Bild – Film – Diskurs: ein neuer integrativer Ansatz

John Bateman / Otthein Herzog / Rainer Malaka / Marion G. Müller
Universität Bremen und Jacobs University, Bremen

Problemstellung und Zielsetzungen

Die Interpretation und Analyse von Bildern, bewegten Bildern und Text-Bild-Kombinationen ist hinsichtlich vieler Aspekte wissenschaftlich äußerst herausfordernd sowie dringend notwendig. Solche Kombinationen entwickeln sich zur Zeit zu dem am meisten verwendeten Kommunikationsmodus überhaupt und sind bereits in traditionellen Dokumenten, Webseiten, Benutzerschnittstellen, Lehrmaterialien u.v.a. üblich. Aber unser theoretisches Verständnis dafür, wie genau diese komplexen kommunikativen Gegenstände Bedeutung tragen oder, warum sie die intendierte Bedeutung *nicht* tragen (zum Beispiel in Gebrauchanweisungen oder in unüblich zusammengesetzten Filmsequenzen, usw.) ist überraschend fragmentarisch. Das beantragte Vorhaben hat die Zielsetzung eine nachweisbare Verbesserung und Erweiterung dieses Grundverständnisses zu liefern, die unserer Meinung nach nur durch ein Zusammenwirken technisch-formaler quantitativer Methoden und qualitativer hermeneutischer Methoden zu leisten ist.

Zielsetzung der Universität Bremen

Die Antragsteller der Universität Bremen erzielen eine erhebliche Verbesserung — sowohl auf der theoretischen sowie der praktischen Ebene, in der automatischen Analyse von Text-Bild-Film-Zusammenhängen. Als Ausgangspunkt dient die Einbeziehung von Domänen- und Weltwissen in die automatische Verarbeitung von Bildern und Filmen, deren Notwendigkeit immer breiter werdende Akzeptanz findet (cf. Blankert et al. 2005). Es ist letztlich kaum möglich, Bildverarbeitung durchzuführen, ohne alltägliches Wissen über die dargestellten Objekte und Erreignisse zu haben. Gleichermaßen ist es nicht möglich auch nicht ansatzweise, Sequenzen von Filmeinstellungen in einer 'Event'-Darstellung automatisch zusammenzuführen, ohne über semantische Information des Inhaltes zu verfügen. Diese Probleme haben auch die Analyse von verbalem Diskurs in der Linguistik viele Jahre beeinträchtigt und eine Zusammenführung dieser Kompetenzbereiche stufen wir als höchst vielversprechend für eine neue Generation von Bild- und Filmbearbeitungstechniken ein. Sowohl Bild als auch Film können auch als 'Diskurse' im linguistischen Sinn betrachtet werden. Wenn wir die spezifischen Strukturen dieser Diskurse entdecken können, verspricht dies generische Organisationsprinzipien für diese Medien offenzulegen, die zu einer dramatischen Verbesserung in deren Verarbeitung führen könnten. Genau dies wird im vorliegenden Vorhaben angestrebt.

Zielsetzung der Jacobs University, Bremen

Die Antragsteller der Jacobs University gehen davon aus, dass visuelle Muster der menschlichen Perzeption und Interaktion sowie des Informationsaustausches ein in der Zukunft zunehmend relevantes Phänomenon sein werden. Durch die einzelnen und in Kooperation durchgeführten Teile des Gesamtvorhabens werden insbesondere Nachrichtenbilder im Kontext der Print- und Online-Berichterstattung, ihre Be- und Untertitelung sowie das Wechselverhältnis zwischen Bild und Text untersucht werden. Das Ziel ist die Bildung einer Typologie von Bildmotiven einerseits und von interpretativ erzeugten Sinnzusammenhängen andererseits. Dabei ist als Projektergebnis ein erhöhtes theoretisches Verständnis für Kategorien und Interpretationen von Nachrichtenbildern sowie eine Verbesserung der Erfassung von visuellen Daten zu erwarten. Diese Ziele stimmen mit den mittel- und langfristigen Forschung- und Entwicklungsplänen der Jacobs University überein. Eine starke Orientierung zur visuellen Kommunikation ist dort bereits etabliert und die konkreten durch dieses Vorhaben zu erreichenden Ergebnisse werden einen wesentlichen Beitrag zur Weiterentwicklung dieser Richtung innerhalb der Institution sowie in nationalen und internationalen Kooperationen darstellen.

Allgemeine Orientierung

Die unterschiedlichen, aber eng verzahnten Zielorientierungen der zwei beteiligten Institutionen bauen auf einer gemeinsamen theoretischen und praktischen Basis auf, wo die eingebrachten Kompetenzen sich in idealer Weise ergänzen. Darüber hinaus wird von beiden Seiten ein neuer und sehr allgemeiner Befund aus dem Bereich der linguistischen Diskurssemantik (Asher & Lascarides 2003) angewendet werden, der ein grundlegender Beitrag der sprach- und geisteswissenschaftlichen Seite darstellt. Hier wird argumentiert, dass es äußerst implausibel wäre, wenn linguistische Diskursverarbeitung sich direkt mit allgemeinem Weltwissen auseinandersetzen müsste. Die Komplexität der dafür benötigten Inferenzen ist einfach zu hoch, um vernünftige Verarbeitung zu erreichen. Asher & Lascarides schlagen deshalb vor, dass Logiken von unterschiedlicher Ausdrucksmächtigkeit für die zwei Bereiche von Diskurs und von Weltwissen zuständig sind.

Dies bringt eine signifikante Vereinfachung der Gesamtproblematik mit sich, die zahlreiche praktische sowie theoretische Anwendungsmöglichkeiten verspricht: kurz zusammengefasst, die verwendeten Diskursstrukturen setzen genaue Bedingungen dafür fest, welche Aspekte des Weltwissens in einer konkreten kommunikativen Situation potenziell von Relevanz sind und welche nicht. Diese Art von Diskursinterpretation ist in der Linguistik bereits sehr gut etabliert, aber eben nur in Bezug zu verbalem Diskurs (ohne Bilder). In diesem Antrag schlagen wir eine radikal neue Verwendung dieser Modularität vor: wir wollen untersuchen, wie entsprechende Diskursstrukturen für Bilder,

Bild-Text-Kombinationen sowie bewegte Bilder wie Film und Video ausgearbeitet werden und für die automatische Analyse zugänglich gemacht werden können. Zu erwarten ist eine Vereinfachung und Verallgemeinerung der automatischen Interpretationsmöglichkeiten und ein tiefgreifenderes Verständnis dafür, wie solche Medien als Bedeutungsträger funktionieren.

Eine solche Untersuchung ist außerdem von großer Bedeutung für unser linguistisches Verständnis von Diskurs und seinen Mechanismen. Es scheint der Fall zu sein, dass Film und Bilder häufig kommunikative Methoden verwenden, die sich wesentlich überlappen mit denen, die wir in verbalen Sprachen beobachten können (cf. Kress & van Leeuwen 1996, Bateman 2007). Deswegen behandeln zwei Teilprojekte jeweils konkrete Forschungsfragen, die aus der hermeneutischen Interpretation von Film und Bild in Anlehnung an Interpretation von Sprache definiert werden.

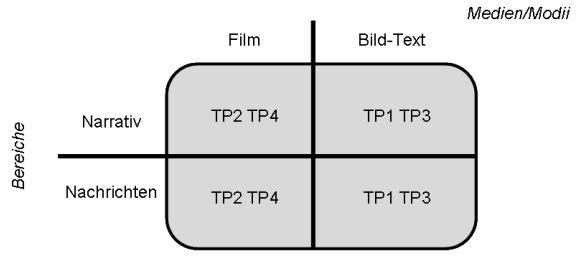
Konkret sollen Bilder, Bild-Text-Kombinationen und Filme aus zwei Bereichen analysiert werden: Narrationen und Nachrichten (inkl. Zeitungs- und Zeitschriftenartikel). Für den ersten Bereich wird primär dargelegt werden, wie filmische Mittel aktiviert werden, um narrative Vorgänge zu erzählen und wie dieses Diskurswissen für die automatische Analyse eingesetzt werden kann. Für den zweiten Bereich wird untersucht werden, wie kulturelle Ereignisse und Individuen visuell dargestellt werden und wie dies ihre automatische Verarbeitung und Annotation verbessern kann. Darüber hinaus wird quer untersucht, wie Bilder auch als 'Erzählungen' betrachtet werden müssen, um ihren Inhalt in Objekte, Personen und Ereignisse stufenweise zu dekomponieren, und wie 'Filme' wie Magazinsendungen, Dokumentarfilme usw. die gewonnenen Kategorien aus dem Bereich Kultur einsetzen.

Die beiden Themenbereiche Narrativ/Erzählungen und Presseartikel/Magazin- und Nachrichtensendungen sind aus folgenden Gründen als besonders repräsentativ und nützlich für beide beteiligten Institutionen ausgewählt worden.

