem Videomaterial trainiert.
Nach der framebasierten Klassifikation folgt ein Abgleich mit dem Sendemodell. Der Abgleich mit einem Sendemodell ist in dieser Form bisher neu. Es gibt jedoch einen Ansatz von Ivanov et al. [30] zum Parsen von Ereignissen in Überwachungsvideos, der vergleichbar ist. Er verbindet ebenfalls die Vorteile von manuell und automatisch erstellten Modellen, wobei das manuell erstellte Modell in Form einer regulären Grammatik vorliegt. Wir verfolgen hier allerdings einen anderen Ansatz auf Basis des Relaxation Labellings. Hierbei werden die im Sendemodell vorgegebenen zeitlichen Beziehungen in geeignete Kompatibilitätskoeffizienten für das Relaxation Labelling übersetzt.
Zunächst wird jedoch ein klassisches Relaxation Labelling zur Nachverarbeitung einer Segmentierung verwendet, um einzelne Fehler in der Klassfikation zu unterdrücken, und längere zusammenhängende Videosequenzen derselben Klasse zu erhalten. Der Kompatibilitätskoeffizient für diesen ersten Schritt ist somit nach [36], S. 229:

$$r_1((p_i, l_k), (p_j, l_l)) = \begin{cases} 1 & \text{falls } l_k = l_l \\ 0 & \text{sonst} \end{cases} \quad (1)$$

Dabei entsprechen p_i den betrachteten Einzelbildern und l_k den Klassen mit $l_k \in \{\text{Vorspann}, \text{Moderation}, \text{Landkarte}, \text{Bericht}, \text{Abspanntafel}\}$. Die Anfangswahrscheinlichkeiten für den iterativen Relaxationsprozeß werden direkt aus dem Ergebnis des Bayer-Klassifikators übernommen.
Im zweiten Relaxationsschritt werden zunächst aufeinanderfolgende Einzelbilder p_i derselben Klasse zu Sequenzen s_m zusammengefaßt. Die Labelwahrscheinlichkeiten der Sequenzen ergeben sich dabei als einfaches arithmetisches Mittel der Labelwahrscheinlichkeiten der einzelnen Bilder der Sequenz. Durch die im Sendemodell gegebenen zeitlichen Beziehungen zwischen Sendestrukturprimitiva ergab sich für das Magazin “ZDF Auslandsjournal” die in Tabelle 1 gezeigte Matrix von Kompatibilitätskoeffizienten, unter der Annahme, dass s_m zeitlich vor s_n liegt.
Die Koeffizienten ergeben sich aus folgenden modellierten zeitlichen Beziehungen:

- Nach dem Vorspann folgt immer eine Moderation.
- Nach einer Moderation kann eine Landkarte, ein Bericht, oder eine Abspanntafel folgen, wobei eine Moderation direkt gefolgt von einem Bericht weniger häufig vorkommt.

$r'_2((s_m, l_k), (s_n, l_l))$	l_l = Vorsp.	Mod.	Landk.	Bericht	Absp.
l_k = Vorspann	0.5	0.5	-1	-1	-1
Moderation	-1	0.5	0.5	0.25	0.5
Landkarte	-1	-1	0.5	0.5	-1
Bericht	-1	-1	-1	0.5	0.5
Abspanntafel	-1	0.5	-1	-1	-1

Tabelle 1: Aus dem Sendemodell für "ZDF Auslandsjournal" extrahierte Kompatibilitätskoeffizienten

- Nach einer Landkarte folgt immer ein Bericht.
- Nach einem Bericht folgt immer eine Abspanntafel.
- Nach einer Abspanntafel folgt immer eine Moderation.