1. *Narrativ* ist ein Kernstück unseres Verständnisses von Diskurs, Literatur und Film und findet immer häufigeren Gebrauch in anderen Zusammenhängen, wie in Dokumentarfilmen, didaktischem Stoff und sogar in der Werbung. Die vorgenommenen Untersuchungen werden dieses Verständnis einerseits vertiefen und andererseits für verbesserte automatische Analyseverfahren zur Filmsegmentierung und -klassifikation sorgen.

2. In *Zeitungs- und Zeitschriftenartikeln* sowie *Nachrichtensendungen* werden jeden Tag kulturelle Ereignisse und kulturell signifikante Individuen visuell dargestellt und konstruiert. Dieser Prozess ist nicht neutral, sondern sowohl in politischer als auch in sozialer Hinsicht bedeutungstragend. Die für die Bedeutungsstiftung verwendeten visuellen Methoden sollen offengelegt und für die automatische Analyse zugänglich gemacht werden.

Fortschritte in beiden Bereichen benötigen eine empirische Grundlage, die nur durch verbesserte automatische Verfahren möglich sein wird. Das Zusammenspiel von gewählten Themenbereichen und deren Medien ist in folgendem Bild veranschaulicht; hier sieht man schon die Zuordnung der Teilprojekte, deren detaillierte formale und quantitative Grundlagen unten dargestellt werden.



Bezug zu den förderpolitischen Zielen des Förderschwerpunktes

In diesem Vorhaben werden quantitative analytische Methoden für die Verarbeitung von Bild und Film aus natur- und technikwissenschaftlichen Bereichen und qualitativ hermeneutische analytische Methoden für die linguistische und sozialkulturelle Interpretation von Text, Text-Bild-Kombinationen und Film aus geisteswissenschaftlichen Bereichen zum ersten Mal kombiniert.

Besondere Ergebnisse werden im **Umgang mit wissenschaftlichen Informationen und Daten** erwartet. Die Schwerpunktsetzung in der Kombination von Bild und Text wird die **methodische und technische Erfassung, Speicherung, Bearbeitung und Auswertung** von diesen Datentypen signifikant verbessern. Das streng empirisch und daten-orientiert ausgelegte Forschungsprogramm zusammen mit den herangezogenen Kultur-, Sprach- und Literaturrelevanten Analysekategorien wird ganz gezielt zur **Reflexion von Daten und Messungen** bzgl. **kulturwissenschaftlicher Fragestellungen** hinführen. Darüber hinaus werden traditionell text-bezogene Fächer in vielen Bereichen aus Literatur-, Sprach-, Kultur- und Sozialwissenschaften immer häufiger mit Gegenständen konfrontiert, die jenseits der Grenzen von Text zu platzieren sind. Von der genaueren technischen Erfassung dieser Gegenstände werden daher auch **weitere geistes- wie naturwissenschaftliche Fächer stark profitieren**. Projektergebnisse werden direkt diskurs-bezogene Probleme in der **Modellierung von sprach- und literaturwissenschaftlichen Phänomenen** behandeln, während die Speicherung und Bearbeitung der gesammelten Daten die **technische Umsetzung für die Bearbeitung linguistischer Korpora** (hier mitsamt in bildlicher Form) wesentlich vorantreiben wird. Dabei werden die meisten Themen aus dem Schwerpunkt (b) der Förderrichtlinien des BMBFs im Programm "Wechselwirkungen zwischen Natur- und Geisteswissenschaften" mehrmals adressiert und in jedem Bereich sind signifikante Fortschritte zu erwarten.

Das breite Spektrum von gezielten Ergebnisse lässt sich zurück zu folgenden Tatsache zurückführen.

Bedeutungsträger in allen Bereichen, und insbesonders in der Wissenschaft, Wirtschaft, Politik und den Medien, sind immer mehr aus Texten, Bildern, bewegten Bildern und deren Kombinationen komponiert. Diese Kombinationen sind mit erheblichen Problemen für text-basierte Analyseansätze verbunden. Um den Erfordernissen der modernen sogenannten 'multimodalen Texte' gewachsen zu sein, sind neue analytische Methoden dringend notwendig. Auf der einen Seite sind die vorhandenen Frameworks zur Analyse solcher multimodaler Texte aus hermeneutischen Traditionen der Geisteswissenschaften oft nicht allgemein gültig und nicht in der Lage, bedeutungsstiftende Prozesse zu erläutern. Auf der anderen Seite sind vorhandene formale und technische Ansätze unzufriedenstellend wegen der immer noch sehr niedrigen Abstraktionsebene der möglichen Analysen und Klassifikationsmerkmale. Fortschritt auf beiden Seiten ist jetzt nur möglich durch eine tiefgreifende Integration und Kommunikation zwischen dem formalen, technischen Umgang und dem qualitativen Umgang mit multimodalen Texten. Genau diese Wechselwirkung wird in dem hier beantragten Projekt vorgeschlagen.

Eine sinnvolle Zusammenarbeit zwischen den eher quantitativen, formalen und technischen Ansätzen und den eher qualitativen, hermeneutischen Ansätzen ist in den verschiedenen Bereichen vorstellbar. Für das akute und höchst aktuelle Problem der Analyse von multimodalen Texten ist es jedoch absolut unabdingbar. Nichtsdestotrotz gestaltet sich die notwendige Kombination außerordentlich schwierig, weil der Umgang mit und die Auffassung von den zu analysierenden Gegenständen, sowie die analytischen Frameworks selbst so unterschiedlich sind. Im beantragten Projektverbund wird diese Problematik in einer völlig neuen Art und Weise gelöst. Durch den Einsatz einer qualitativ ausgerichteten, aber trotzdem quantitiv-nahen Analysemethodik wird eine neuartige Wechselwirkung erreicht.

Wissenbasis

Zwei Wissenbasen werden die zentrale Rolle in allen Teilprojekten übernehmen. Die erste ist die Sammlung von Filmen und Videosequenzen, die bereits in dem früheren Projekt AVAnTA (DFG-SPP V3D2: Miene & Herzog 2000) angefangen worden ist. Diese Sammlung soll ergänzt werden durch narrative Filmsequenzen. Während des Projekts wird diese Sammlung stets mit neu gewonnenen Annotationen angereichert. Die zweite ist der seit 10 Jahren in Entwicklung befindliche Bildkorpus von Presse- und Zeitschriftenbildern von Prof. Müller an der Jacobs University Bremen. Diese Bilder sind auch teilweise digitalisiert und annotiert. Die Ergebnisse von TP1 und TP3 beziehen sich direkt auf diesen Korpus.

Vorarbeiten der Antragsteller

Die vier Hauptantragsteller sind alle international erfolgreich ausgewiesene Forscher in den beteiligten Disziplinen. Die Vorarbeiten im Einzelnen lassen sich kurz wie folgt darstellen.

Prof. John Bateman, PhD, Lehrstuhl für Angewandte Englische Sprachwissenschaft an der Universität Bremen, arbeitet seit 1994 im Bereich Multimodalität und linguistische Behandlung von multimodalen Phänomenen. Er hat seit 1989 Forschungsprojekte auf nationaler und internationaler Ebene geleitet. 1999–2002 war er Mitantragsteller in einem von dem Britischen EPSC geförderten Projekt, das ein analytisches Verfahren für multimodale linguistische Analyse ausgearbeitet hat. Der Ansatz ist in mehreren internationalen Zeitschriften und Sammelbänden erschienen. Bateman ist durch eine Mehrzahl von durchgeführten Forschungsprojekten innerhalb nationaler und internationaler Forschungsverbünde (z.B., EU-Projekte Dandelion, Agile, OASIS; DFG SFB/TR8 “Spatial Cognition”; UK EPSC Projekt GEM; US NSF Projekte zu Sprachgenerierung) hervorragend für kooperative Forschung und Entwicklung ausgewiesen.

Prof. Dr. Otthein Herzog, seit 1993 Inhaber des Lehrstuhls für Informatik und Grundlagen der künstlichen Intelligenz und Expertensysteme im Fachbereich Mathematik und Informatik an der Universität Bremen, Sprecher des Technologie-Zentrums Informatik (TZI), Leiter der Arbeitsgruppen: Intelligente Systeme und Bildverarbeitung. Forschungsschwerpunkte: u. a. mobile, tragbare und räumlich universell einsetzbare EDV-Lösungen in Produktion, Logistik, Gesundheitswesen und First Responder, Multi-agent-Systeme zur Darstellung und Modellierung flexibler Arbeitsabläufe und zur Integration heterogener Datenquellen, insbesondere im Bereich Informationslogistik. Forschungsschwerpunkte im Bereich Bildverarbeitung ist die Analyse von Bildfolgen und hier insbesondere die Bewegungsanalyse und -interpretation durch qualitative und quantitative Methoden. Durch zahlreiche erfolgreich durchgeführte Projekte mit Partnern aus Forschung und Industrie (z.B. AVAnTA, ADViSOR oder AUTOZELL — aktuell: DfG-Projekt KonPro, EU-Projekt GAMA) hat das TZI hier eine breite Expertise.