Um Klassifikationsfehler, die zu zerteilten Sequenzen führen, zu berücksichtigen, werden aufeinanderfolgende Sequenzen mit derselben Klasse ebenfalls zugelassen, wenn die Sequenz nicht immer sehr kurz ist, wie hier die Abspanntafel. Um die Tatsache zu berücksichtigen, dass der Vorspann immer am Anfang der Sendung liegt, wird außerdem festgelegt:

$$r_2((s_m, l_k), (s_n, l_l)) = \begin{cases} -1 & \text{falls } (l_k \neq \text{Vorspann}) \text{ und } s_m \text{ erste Sequenz} \\ r'_2((s_m, l_k), (s_n, l_l)) & \text{sonst} \end{cases} \quad (2)$$

Dieses zweistufige Vorgehen erzielt auf den untersuchten Sendungen eine Precision von 100% für alle Klassen, sowie einen Recall von 100% für die Klassen "Vorspann", "Moderation", "Bericht" und "Abspanntafel" und einen Recall von 90% für die Klasse "'Landkarte". Die Werte für Precision und Recall auf Einzelbildbasis sind etwas niedriger, was auf die Übergänge zwischen den verschiedenen Klassen zurückzuführen ist, welche keiner der Klassen zuzuordnen sind, da sie nicht modelliert wurden. Diese Ungenauigkeiten an den Sequenzgrenzen sollten aber in der Praxis irrelevant sein, und eine Modellierung der Übergänge würde den Modellierungsaufwand beträchtlich erhöhen. Tabelle 2 zeigt die Precision- und Recall-Werte auf Einzelbildbasis.

AP4: Automatische Klassifikation der strukturellen Einheiten

Im Laufe der Zusammenarbeit mit dem Fraunhofer IMK im Delos Network of Excellence on Digital Libraries (siehe auch Abschnitt 3.2.3) wurde klar, dass das Fraunhofer-Institut für das vom TZI geleitete Arbeitspaket eine Komponente zur Spracherkennung und Transkriptionsbasierten Themenklassifikation entwickeln würde. Unter dem Gesichtspunkt, dass die vom IMK erwartete Komponente anstelle der Ergebnisse des vierten Arbeitspaketes verwendet werden

Klasse	Precision	Recall
Vorspann	1	1
Moderation	0.958	0.953
Landkarte	0.861	0.912
Bericht	0.993	1
Abspanntafel	1	1

Tabelle 2: Precision- und Recall-Werte auf Einzelbildbasis auf den Sendungen des "ZDF Auslandsjournal"

kann, wurde dieses Arbeitspaket daraufhin zugunsten der Arbeitspakete drei und fünf zurückgestellt. In Arbeitspaket drei wurden zusätzlich zu den geplanten Vorhaben Möglichkeiten untersucht, Modelle für die Struktur einer Sendereihe automatisch zu ermitteln. In Arbeitspaket fünf wurde außerdem eine Möglichkeit zur nichtlinearen Navigation in Videos basierend auf visueller Ähnlichkeit untersucht.

AP5: Entwicklung neuer Navigationsstrukturen für Videodokumente

Im Rahmen dieses Arbeitspaketes wurde das im Projekt „Notebook-Universität" (siehe Abschnitt 3.2.1) in Zusammenarbeit mit dem Fachbereich Kulturwissenschaften der Universität Bremen entwickelte Video-Content-Management-System um Funktionen zur Navigation in und zur Unterstützung der Annotation von Videos erweitert. In einer Keyframe-Ansicht wird zu jedem Shot der integrierten automatischen Shotgrenzenerkennung eine wählbare Anzahl an Keyframes gezeigt, um dem Dokumentar eine erste grobe Übersicht über das Video zu geben. Da die Shotlänge i.A. begrenzt ist und nicht mit der Länge des Videos zunimmt, eignet sich die Keyframe-Ansicht allerdings nicht zur Navigation durch ein ganzes Video. D.h., die Anzahl der Keyframes ist i.A. proportional zur Länge des Videos. Aus diesem Grund enthält das entwickelte System außerdem eine hierarchisch geordnete Übersicht der betrachteten Videos, basierend auf der automatisch erstellten Segmentierung des Videos auf Basis des Sendemodells.
Als weitere nicht-lineare Navigationsmöglichkeit wurden die im Projekt "PictureFinder" [29] entwickelten Bildvergleichsalgorithmen integriert, um eine beispielbasierte visuelle Suche innerhalb des Videomaterials zu ermöglichen. So kann z.B. direkt zu den Beiträgen gesprungen werden, die bereits in Ausschnitten in der Einleitungssequenz der Sendung gezeigt wurden. Die visuelle Suche erleichtert zudem die Erstellung von Sendemodellen, indem nach Ähnlichkeiten innerhalb und zwischen Sendungen eines Formats gesucht werden kann. Abbildung 1 zeigt einen Screenshot des Systems.
Zusätzlich zu den Navigationsmöglichkeiten beinhaltet das System einige der in den anderen Arbeitspaketen entwickelten Analysemodule. Weiterhin steht eine generische Annotationskomponente zur Verfügung sowie eine Suchfunktion auf