Prof. Dr. Rainer Malaka, Lehrstuhl für digitale Medien im Fachbereich Informatik an der Universität Bremen. Schwerpunkte an diesem neu entstehenden Lehrstuhl sind Forschungsarbeiten zur virtuellen und augmentierten Realität, insbesondere Benutzerschnittstellen und intelligente mobile Systeme. Dazu wird zur Zeit erforscht, wie das automatisierte Erkennen von Alltagsgegenständen realisiert werden kann. Vor seiner Zeit in Bremen war Rainer Malaka seit 1997 Leiter einer Forschungsabteilung am European Media Laboratory (EML) in Heidelberg. Im Fokus seiner

Forschungstätigkeit dort waren mobile Assistenzsysteme. Sprachverstehen, Geografische Informationssysteme und Bildverarbeitung. Er entwickelte neue Verfahren, die es erlaubten, komplexe Objekte in Bildern zu erkennen. Dies wurde ermöglicht durch eine hierarchische Analyse der Objektbestandteile. Seine Forschungstätigkeit im Bereich maschinelles Bildverstehen wurde 2004 mit dem Forschungs- und Innovationspreis Rhein-Neckardreieck ausgezeichnet.

Prof. Dr. Marion G. Müller, Lehrstuhl für Mass Communication an der Jacobs Universität Bremen, hat eine Vielzahl von Veröffentlichungen in dem Spannungsgebiet zwischen visueller Kommunikation, Politik und Kultur veröffentlicht. Sie hat seit 10 Jahren bildanalytische Korpusarbeit betrieben und sammelt einen einzigartigen, motivisch strukturierten Pressebildkorpus. Wichtiger Teil ihrer Arbeit ist ein detailliertes Kategoriensystem für die Annotation und Analyse dieser Bilder. Im Mai diesen Jahres wurde Müller zum Chair der Visual Communication Studies Division der International Communication Association gewählt, der größten internationalen Gruppe von Forschern in visueller Kommunikation weltweit. Von 2004 bis 2007 war sie zudem Sprecherin der Fachgruppe Visuelle Kommunikation der Deutschen Gesellschaft für Publizistik und Kommunikationswissenschaft (DGPuK)

Kooperationen der Forschungs- und Praxispartner im Verbund

Auf naturwissenschaftlicher Seite sind bezüglich der automatischen Erkennung und Klassifizierung von Bildern und bezüglich der automatischen Filmsegmentierung und -analyse bereits gute Ergebnisse erzielt worden. Auf geisteswissenschaftlicher Seite haben neue Ansätze im semiotischen und linguistich fundierten Umgang mit bildlichen kommunikativen Artefakten die Forschungsrichtung in den letzten 10 Jahren neu belebt. Die Grenzen dieser Bestrebungen sind jedoch bereits erkennbar. Automatische Erkennung und Analyse kann nur gelingen, wenn ausreichend Wissen über die zu analysierenden Artefakte vorhanden ist; eine eher hermeneutisch und diskursiv orientierte Analyse kann jetzt nur weiterkommen, wenn ein effektiver Umgang mit größeren Datenmengen und gezielte Zugriffmechanismen auf passende Beispiele und repräsentive Datensätze vorhanden sind. Genau dies wird das vorliegende Vorhaben ermöglichen: ein neuartiger bidirektionaler Kanal wird eröffnet zwischen dem technischen Umgang mit multimodalen Daten und geistes-, kultur- sowie sozialwissenschaftlichen Forschungsmethoden und -ergebnissen.

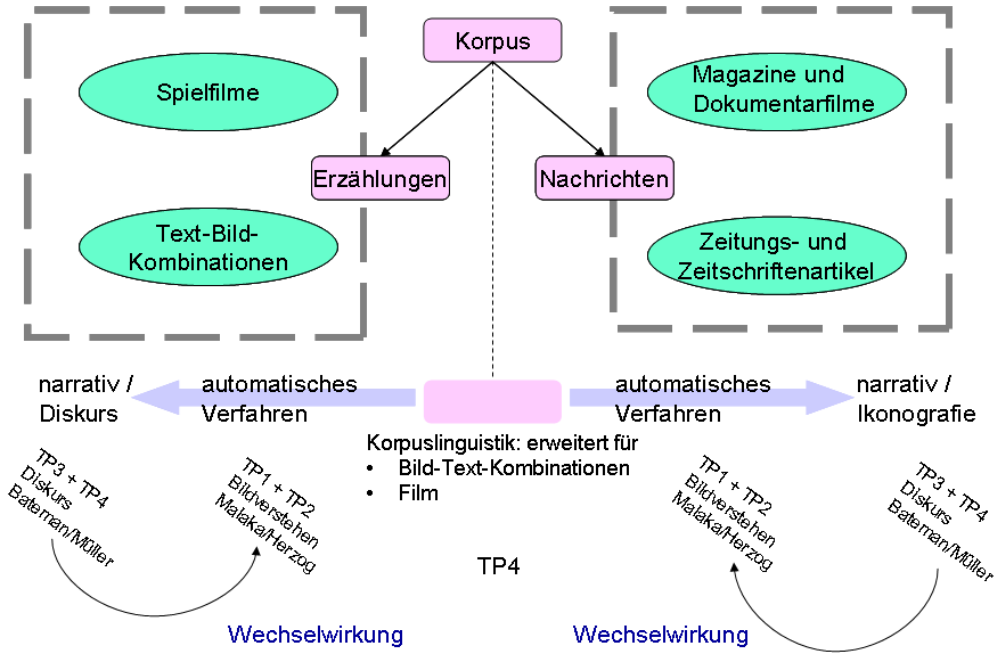
Dieser Kanal wird auf moderne Entwicklungen in der Sprachwissenschaft aufbauen. Erprobte Anwendungen von sogenannten Korpusmethoden in der Linguistik, die spezielle Lösungen für den Umgang mit linguistischen Daten anbieten, werden eingesetzt, um eine Kommunikation zwischen der natur- und der geisteswissenschaftlichen Seite zu ermöglichen. Korpusmethoden

haben sich in der Linguistik als besonders effektiv erwiesen. Der gezielte Umgang mit größeren Mengen von linguistischen Daten hat linguistische Theorie und Praxis, sowie die Zahl von Anwendungsmöglichkeiten, vorangetrieben und erweitert. Solche Methoden sind bisher nur ansatzweise in die geisteswissenschaftliche Forschung sowie in die Behandlung von Bildern, bewegten Bildern und Text-Bild-Kombinationen vorgedrungen. Dies soll jetzt in mehreren Projekten des Verbundes vorangetrieben werden. Eine multimodale Artefakte analysierende Linguistik fungiert bei allen Kooperationen als Mediator zwischen den beteiligten Disziplinen.

In den konkreten Forschungsprojekten des Verbundes wird gezeigt, dass der Einsatz naturwissenschaftlicher Methoden zur wissensbasierten Bildverarbeitung erhebliche Erleichterungen und qualitative Fortschritte für das geistes-, sozial- und kulturwissenschaftliche Verständnis von Bildern, bewegten Bildern und Bild-Text-Kombinationen und deren Funktion als Bedeutungsträger im gesellschaftlichen und persönlichen Umfeld mit sich bringt. Außerdem wird gezeigt, wie das nichtnaturwissenschaftliche Verständnis von Bedeutung zu erheblichen Verbesserungen für die technische und formale Modellierung sowie Analyse von solchen bildhaften und bebilderten Darstellungen führen kann.

Die Kooperation der Forschungspartner im Verbund beruht auf den einzelnen Kompetenzen der Partner und wird durch die linguistischen Korpusmethoden als das allgemeine Kommunikationsmedium vermittelt. Die Kooperation besteht in erster Linie aus einer gemeinsamen Ausarbeitung der methodischen und technischen Erfassung und Speicherung der Daten durch Linguisten und der Beteiligten in der Bildinterpretation und Filmanalyse, und einer gemeinsamen Ausarbeitung der Verarbeitungs- und Auswertungserfordernisse durch die Beteiligten aus den Geisteswissenschaften sowie der Ikonographie. Die Vermittlungsrolle der Linguistik und der Korpusmethoden wird im Laufe des Projekts dadurch ergänzt, dass direkte Ergebnisse in beide Richtungen fließen werden: von der automatischen Seite werden Probleme oder Erfolge in die vorgeschlagenen Klassifikationen einfließen und von der qualitativen Seite werden sich Änderungen oder Ergänzungen der Klassifikationen auf Grund von weiteren theoretischen und empirischen Auswertungen der durchgeführten Analysen entwickeln. Durch den wesentlich verbesserten Zugang zu automatisierter empirischer Auswertung der qualitativen aufgefassten Hypothesen für Klassifikationen sind nützliche und sonst gar nicht realisierbare Veränderungen in den literatur-, sprach- und kulturwissenschaftlichen Fragestellungen zu erwarten.

Die Abhängigkeiten und Beziehungen unter den Teilprojekten sind im detaillierten Arbeitsplan unten geschildert. Der gesamte Informationfluss im Vorhaben lässt sich wie in der folgenden Grafik darstellen.



Design und Methodik des Vorhabens

Jedes Teilprojekt des Vorhabens widmet sich einem Kernproblem, das eine wesentliche Herausforderung in den einzelnen beteiligten Disziplinen darstellt. Die Lösungen für diese Probleme werden durch eine eng verzahnte Zusammenarbeit der vier Teilprojekte erzielt. Insbesondere gibt es Abhängigkeiten zwischen TP1 und TP3, und zwischen TP2 und TP4. Darüber hinaus definieren wir eine weitere Reihe von Forschungsfragen, an der alle vier Teilprojekte beteiligt sind, um einen noch höheren Grad an Synergie zu erzielen.

Sowohl in Bezug auf Film als auch auf Bilder liegen bereits substantielle, aber meistens informelle hermeneutisch ausgerichtete Beschreibungsmethoden vor. Diese Beschreibungsmethoden sollen formal ausformuliert werden und in ein allgemeines Beschreibungsframework aus der multimodalen Linguistik eingesetzt werden. Dieses Beschreibungsframework wird dann als eine Wissensquelle in angereicherte automatische Verfahren in der Bildinterpretation und Filmanalyse importiert. Durch den Einsatz dieses Wissens sind wesentlich bessere und genauere Ergebnisse bei den automatischen Verfahren zu erwarten. Letztlich um den Kreis zu schließen, werden diese Verfahren wiederum der empirischen Verifikation und Verbesserung der ursprüngliche Beschreibungen zugänglich gemacht werden. Das Gesamtvorhaben wird so zum erstenmal einen Kreislauf ermöglichen, wobei Klassifikationen und analytische Methoden aus Geistes- und Sozialwissenschaften für Film, Bild und Text in automatischen Verfahren angewandt werden können, um damit eine empirisch besser verifizierte Qualität erzielen zu können.

TP1: Bildverstehen

Digitale Bilder sind ein wesentlicher Bestandteil heutiger Arbeitsumgebungen. Die Leichtigkeit, mit der sie produziert werden, führt im geschäftlichen und privaten Bereich zu riesigen Beständen, die verwaltet werden müssen. Während sie mit aktuellen Datenbanken problemlos gespeichert werden können, ist das Wiederauffinden (Retrieval) eine bisher unbefriedigend gelöste Aufgabe. Meist ist man an einem Konzept interessiert, was auf den gesuchten Bilden abgebildet ist. Das kann z.B. ein bestimmter Gegenstand, eine Person oder auch ein Vorgang sein. Abbildungen eines Konzeptes sind in ihrer Bildstruktur meist sehr unterschiedlich. Das liegt an Varianzen, wie unterschiedliche Beleuchtung, Perspektive oder Texturausprägungen. Eine wesentliche Schwierigkeit für das Erkennen von Gegenständen oder Vorgängen liegt darüber hinaus darin, dass diese oft ein völlig unterschiedliches Aussehen haben. Was solche Abbildungen eint, ist das vom Menschen zugeordnete Konzept, dass sich oft eher nach der Funktion eines Gegenstandes oder Vorgangs und nicht so sehr an dessen Aussehen orientiert. So findet man Abbildungen mit extrem unterschiedlichen Bildstrukturen zu Konzepten, wie z.B. Lastwagen, Palast oder Geburtstag. Bildunterschiede dieser Art sollen im Folgenden “abstrakte Varianzen” genannt werden. Bisherige Verfahren zum Bildverstehen sind kaum in der Lage mit abstrakten Varianzen umzugehen.

Um dennoch in Bildbeständen automatisiert suchen zu können, hat man sich bisher damit beholfen, die Bilder mit textuellen Kennzeichnungen (Tags) zu versehen (zu annotieren). Das bringt jedoch das Problem mit sich, dass die Kennzeichnungen manuell zugeordnet werden müssen und dass deren Bedeutungen für die typischen Selektionen der Anwendung geeignet sein müssen. Gewöhnlich ändern sich diese aber mit der Zeit. Für wachsende Bilddatenbanken ergibt sich z.B. regelmäßig der Wunsch, dass noch feiner unterscheidende Konzepte zur Abfrage zu Verfügung stehen sollen. Dazu muss der gesamte Bestand neu annotiert werden.

In diesem Teilprojekt sollen neue Techniken des Bildverstehens entwickelt werden, die mit klassischen Annotationstechniken kombiniert werden können und so erheblich zur Nutzerfreundlichkeit von Bilddatenbanken beitragen. Bei der Neuaufnahme eines Bildes sollen automatisch Vorschläge für passende Kennzeichnungen gemacht werden. Wird eine neue Kennzeichnung eingeführt, sollen aus dem Bildbestand automatisch mögliche Kandidaten vorgeschlagen werden, die diese Kennzeichnung auch tragen können. Ganz wesentlich verbessert werden soll die Suchtechnik, die nun auch bisher nicht gekennzeichnete Konzepte erkennen soll. Dazu sollen Kombinationen von Bildeigenschaften und bisherige Kennzeichnungen ausgewertet werden.

Die zu erstellende Architektur zum Bildverstehen soll Abbildungen Konzepte zuordnen können, selbst, wenn diese abstrakte Varianzen zeigen. Die Verarbeitung orientiert sich dabei am biologischen Vorbild. Dieses realisiert offenbar einen Verarbeitungspfad mit sehr vielen Stufen. Entlang des Pfades

beschreiben die Repräsentationen der Stufen immer komplexere (Teil-)Bildinhalte. Zusätzlich findet man, dass die Repräsentationen immer invarianter auf Varianzen einer Konzeptabbildung reagieren - sie also spezifischer auf das abstrakte Konzept werden (Tanaka et al. 1991). Abbildungsvarianzen von Konzepten sind im Allgemeinen sehr komplex. Um sie detektieren und repräsentieren zu können, werden offenbar effiziente Teilrepräsentationen eingesetzt, die flexibel zu einer Szenenbeschreibung kombiniert werden können (Oram & Perrett 1994).

Untersuchungen aus der biologischen visuellen Informationsverarbeitung lassen vermuten, dass Varianzen stets dort verarbeitet werden, wo Bildinhalte repräsentiert werden, die von der Varianz betroffen sind (Oram & Perrett 1994). Bisher wurde das meist so interpretiert, dass alle Informationen, die zur Detektion eines Teilkonzeptes beigetragen haben, für Folgestufen nicht mehr sichtbar sein müssen. Mit Varianzen dieser Signale muss nur in der Detektionsstufe umgegangen werden. Folgende Detektionsstufen bekommen nur noch die Information, dass das Teilkonzept gegeben ist - nicht aber dessen Ausprägung. Ob z.B. eine Uhr mit Zeigern oder Digitalanzeige ausgestattet ist, muss für das Konzept “Uhr” keine Rolle spielen. Diese Varianz kann auf der Stufe zur Erkennung der Gehäuse Frontansicht aufgelöst d.h. invariant gemacht werden. Aus dieser Interpretation wurden etliche biologisch motivierte Architekturen zum Bildverstehen erstellt. Sie verwenden eine Hierarchie von Teilkonzepten, die stets nur invariante Signale zur nächsthöheren Detektionsstufe weiterleiten (Fukushima 1975, Biederman 1987, LeCun et al. 1990, Olshausen et al. 1993, Wallis & Rolls 1997, Teichert & Malaka 2003). Das Erkennen und explizierte Repräsentieren von varianten Signalanteilen ist jedoch von großer Bedeutung für das Bildverstehen. Zum Beispiel ist es häufig wichtig zu wissen, ob die Ausprägung eines Teilkonzeptes innerhalb eines bestimmten Bereiches gegeben ist. Beim Erkennen einer Uhr ist es wesentlich, die Information zu erhalten, dass entweder Zeiger oder eine Digitalanzeige in der Frontansicht geben sind. Sonst wäre eine Armbanduhr beispielsweise nicht von einem Armreif zu unterscheiden. Generell kann die explizite Repräsentation von Varianzen für benachbarte Bildregionen hilfreich sein, wenn dort keine eindeutige Detektion möglich ist. Dies ist für Varianzen der Fall, die typischerweise regionale Ausbreitungen haben, wie z.B. perspektivische Verzerrungen, Texturausprägungen oder Stilausprägungen. Schließlich kann es wesentlich sein, eine bestimmte Ausprägung eines vorgeschalteten Teilkonzeptes zu erkennen. Wenn z.B. die Konzepte Digitaluhr und Analoguhr unterschieden werden sollen, müssen sowohl das übergeordnete Konzept “Uhr” als auch die Teilkonzepte “analoges Zifferblatt” und “digitales Zifferblatt” erkannt werden können. Die Detektion eines Teilkonzeptes kann sowohl von invarianten, als auch von varianten Signalen abhängen.

Für die zu erstellende Architektur sollen daher variante als auch invariante Signale expliziert repräsentiert und verarbeitet werden. Zur Kodierung von Bildstruktur soll ein geeignetes iteratives

Verfahren eingesetzt werden (Teichert & Malaka 2006). In Anlehnung an die biologische Verarbeitung soll damit versucht werden, einen Verarbeitungspfad zu realisieren, der stufenweise abstraktere Teilrepräsentationen detektiert.