Abbildung 1: Ansichten des System mit hierarchischer Strukturansicht (links), Annotationsansicht (Mitte), Keyframeansicht und Ergebnissen der visuellen Suche (oben rechts) sowie Videoansicht (rechts)

Basis der textuellen Annotationen und den Ergebnissen der Analyse.

3.6 Dissertationen und Diplomarbeiten

Die folgende Disseration wurde in der zweiten Phase des Projektes AVAnTA begonnen und in der dritten Phase fertiggestellt:

- Andrea Miene (2004), *Räumlich-zeitliche Analyse von dynamischen Szenen*:

  Die Dissertation von Frau Miene beschäftigt sich mit der qualitativen Beschreibung von Bewegung und der Interpretation und Vorhersage von Bewegungssituationen.

Die folgenden Diplomarbeiten sind im Kontext des Projektes AVAnTA entstanden:

- Björn Melzer (2002), *Bewegungsanalyse in Bildfolgen auf Basis von Farbregionen*:

  Diese Diplomarbeit wurde im Rahmen der zweiten Projektphase erstellt. Sie beschäftigt sich mit der Verfolgung von durch eine Farbsegmentierung erhaltenen Regionen in Videos.

- Arne Jacobs (2003), *Mosaicing auf Szenen mit bewegten Objekten*:

  Das Thema dieser Arbeit ging ursprünglich aus der zweiten Phase des Projektes hervor. Die Ergebnisse können jedoch auch im Kontext des zweiten Arbeitspaketes der dritten Projektphase zur bewegungsbasierten Merkmalsextraktion genutzt werden.

- Norbert Wilkens (2003), *Detektion von Videoframes mit Texteinblendungen in Echtzeit*:

  Das Thema dieser Arbeit ergibt sich direkt aus dem Arbeitsplan des Projektes (AP2). Ziel ist das schnelle Detektieren von Textstellen mitsamt Angaben über Position und Größe, um eine effizientere Texterkennung zu ermöglichen.

- Lars Bankert (2005), *Entwicklung einer Modellierungsumgebung für Sendestrukturen*:

  Das Thema dieser Arbeit resultiert ebenfalls aus dem Arbeitsplan des Projektes (AP1). Ein Ergebnis der Arbeit ist u.A. eine Modellierungssprache für strukturierte Sendungen

- Dominik Ströhlein (Arbeitstitel, voraussichtliche Beendigung: 2006), *Bewegungsanalyse auf Basis von Blockvergleichsverfahren zur MPEG-7 Visual-Descriptor Extraktion*

  Die Arbeit beschäftigt sich mit der Extraktion von MPEG-7-Metainformationen zur Unterstützung von Bewegungsanalyse-Systemen.

4 Zusammenfassung

In der dritten Phase des Projekts AVAnTA wurde eine Modellierungssprache zur Beschreibung von Sendeformaten für strukturierte Fernsehsendungen entwickelt. Von gängigen Beschreibungssprachen für Videos wie MPEG-7[7] oder NEWS-ML[8] unterscheidet sie sich dahingehend, dass sie geeignet ist, ganze Klassen von Sendungen anstatt einzelner Videodokumente zu beschreiben. Die Nützlichkeit dieser Sprache für Dokumentare, die viele Videos desselben Formats dokumentieren müssen, ergibt sich aus der Tatsache, dass Sendungen eines Formats i.A. eine starke Sturktur besitzen, die sich auch visuell äußert, um dem Zuschauer das Verfolgen der Sendung und ihrer Struktur zu ermöglichen.