In der ersten Phase des Teilprojekts sollen Konzepte von Alltagsgegenständen mit diesem Verfahren erkannt und unterschieden werden können. Die zu analysierenden Bilddaten werden primär aus dem Korpus von Pressebildern entnommen, die in TP3 bearbeitet werden. Die erste Stufe der Erkennung wird auf einer Modellierung des Weltwissens einer ausgewählten Testmenge der Daten basieren. In der zweiten Phase wird weitere Information aus dem 'Kontext' für die Analyse herangezogen. Diese Information ist als Text in den entsprechenden Presseartikeln zu finden. Die möglichen Ergänzungsmodi, die es erlauben, eine spezifischere Interpretation zu finden, wenn Bild und Bildunterschrift vorhanden sind, werden von TP3 importiert. Letztlich wird eine weitere Kategorieebene von TP3 erprobt werden, um die Möglichkeiten einer kulturellen Zuordnung und darauffolgenden Annotation von Bilddaten zu untersuchen und ggf. für die automatische Bildkorpusklassifizierung einzusetzen.

TP1: Bildverstehen – Arbeitspakete

In diesem Teilprojekt sollen Techniken entwickelt werden, die es ermöglichen, Konzepte in Nachrichtenbildern zu erkennen. Die Herausforderung liegt dabei darin, dass diese Bilder komplex sind und einer sehr hohen Ausprägungsvarianz unterliegen. Die Komplexität ergibt sich aus der Quantität der Bildstrukturen, die auszuwerten sind, um sie von Abbildungen anderer Konzepte abgrenzen zu können. Die Ausprägungsvarianz beschreibt die Vielfalt unterschiedlicher Anordnungen von Bildstrukturen, die für Abbildungen eines Konzeptes gegeben sein können.

Bisher ist es nicht möglich, komplexe Konzepte mit hoher Ausprägungsvarianz in Bildern automatisiert zu erkennen. Den meisten bisherigen Ansätzen gelingt ein Konzepterkennen nur, wenn starke Einschränkungen, bezüglich besonderer Eigenschaften, wie z.B. Hintergrund, Beleuchtung oder Objektausprägungen gemacht werden können. Daher sind diese Anwendungen meist spezialisiert auf die jeweilige Anwendung und können kaum auf andere Anwendungen übertragen werden. Die Arbeiten in diesem Teilprojekt sollen dazu beitragen, universellere Architekturen für das Bildverstehen zu finden und zu realisieren. Die Verarbeitung soll möglichst so erfolgen, dass beliebige Konzepte gelernt werden können. Das ist für die Analyse von Nachrichtenbildern auch notwendig, denn diese wird bestimmt von wechselnden Fragestellungen nach Bildinhalten.

Konzepte sollen anhand von Beispielbildern gelernt werden können, die die Konzepte in unterschiedlichen Ausprägungen zeigen. Dies soll für eine Reihe von Konzepten durchgeführt werden, die später in Bildern unterschieden werden sollen. Werden bisher nicht gelernte Bilder ausgewertet,

soll eine Zuordnung eines detektierten Konzeptes zu Bildregionen möglich sein. Als Ausgangsbasis für Beispiel- und Testbilder sind die in TP3 erstellten Datenbanken hervorragend geeignet, da sie bereits nach inhaltlichen Konzepten geordnet sind. Sie sind jedoch sehr komplex und zeigen eine extrem hohe Ausprägungsvarianz. Im Rahmen eines ersten Arbeitspakets wird ein Arbeitsablauf definiert, bei dem die Bilder so angepasst werden, dass sie einerseits nicht zu komplex sind und anderseits ausreichend Komplexität besitzen, um einen relevanten Forschungsschritt leisten zu können. Das Gleiche trifft auch auf die Ausprägungsvarianz zu.

— Arbeitspakete (TP1) —

Arbeitspaket	*Kurztitel*	*Jahr 1*			M1	*Jahr 2*			M2	*Jahr 3*			M3
TP1.1 a,b	**Erstellung einer Wissensbasis zum Konzeptlernen in Bildern**	XX X	X										
TP1.2	**Identifikation relevanter automatischer Analyseverfahren und Evaluation (Bild)**		X	XX									
TP1.3	**Erstellung eines Verfahrens zum Bildverstehen**			XX	XXX								
TP1.4	**Erstellung einer erweiterten Wissensbasis (Bild)**					XX							
TP1.5	**Implementierung und Evaluierung der Architektur zum Bildverstehen**					X	XXX	XXX					
TP1.6	**Identifikation und Formalisierung von Diskurswissen**								XXX				
TP1.7	**Verbesserung der Architektur zum maschinellen Bildverstehen**									XXX	X		
TP1.8	**Evaluation und Anwendung der adaptierten Architektur**										XX	XXX	
TP1.9	**Evaluierung: Generalisierbarkeit der Fragestellung (Bild)**												XXX

TP1.1 Erstellung einer Wissensbasis zum Konzeptlernen in Bildern

a) Erstellung einer Wissensbasis zum Konzeptlernen in Bildern. Zusammen mit TP3 (TP3.1–TP3.2) wird bestimmt, welche Konzepte erkannt werden sollen. Dabei soll eine angemessene Komplexität und Ausprägungsvarianz in den zugehörigen Bildern gegeben sein, die es ermöglichen, gegebene Architekturen bezüglich ihrer Erkennungsleistungen zu beurteilen. (2PM)

b) Definition eines Arbeitsablaufes, der es ermöglicht, Bilder aus der Sammlung des TP3 so anzupassen, dass sie zum Lernen und Testen von Verfahren zum Bildverstehen geeignet sind. Insbesondere muss ein Übergabeformat erstellt werden, welches ermöglicht, Konzepte konkreten Bildregionen zuzuordnen. (2PM)

TP1.2 Identifikation relevanter automatischer Analyseverfahren und Evaluation auf initialer Wissensbasis

Es soll untersucht werden, welche bestehenden Architekturen zum Bildverstehen einen guten Ausgangspunkt für die Bewältigung der beschriebenen Anforderungen bieten. Diese Architekturen sind hinsichtlich ihrer Erkennungseigenschaften bezüglich der Bildzusammen-

stellung aus Arbeitspaket TP1.1 zu bewerten. Für geeignete Kandidaten sind Evaluationen durchzuführen. (3PM)

TP1.3 Erstellung eines Verfahrens zum Bildverstehen

Für die im vorigen Arbeitspaket identifizierten Verfahren zum Bildverstehen soll untersucht werden, inwieweit Erweiterungen realisiert werden können, die ein universelles Konzepterkennen ermöglichen. Gegebenenfalls sind geeignete Teilarchitekturen zu integrieren und für fehlende Eigenschaften neue Funktionalitäten zu entwickeln. (5PM)

Meilenstein 1: Architekturdefinition (M12). Zum ersten Meilenstein steht eine umfassende Architekturdefinition zum Konzepterkennen bereit.

TP1.4 Erstellung einer erweiterten Wissensbasis

Für die neu zu erstellende Architektur zum Bildverstehen sollen weitere Musterpaare aus Konzept und Beispielbild zusammengestellt werden, damit umfangreichere Evaluationen durchgeführt werden können. (2PM)

TP1.5 Implementierung und Evaluierung der Architektur zum Bildverstehen

Die in Meilenstein 1 definierte Architektur zum Bildverstehen soll implementiert und evaluiert werden. Dabei soll die erweiterte Wissensbasis aus Arbeitspaket TP1.4 Verwendung finden. (7PM)

TP1.6 Identifikation und Formalisierung von Diskurswissen enthalten in entwickelten/-existierenden automatischen Verfahren

Für das Teilprojekt TP3 sollen die detektierten Konzepte nutzbar gemacht werden. Dazu sollen in diesem Arbeitspaket die notwendigen Exportfunktionen und Übergabeformate definiert und implementiert werden. (3PM)

Meilenstein 2: Architekturevaluation (M24). Zum zweiten Meilenstein steht die im ersten Meilenstein definierte Architektur bereit. Ihre Eigenschaften sind evaluiert worden und Detektionsergebnisse können über eine Schnittstelle zum TP3 weitergeleitet werden.

TP1.7 Verbesserung der Architektur zum maschinellen Bildverstehen anhand von Ergebnissen der Diskursanalyse

Es soll untersucht werden, wie Ergebnisse der Diskursanalyse aus TP4 für den Konzepterkennungsvorgang gewinnbringend eingesetzt werden können. Entsprechende Architekturerweiterungen sind zu definieren. (4PM)

TP1.8 Evaluation und Anwendung der adaptierten Architektur

Die in Arbeitspaket TP1.7 definierte Architektur zum Bildverstehen soll implementiert und evaluiert werden. Dabei findet wieder die in Arbeitspaket TP1.4 definierte erweiterte Wissensbasis Verwendung. (5PM)

TP1.9 Evaluierung: Generalisierbarkeit der Fragestellung auf neues, nicht betrachtetes Material

Die Architektur soll auf neuen Bildersätzen überprüft werden. Diese Bilder sollen nur bisher nicht verwendete Konzepte zeigen. Es soll evaluiert werden, wie die Architektur auf völlig neue Konzepte angewendet werden kann. (3PM)

Meilenstein 3 und Projektende (M36). Vollständiges integriertes System: Bildaspekte

TP2: Filminterpretation

Die Bedeutung von Videos als Medien in digitalen Bibliotheken nimmt immer weiter zu. Ein sinnvolles und erfolgversprechendes Retrieval großer digitaler Bild- und Videoarchive setzt eine systematische Erschließung in Form einer inhaltlichen Annotation und Strukturierung der im Archiv enthaltenen Dokumente voraus. Die systematische Erschließung insbesondere von Videodokumenten stellt eine extrem zeitintensive Aufgabe dar, da die rein manuelle Annotation eines einstündigen Filmbeitrages bis zu acht Stunden in Anspruch nimmt.[1] Methoden aus dem Bereich der inhaltsbezogenen Bildanalyse und der Künstlichen Intelligenz können sowohl die Archivare bei der inhaltlichen Erschließung von Bildern und Videodokumenten unterstützen, als auch eine umfangreiche Recherche auf Seiten der Benutzer der Bild- bzw. Videobibliothek signifikant erleichtern Hermes et al. (1999), Miene & Herzog (2000), Miene, Hermes & Ioannidis (2001), Christel et al. (2001).

Die Formulierung von Anfragen wird vereinfacht und die Suchergebnisse können direkt am eigenen Arbeitsplatz visualisiert werden. Ein besonders schwieriges, aber zugleich äußerst interessantes Forschungsgebiet besteht dabei in der Entwicklung von Verfahren zur automatischen inhaltlichen Analyse und Strukturierung der Videodokumente. In diesem Teilprojekt wird eine Architektur für die automatische inhaltsbezogene Strukturierung von Filmen entworfen und evaluiert werden.

Grundsätzlich werden zwei Anwendungsszenarien bzw. - phasen unterschieden: die Archivierungs- bzw. Annotationsphase und die Retrievalphase. Eine Annotation stellt eine Beschreibung des Inhaltes des annotierten Dokumentes dar. Das Vorliegen dieser Informationen in textueller Form bzw. in Form von Merkmalsvektoren ist für eine akzeptable Antwortzeit beim Retrieval unverzichtbar, da ein direkter Bildzugriff bei der hohen Anzahl von Einzelbildern nicht ausreichend effizient ist. Bei

[1] Diese Zeiten wurden im Fernseharchiv von Radio Bremen gemessen.

Videodokumenten besteht die Annotation sowohl aus syntaktischen Informationen z.B. bezüglich der Schnitte, der Kameraführung, etc. als auch aus Informationen über den Inhalt der Videosequenz. Diese Annotationen können manuell erzeugt werden, d.h. ein Archivar oder ein Benutzer betrachtet das Bild bzw. die Videosequenz und erfasst alle relevanten Informationen in textueller Form. Durch Methoden der automatischen Videoanalyse kann der Dokumentar in vielen Bereichen bei der Annotationsgenerierung unterstützt werden. Hierzu gehören z.B. die automatische Analyse von Kameraschnitten und der Kamerabewegung (Deardorff et al. 1994, Yusoff et al. 1998, Lienhart 1999, Miene, Dammeyer, Hermes & Herzog 2001). Um dem Archivar einen schnellen Überblick über den Bildinhalt des Videos zu verschaffen, kann das Video mittels einer geeigneten Menge von Einzelbildern repräsentiert werden. Hierzu kommen sowohl key frames als auch Mosaicbilder in Frage (Flickner et al. 1995, Mann & Picard 1995, Smolic et al. 1999). Einzelbilder können hinsichtlich Farb-, Textur- und Konturmerkmalen untersucht werden, um beispielsweise menschliche Gesichter oder Logos automatisch zu erkennen. Der Benutzer hat anschließend die Möglichkeit, die automatisch erstellte Annotation manuell zu ergänzen. Bei dieser Form der automatischen Annotation handelt es sich zunächst um rein syntaktische Informationen. Für eine Analyse von Einzelbildern auf semantischer Ebene wurden erste Ergebnisse und ein System in Hermes et al. (1995) vorgestellt, worauf wir in diesem Projekt weiter bauen werden.

Der Retrievalprozeß wird maßgeblich bestimmt durch das Format, über den das Retrieval durchgeführt wird, und durch die Sprache, in der die Anfragen formuliert werden. Das Ergebnis einer Anfrage ist eine Menge bzw. eine Liste von Bildern oder Videosequenzen. Oftmals sind die Ergebnisse nach ihrer Relevanz, d.h. ihrer Ähnlichkeit zu der Anfrage sortiert. In dieser Form werden auch Ergebnisse an Teilprojekt TP4 weitergegeben werden.

Die Annotations- und die Retrievalphase sind notwendige Bestandteile jedes Ansatzes zum Bild- und Videoretrieval. Darüber hinaus werden wir uns vor allen mit der wissenschaftlichen Herausforderung der *Event Detection* auseinandersetzen. Szenen und/oder Event Detektion kann in zwei Klassen unterteilt werden. Zum einen in *modellbasierte* Ansätze und zum anderen in *modellfreie* Ansätze. Modellbasierte Ansätze modellieren oftmals die Struktur des Videos (der Sendung). Die zeitliche Struktur kann als Abfolge von bestimmten Shots – wie Ansager (Sprecher), Werbeunterbrechungen, Wetter, usw. – modelliert werden. Die räumliche Struktur bei TV Nachrichtensendungen zum Beispiel wurde in Zhang et al. (1995) als ein Vier-Framemuster modelliert, wobei dieses Muster entweder einen Sprecher oder zwei Sprecher oder ein Sprecher mit einem Nachrichten-Icon (oben rechts) oder einen Sprecher mit einem Nachrichten-Icon (oben links) aufzuweisen hat.

Modellbasierte Ansätze können oftmals gute Klassifikationsraten erzielen, aber nachteilig ist, dass die für eine bestimmte Anwendung entwickelten Modelle bei anderen Anwendungen gar

nicht oder deutlich schlechter funktionieren. Ein Modell, das für Fußball entwickelt wurde, wird augenscheinlich bei Basketball oder beim Wasserball nicht funktionieren. Des Weiteren ist die Modellentwicklung in der Regel aufwändig. Modellfreie Ansätze hingegen können generisch auf Videos angewendet werden. Solche Ansätze lassen sich in drei Klassen einteilen: (i) Ansätze, die auf *visuellen* Hinweisen/Merkmalen basieren beziehungsweise mit jenen arbeiten; (ii) Ansätze, die auf *audio* Hinweisen/Merkmalen basieren beziehungsweise mit jenen arbeiten, und (iii) hybride Ansätze. die auf so genannten *audio-visuellen* Hinweisen/Merkmalen basieren beziehungsweise mit jenen arbeiten.

Ansätze, die auf visuellen Merkmalen basieren, nutzen sehr häufig Farb- und/oder Bewegungsinformation, um zum Beispiel Shots zu gruppieren, die ähnlich zueinander sind. Einen etwas anderen Weg gehen Aigrain et al. (1995), die heuristische Regeln verwenden, die lokale Eigenschaften wie Transitionseffekte und Shot-Wiederholungs-Muster berücksichtigen. Damit erzielen sie eine Art Makro-Segmentierung des Videos; eine auf linguistischer Grundlage basierende Verfeinerung dieser Richtung wird in TP4 unternommen. Rein auf Audio basierende Ansätze zur Eventsegmentierung lassen sich zum Beispiel in Zhang & Kuo (1999) oder auch in Moncrieff et al. (2001) finden. Erstere unterteilen das Video aufgrund von "'low-level"' Audiomerkmalen (zum Beispiel Frequenzen) in einzelnen Sequenzen von "'Audio-Szenen"'. Diese Sequenzen können zum Beispiel Sprach- oder Musiksequenzen sein.

In der ersten Phase des Teilprojekts ist das Hauptziel unsere Architektur so zu verfeinern, dass ein zwei-stufiges Bearbeitungsverfahren unterstützt wird: beide modellfreie und modellbasierte Methoden werden integriert. Der modellfreie Ansatz wird die erste Segmentierung durchführen, um die Anzahl von möglichen Interpretationen von vornherein einzugrenzen. Ein einfaches Modell des notwendigen Weltwissens für eine ausgewählte Menge von Filmsequenzen wird durch Standard-Wissensrepräsentationsmethoden erstellt, um die modellbasierte Stufe zu unterstützen. Ein hoher Grad an Überschneidung mit dem alltäglichen Wissen, das für TP1 benötigt wird, wird vorausgesetzt. Diese erste Version der Architektur und ihre Implementierung wird gegen eine weitere Menge von ausgewählten Filmsequenzen evaluiert, um eine "Baseline-Performanz" des Verfahrens zu dokumentieren.

In der zweiten Phase des Teilprojekts werden Erweiterungen in der modellfreien Stufe unternommen werden. Makrosegmentierung wird mit Hilfe von weiteren in TP4 ausgearbeiteten Segmentierungsschemen vorgenommen. Hierbei werden die sogenannten *Konstruktionen* von TP4 als Segmentierungsschemen formalisiert und validiert. Eine weitere Evaluation des Gesamtsytems wird dann Verbesserungen und Probleme im Vergleich zu der ersten Version feststellen und dokumentieren.