Durch die Entwicklung neuer Algorithmen zur Detektion von Texteinblendungen und zur Bewegungserkennung und durch Erweiterung bereits existierender Algorithmen aus früheren Phasen des Projektes sowie durch Einsatz existierender Algorithmen zur Detektion von Gesichtern und zur Bestimmung visueller Ähnlichkeit wird es möglich, die charakteristischen Strukturelemente eines Sendeformats zu erkennen. Ein neuer Ansatz zur Untersuchung der Selbstähnlichkeit in Videos automatisiert den Prozeß weiter und vereinfacht zudem die Modellierung von Sendeformaten. Durch Weiterentwicklung dieses Ansatzes wird es in Zukunft evtl. möglich sein, die Struktur einer Sendung auf Basis einiger Beispiele vollautomatisch zu erkennen, so dass das manuelle Erstellen von Sendemodellen entfallen kann.

Zum Abgleich einer Videosendung mit dem zugrundeliegenden Sendemodell auf Basis der extrahierten visuellen Merkmale wurde ein neuer Ansatz auf Basis des Relaxation Labellings entwickelt. Dieser Ansatz basiert auf einer Übersetzung des im Sendemodell enthaltenen Wissen über die zeitlichen Relationen der charakteristischen Strukturelemente des Sendeformats in geeignete Kompatibilitätskoeffizienten für das Relaxation Labelling.

Um eine effiziente Dokumentation von Sendungen zu ermöglichen, wurde ein am TZI der Universität Bremen entwickeltes Softwaresystem erweitert, um u.A. eine nichtlineare Navigation innerhalb der Sendungen auf Basis des Sendemodells zu ermöglichen. Durch Integration geeigneter Suchmethoden erleichtert das System zudem die Erstellung von Sendemodellen, falls noch kein solches Modell für eine Sendereihe existiert, ohne dass beim Dokumentar Wissen über die Struktur der Reihe vorhanden sein muß.

Von anderen Systemen für Videoarchivierung und Retrieval, wie z.B. den sehr bekannten und erfolgreichen Informedia-Projekten der Carnegie Mellon University [26], oder dem MediaMill-System der Universität von Amsterdam [34] unterscheidet sich das Projekt AVAnTA vor allem durch eine andere Zielsetzung: Der Dokumentar eines Fernsehsenders soll bei der Archivierung und Annotation seiner Sendungen unterstützt werden. Durch diese Fokussierung wird möglich und leichter, was bei einer Betrachtung von Videos im allgemeinen Fall nicht

[7] Online unter `http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm`, zuletzt geprüft am 28.12.2005

[8] Online unter `http://www.newsml.org`, zuletzt geprüft am 28.12.2005

oder nur sehr schwer zu erreichen ist: Eine verläßliche und semantisch bedeutsame Analyse der Sendungen.

Verzeichnis der Anlagen

1. Andrea Miene, Ubbo Visser und Otthein Herzog. *Recognition and Prediction of Motion Situations Based on a Qualitative Motion Description.* In D. Polani, B. Browning, A. Bonarini und K. Yoshida (Hrsg.), RoboCup 2003: Robot Soccer World Cup VII, Vol. 3020, Lecture Notes in Computer Science, S. 77-88, Springer, 2004.

2. A. Jacobs, Th. Hermes und O. Herzog. *Hybrid Model-based Estimation of Multiple Non-dominant Motions.* In Proceedings of the 26th DAGM Symposium on Pattern Recognition, S. 87-94, September 2004.

3. L. Bankert, A. Jacobs, A. Miene, Th. Hermes, G.T. Ioannidis und O. Herzog. *An Environment for Modelling Telecast Structures.* In AVIVIDiLib'05 Proceedings, S. 176-179, Mai 2005.

4. A. Jacobs. *Using Self-similarity Matrices for Structure Mining on News Video.* In Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06), Mai 2006 (noch nicht erschienen).

Projektpublikationen

[1] L. Bankert, A. Jacobs, A. Miene, Th. Hermes, G.T. Ioannidis, and O. Herzog. An environment for modelling telecast structures. In T. Catarci, S. Christodoulakis, and A. Del Bimbo, editors, *AVIVDiLib'05 Proceedings*, pages 176–179, May 2005.