— Arbeitspakete (TP2) —

Arbeitspaket	*Kurztitel*	*Jahr 1*			**M1**	*Jahr 2*			**M2**	*Jahr 3*			**M3**
TP2.1 a,b	**Erstellung einer Wissensbasis für Filmmaterial**	XX X	X										
TP2.2	**Identifikation relevanter automatischer Analyseverfahren und Evaluation (Film)**		XX	X									
TP2.3	**Definition der Architektur zur Event Detektion und Integration Verfahren**			XX	XXX								
TP2.4	**Erweiterung der Wissensbasis (Film)**					XX							
TP2.5	**Implementierung und Test der Architektur zum Filmverstehen**					X	XXX	XXX					
TP2.6	**Identifikation und Formalisierung von im System enthaltenen Diskurswissen**								XXX				
TP2.7	**Adaptierung der Architektur anhand von Ergebnisse der Diskursanalyse**									XXX	X		
TP2.8	**Implementierung und Text der adaptierten Architektur**										XX	XXX	
TP2.9	**Evaluierung: Generalisierbarkeit der Fragestellung (Film)**												XXX

TP2.1 Erstellung einer initialen Wissensbasis für Filmmaterial

a) In Zusammenarbeit mit TP4 wird relevantes Filmmaterial aus den Bereichen Nachrichten und Narrativ gesammelt und für das Projekt zugreifbar gemacht. Die Zusammenstellung richtet sich dabei nach der Art der Szenen und Events, die später automatisch erkannt werden sollen. (2PM)

b) Bereitstellen von Werkzeugen zur Annotation von Videodaten. Diese Werkzeuge werden benötigt, um eine sog. Ground Truth auf der Wissensbasis zu erstellen, mit deren Hilfe die zu entwickelnden automatischen Verfahren evaluiert und später ggf. auch trainiert werden können. (2PM)

TP2.2 Identifikation relevanter automatischer Analyseverfahren und Evaluation auf initialer Wissensbasis

Existierende Verfahren zur Event Detektion werden hinsichtlich ihrer Eignung für die hier betrachteten Szenen/Events untersucht. Hierzu wird die initiale Wissensbasis und die darauf erstellte Ground Truth herangezogen. Die betrachteten Verfahren gliedern sich in vier Bereiche: Modellbasierte bzw. modellfreie Verfahren jeweils in den Bereichen Nachrichten bzw. Narrativ. (3PM)

TP2.3 Definition der Architektur eines Systems zur Event Detektion und Integration selektierter relevanter Verfahren

Auf Basis der vorangegangenen Evaluation werden geeignete Verfahren selektiert und in einem Gesamtsystem integriert. Gegegenfalls werden Erweiterungen und, wenn möglich, Verbesserungen vorgesehen. Das Ergebnis ist der erste Meilenstein des TP2, die Architekturdefinition. (5PM)

Meilenstein 1: Architekturdefinition. Ergebnisse aus Arbeitspaket TP2.3.

TP2.4 Erweiterung der Wissensbasis

Die bereits bestehende Wissensbasis und die zugehörige Ground Truth werden um weitere Beispiele erweitert, um umfangreichere Evaluationen zu ermöglichen und ggf. eine angemessenere Aufteilung in Trainings- und Testset zu erhalten. (2PM)

TP2.5 Implementierung und Test der Architektur Die in Meilenstein 1 definierte Architektur wird implementiert und auf der erweiterten Wissensbasis evaluiert. (7PM)

TP2.6 Identifikation und Formalisierung von im entwickelten System enthaltenen Diskurswissen

Das noch implizit in dem entwickelten System enthaltene Diskurswissen soll formalisiert und so für das TP4 nutzbar gemacht werden. Auf Basis der formalen Definition werden Schnittstellen implementiert, über die Ergebnisse der automatischen Analyse mit TP4 (insb. TP4.5 u. TP4.6) ausgetauscht werden können. (3PM)

Meilenstein 2: Architekturevaluation Zum zweiten Meilenstein steht die im ersten Meilenstein definierte Architektur bereit. Ihre Eigenschaften sind evaluiert worden und Ergebnisse der automatischen Analyse können über eine Schnittstelle zum TP4 weitergeleitet werden.

TP2.7 Adaptierung der entwickelten Architektur anhand von Ergebnissen der Diskursanalyse

Aus den im TP4 entwickelten Konstruktionen (insb. TP4.5 und TP4.8) werden mögl. Verbesserungen der bestehenden Architektur abgeleitet. Diese können zum einen neue Modelle für die modellbasierten Ansätze beinhalten, aber auch neue modellfreie Verfahren. (4PM)

TP2.8 Implementierung und Test der adaptierten Architektur

Die in Meilenstein 1 definierte Architektur wird implementiert und auf der in Arbeitspaket TP2.4 definierten Wissensbasis evaluiert. (5PM)

TP2.9 Evaluierung: Generalisierbarkeit der Fragestellung auf neues, noch nicht betrachtetes Material

Die Architektur wird auf ihre Generalisierbarkeit geprüft. Hierbei soll anhand neuen Filmmaterials evaluiert werden, wie gut die entwickelte Architektur auf ursprünglich nicht berücksichtigte Arten von Szenen/Events anwendbar ist. (3PM)

Meilenstein 3 und Projektende (M36). Vollständiges integriertes System: Filmaspekte

TP3: Bild-Text in der Presse und Nachrichten

Die Gruppe der Antragsteller geht davon aus, dass visuelle Muster der menschlichen Perzeption und Interaktion sowie des Informationsaustausches ein in der Zukunft zunehmend relevantes Phänomen sind. Wahrnehmung, zwischenmenschliche und besonders massenmediale Kommunikation, vollziehen sich mehr und mehr in bildhafter Form. Zeitungen und Zeitschriften drucken häufiger Bilder ab, um ihre Attraktivität zu steigern. Das einstige prototypische Textmedium nimmt immer stärker eine visuelle Form an und verändert so durch die Kombination mit Bildern seinen Charakter. Dieser Wandel in der Nachrichtenvermittlung in zunehmend multimodaler Form wird im Vergleich von traditionellen Print- mit Online-Zeitungen besonders deutlich (Bateman et al. 2007, Knox 2007, de Vries 2008).

In diesem Teilprojekt untersuchen wir insbesondere Nachrichtenbilder im Kontext der Print- und Online-Berichterstattung, ihre Be- und Untertitelung sowie das Wechselverhältnis zwischen Bild und Text. Ziel ist die Bildung einer Typologie von Bildmotiven (Müller 1997, 2006, 2007) einerseits und von interpretativ erzeugten Sinnzusammenhängen andererseits. Dabei untersuchen wir die Annotation von Pressebildern (Fotografie, Karikatur, Infografiken) mit einem detaillierten und breitgefächerten System von kulturwissenschaftlichen Kategorien, das die herkömmlichen Annotationen in neuer Weise ergänzen soll und sich für sprach- und kulturwissenschaftliche Untersuchungen besonders eignet. Alle genannten Aspekte fügen dem Bildverstehen und der Interpretation von Bild-Text-Kombinationen eine wichtige kulturell-ideologische Dimension hinzu (Kress & van Leeuwen 1996, Müller 2003), die eine weitere Stufe für die Verarbeitung von TP1 anbietet. Ein besseres Verständnis für das Wechselverhältnis von Bild, Text und sozial-kulturellem Kontext wird auch ermöglichen, Aussagen über die Verschiebung von Interpretationen zu machen, sobald Nachrichtenbilder “enttextualisiert” oder “entkontextualisiert” präsentiert werden (Müller & Özcan 2007). Die Interpretation von Bildern anhand solcher Kategorien hat eine lange Tradition in verschiedenen Disziplinen. Bildinterpretation in diesen disziplinären Kontexten ist jedoch häufig idiosynkratisch und auf das konkrete Forschungsprojekt bezogen. Meist gelingt keine Übertragung der Bildkategorien auf andere Bildkorpora, da die hermeneutisch generierten Kategorien zu sehr mit dem beschriebenen Material verhaftet sind. In diesem Projekt versuchen wir diese hermeneutisch inspirierten Interpretationsmuster und Kategorisierungsschemata mit generalisierbaren, visio-perzeptiven Merkmalen zu verbinden, auf denen die eigentliche Interpretationsleistung basiert. Dies bedeutet, dass einerseits ein Korpus von Nachrichtenbildern in Printmedien kategorial erschlossen wird, andererseits die Kategorien im Hinblick auf ihre mögliche automatische Erkennbarkeit weiterentwickelt werden. Auf dem breiten Kategorisierungsschema von Müller aufbauend, wird in diesem Teilprojekt eine Untergruppe an Kategorien ausgewählt und analysiert, um mögliche automatisch erkennbare visuelle Merkmale