[2] Lars Bankert. Entwicklung einer Modellierungsumgebung für Sendestrukturen. Master's thesis, Universität Bremen, 2005.

[3] Th. Hermes, C. Klauck, and O. Herzog. Knowledge-based image retrieval. In B. Jähne, H. Haussecker, and P. Geissler, editors, *Handbook of Computer Vision and Application, Vol. 3*, chapter 25, pages 517–532. Academic Press, 1999.

[4] Th. Hermes, A. Miene, and P. Kreyenhop. On textures: A sketch of a texture–based image segmentation approach. In R. Decker and W. Gaul, editors, *Classification and Information Processing at the Turn of the Millenium. (Proc. 23rd Annual Conference Gesellschaft für Klassifikation e.V. 10.-12. März 1999, Bielefeld)*, pages 210–218. Springer, 2000.

[5] Th. Hermes, A. Miene, and O. Moehrke. Automatic texture classification by visual properties. In R. Decker and W. Gaul, editors, *Classification and Information Processing at the Turn of the Millenium. (Proc. 23rd Annual Conference Gesellschaft für Klassifikation e.V. 10.-12. März 1999, Bielefeld)*, pages 219–226. Springer, 2000.

[6] A. Jacobs. Using self-similarity matrices for structure mining on news video. In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*, May 2006. (noch nicht erschienen).

[7] A. Jacobs, Th. Hermes, and O. Herzog. Hybrid model-based estimation of multiple non-dominant motions. In Rasmussen, Bülthoff, Giese, and Schölkopf, editors, *Proceedings of the 26th DAGM Symposium on Pattern Recognition*, pages 87–94. Springer-Verlag Berlin Heidelberg, September 2004.

[8] A. Jacobs, A. Miene, G.T. Ioannidis, and O. Herzog. Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. In E.M. Voorhees and L.P. Buckland, editors, *TRECVID 2004 Workshop Notebook Papers*, November 2004.

[9] Arne Jacobs. Mosaicing auf Szenen mit bewegten Objekten. Master's thesis, Universität Bremen, 2003.

[10] Andreas D. Lattner, Andrea Miene, Ubbo Visser, and Otthein Herzog. Sequential pattern mining for situation and behavior prediction in simulated robotic soccer. In *RoboCup International Symposium 2005*, 2005. (noch nicht erschienen).

[11] Björn Melzer. Bewegungsanalyse in bildfolgen auf basis von farbregionen. Master's thesis, Universität Bremen, 2002.

[12] A. Miene, A. Dammeyer, Th. Hermes, and O. Herzog. Advanced and adapted shot boundary detection. In D. W. Fellner, N. Fuhr, and I. Witten, editors, *Proc. of ECDL WS Generalized Documents*, pages 39–43, 2001.

[13] A. Miene, Th. Hermes, and G. T. Ioannidis. Extracting Textual Inserts from Digital Videos. In *Proceedings of the ICDAR 2001, Sixth International Conference on Document Analysis and Recognition*, Seattle, Washington, USA, September 10-13, 2001. (to appear).

[14] A. Miene, Th. Hermes, G.T. Ioannidis, R. Fathi, and O. Herzog. Automatic shot boundary detection and classification of indoor and outdoor scenes. In E. M. Voorhees and L. P. Buckland, editors, *Information Technology: The 11th Text Retrieval Conference, TREC 2002*, volume 500-251 of *NIST Special Publication*, pages 615–620. NIST - National Institut of Standards and Technology, May 2003.

[15] A. Miene and O. Herzog. AVAnTA – Automatische Video Analyse und textuelle Annotation. *it + ti – Informationstechnik und Technische Informatik*, 42(6), 2000.

[16] A. Miene and U. Visser. Interpretation of spatio-temporal relations in real-time and dynamic environments. In *Proceedings of the RoboCup 2001 International Symposium (colocated with IJCAI 2001)*, Seattle, USA, August 7-10 2001.

[17] Andrea Miene. *Räumlich-zeitliche Analyse von dynamischen Szenen.* PhD thesis, Universität Bremen, 2004.

[18] Andrea Miene, Andreas D. Lattner, Ubbo Visser, and Otthein Herzog. Dynamic-preserving qualitative motion description for intelligent vehicles. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '04)*, pages 642–646, June 2004.

[19] Andrea Miene, Ubbo Visser, and Otthein Herzog. Recognition and prediction of motion situations based on a qualitative motion description. In D. Polani, B. Browning, A. Bonarini, and K. Yoshida, editors, *RoboCup 2003: Robot Soccer World Cup VII*, volume 3020 of *Lecture Notes in Computer Science*, pages 77–88. Springer, 2004. Diese Publikation hat den Scientific Challenge Award RoboCup 2003 gewonnen.

[20] Andrea Miene and Thomas Wagner. Static and dynamic qualitative spatial knowledge representation for physical domains. *KI*, 2, 2006. (noch nicht erschienen).

[21] Dominik Ströhlein. Bewegungsanalyse auf Basis von Blockvergleichsverfahren zur MPEG-7 Visual-Descriptor Extraktion. Master's thesis, Universität Bremen, 2006. (voraussichtliche Beendigung: 2006).

[22] N. Wilkens. Detektion von Videoframes mit Texteinblendungen in Echtzeit. Master's thesis, Universität Bremen, 2003.

Literatur

[23] P. Alshuth, T. Hermes, J. Kreyß, and M. Röper. Gesucht - gefunden?! Wissensbasiertes Bildretrieval. *KI*, 10(4), 1996.

[24] P. Alshuth, Th. Hermes, Ch. Klauck, J. Kreyß, and M. Röper. IRIS — image retrieval for images and videos. In *First International Workshop, IDB-MMS (Image Databases and Multi-Media Search)*, pages 170–178, Amsterdam, The Netherlands, 1996.

[25] A. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H.Lin, T. Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yang, and H.D. Wactlar. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of the 12th Text Retrieval Conference (TREC)*, November 2003.

[26] A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 skirmishes. In *Proceedings of the 3rd TRECVID Workshop*. NIST, 2005.

[27] Th. Hermes, Ch. Klauck, J. Kreyß, and J. Zhang. Content-based Image Retrieval. In *Proceedings of CASCON '95 (CD-ROM*, Toronto, Canada, 7 - 9 November 1995.

[28] Th. Hermes, Ch. Klauck, J. Kreyß, and J. Zhang. Image Retrieval for Information Systems. In *Proceedings of IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, San Jose, CA, USA, 5 - 10 February 1995.

[29] Th. Hermes, A. Miene, and O. Herzog. Graphical search for images by picturefinder. *Int. J. Multimedia Tools and Applications. Special Issue on Multimedia Retrieval Algorithmics*, 27(2):229–250, November 2005.

[30] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872, 2000.

[31] Vikrant Kobla, David S. Doermann, and Christos Faloutsos. VideoTrails : Representing and visualizing structure in video sequences. In *ACM Multimedia*, pages 335–346, 1997.

[32] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings of ICIP 2002*, volume 1, pages 900–903, September 2002.

[33] M. Schieren and A. Jacobs. iMediathek. Internet platform for a video art archive. In Vito Cappellini and James Hemsley, editors, *Proceedings EVA 2005 Florence*, pages 100–105, March 2005.

[34] C. G. M. Snoek, J. van Gemert, J. M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. de Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, , and M. Worring. The MediaMill TRECVID 2005 semantic video search engine. In *Proceedings of the 3rd TRECVID Workshop*. NIST, 2005.

[35] Günther Specht and Michael G. Bauer. OMNIS/2: A multimedia meta system for existing digital libraries. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000)*, volume 1923, pages 180–189, 2000.

[36] Klaus D. Tönnies. *Grundlagen der Bildverarbeitung*. Pearson, München, Boston, San Francisco, 2005.

[37] Weiqiang Wang and Wen Gao. A fast anchor shot detection algorithm on compressed video. In *IEEE Pacific Rim Conference on Multimedia*, pages 873–878, 2001.

[38] Laurenz Wiskott. Segmentation from motion: combining gabor- and mallat-wavelets to overcome the aperture and correspondence problems. *Pattern Recognition*, 32(10):1751–1766, 1999.