zu finden, die wiederum eine Anreicherung für die in TP 1 ausgearbeiteten Annotationsmöglichkeiten darstellen. Die Annotationstechnik soll im Laufe des Projektes durch ein von TP 1 entwickeltes visuelles Annotationstool ergänzt werden, das Bildmarkierungen auf den ausgewählten Nachrichtenbildern ermöglicht, die wiederum automatisch erkannt werden können. Das Teilprojekt wird zum einen Annotationskategorien entwickeln, die bestimmte häufig vorkommende Motivtypen der Pressefotografie, aber auch anderer, grafischer Bildelemente der Printberichterstattung typologisch erfassen. Diese Kategorienbildung erfolgt in Hinblick auf ihre zukünftige automatische Erfassbarkeit mit den durch TP 1 entwickelten Tools. Zum anderen untersucht TP 3 die interpretative Erschließung von Bild-Text-Bedeutungen sowohl hinsichtlich ausgewählter Print-Nachrichtenmedien als auch hinsichtlich ihrer Online-Varianten. Dafür werden erstmalig zwei Methoden zusammengeführt: Das von Bateman et al. (2004) entwickelte 'multi-layered' linguistische Annotationsframework für die Beschreibung und Analyse von Seitenlayout und rhetorischer Organisation und das von Müller (2003) entwickelte ikonologische Verfahren der visuellen Kontextanalyse. Das von Bateman et al. entwickelte Framework verfolgt eine linguistische Korpusmethodologie, um Bild-Text-Seiten mit struktureller Information für eine empirische Analyse anzureichern. Darüber hinaus ist eine detaillierte Klassifizierung von möglichen Bild-Text-Beziehungen auf semiotisch-linguistischer Basis von Martinec & Salway (2005) entwickelt worden. Wir werden diese Klassifizierung auch in unsere Analyse einbeziehen, um festzustellen: (a) ob solche Beziehungen automatisch erkennbar sind (was Martinec und Salway postulieren, aber nicht beweisen) und (b) ob spezifische Arten von Beziehungen die Interpretation der Bilder in besonderer Art und Weise beeinflussen. Dieses semiotisch-linguistische Vorgehen wird ergänzt um die ikonologische Bildmotivanalyse, die sich weniger auf den das Bild umgebenden Text und stärker auf die historischen Motivtraditionen sowie die sozial-kulturellen Kontextbedingungen bezieht. Dabei spielt der spezifische journalistische Produktionskontext der untersuchten Nachrichtenbilder eine wichtige Rolle. Selektion und Platzierung von Pressebildern, Karikaturen und Infografiken sind das Ergebnis von Routineentscheidungen, die zu einer Auswahl ganz bestimmter Typen von Bildern führen. Diese Bildkontexte können nur durch eine empirische Untersuchung dieser journalistischen Produktionsroutinen erfolgen. Die qualitativen Leitfadeninterviews mit Pressefotografen, Bildagenturen und Zeitungsredakteuren (Print und Online) verfeinern die Kategorienbildung und ergänzen das semiotisch-linguistische Vorgehen um einen Vergleich mit den Kriterien, die in der Produktionspraxis angelegt werden. Im Rahmen dieser abschließenden Studie erfolgt auch ein Validitätstest des entwickelten Kategorien- und Annotationssystems.

— Arbeitspakete (TP3) —

Arbeitspaket	*Kurztitel*	*Jahr 1*			**M1**	*Jahr 2*			**M2**	*Jahr 3*			**M3**
TP3.1	**Erstellung einer initialen Materialbasis: PIAV**	XX											
TP3.2 a,b	**Erstellung einer Materialbasis: (a) neuer Print-Korpus (b) neuer Online-Korpus**	X	XXX	XX X	XXX								
TP3.3	**Methodische Integration von ‚multi-layered' Annotation**					X							
TP3.4	**(Weiter)Entwicklung von Kategorien, Annotationen, Kodierbogen**					XX							
TP3.5	**Annotation/Kodierung alter Korpus und neue Korpora**						XXX	XXX					
TP3.6	**Produktionsanalyse: Leitfaden-Interviews**								XXX				
TP3.7	**Auswertung Annotation, Kodierbögen, Interviews**									XXX	XXX		
TP3.8	**Validitätstest Prototyp**											XXX	
TP3.9	**Ergebnisauswertung und –kompilation**												XXX

TP3.1 Erstellung einer initialen Materialbasis: Alter Korpus

Auswahl und Erfassung von Motivgruppen aus dem bereits digitalisierten Bestand Archiv PIAV (Müller/Jacobs University Bremen). Lieferung von Bildmaterial an TP1 (TP1.2 und TP1.3) für Entwicklung des Annotationstools. (2 PM)

TP3.2 Erstellung einer initialen Materialbasis: Neuer Korpus

a) Neuer Korpus Print:Auswahl, Scannen und Erfassen von Bildern und Texten aus Printausgaben aktueller Zeitungen. Anwendung des Annotationstools aus TP1. (6 PM, 60 StuMi)

b) Neuer Korpus Online: Auswahl und Erfassung von Bildern und Texten aus Online-Zeitungen. Feedback und erste Evaluierung des Annotationstools an TP1. (4 PM, 40 StuMi)

Meilenstein 1 (M12). Neuer Annotationskatalog.

TP3.3 Inhaltsanalyse Bild-Text in Presse und Nachrichten

Methodische Integration von ‘multi-layered’ linguistischem Annotationsframework und visueller Kontextanalyse. Import des ‘multi-layered’ Annotationsframeworks aus TP4 (insb. TP4.3 u. TP4.4). (1 PM)

TP3.4 (Weiter)Entwicklung von Kategorien, Annotationen, Kodierbogen

Zusammenstellung eines integrierten linguistisch-visuellen Annotationsframeworks. Erweitung der notwendigen Annotationskategorien und Strukturen als Export in TP4. (2 PM)

TP3.5 Annotation/Kodierung

Der alte Korpus und die neuen Korpora werden annotiert unter Verwendung der erweiterten Kategorien aus TP3.4. (6 PM, 100 StuMi)

TP3.6 Produktionsanalyse

Leitfaden-Interviews werden durchgeführt und ausgewertet. (3 PM)

Meilenstein 2 (M24). Beitrag zur Systemevaluierung.

TP3.7 Empirische Studie: Produktionsanalyse und Validitätstest

Auswertung von Annotation, Kodierbögen, Interviews. Export in TP1 und in TP4 für die Auswertung der empirischen Ergebnisse der Inhalts- und Produktionsanalyse. (6 PM, 100 StuMi)

TP3.8 Validitätstest Prototyp

Feedback an TP1 bzgl. der laufenden Tests. (3PM)

TP3.9 Ergebnisauswertung und -kompilation Fertigstellung von Berichten und Veröffentlichungen. (3 PM)

Meilenstein 3 und Projektende (M36). Vollständiges integrietes System: Bild- und Bild-Text-Aspekte.

TP4: Filmnarrativ und Erzählungen, die Diskursstruktur des Films und des Bildes

Über den unter Teilprojekt TP2 beschriebenen Stand der Forschung in automatischer Filmsegmentierung hinaus gibt es Forschungsrichtungen, die sich mit Filmstrukturen und Filmstrukturierungsmechanismen aus der Filmsemiotik (Metz 1974) und Filmkritik (Bordwell 1996) beschäftigen. In letzter Zeit haben diese Richtungen immer intensiver versucht, frühere Mängel bzgl. der Genauigkeit und Reproduzierbarkeit solcher Analysen dadurch zu beseitigen, dass chronologische und räumliche Eigenschaften der analysierten Filme herangezogen werden (cf. Schmidt & Strauch 2002). Diese Untersuchungen sind aber immer noch eingeschränkt, indem viele aus linguistischem Diskurs bekannte Eigenschaften noch nicht für Film geltend gemacht worden sind. In diesem Teilprojekt unternehmen wir diesen wichtigen weiteren Schritt.

Wie ursprünglich von Metz vorgeschlagen, ist es möglich größere Filmstrukturen zu definieren, die besondere zusätzliche und diskurs-bezogene Bedeutung zu Sequenzen von filmischen Einstellungen geben. Diese Bedeutungen schließen typischerweise Begriffe wie Alternationen, chronologische Folgen, illustrative Beispiele usw. ein, die wesentliche Interpretationseinschränkungen für die einzelnen Einstellungen, woraus die Sequenzen aufgebaut sind, implizieren. Die erste Arbeitsphase in diesem Teilprojekt wird dementsprechend die Verwandtschaft zwischen den in dieser Tradition vorgeschlagenen Einschränkungen und den Eigenschaften für automatische Interpretation, die in

TP2 untersucht werden, festlegen. Die darauffolgende Phase ist dann eine Untersuchung der von Bateman (2007) vorgeschlagenen Erweiterung des Metzschen Ansatzes, sowohl theoretisch als auch in Bezug auf konkrete empirische Beschreibungen von ausgewählten narrativen Filmen. Diese Erweiterung hat zwei Hauptanliegen. Erstens wurde in Batemans Ansatz ein wesentlicher Ausbau der möglichen Beziehungen zwischen filmischen Einstellungen unternommen; die vorhandene Menge von Beziehungen ist jetzt die bisher vollständigste ihrer Art. Zweitens wurde die unabdingbare Rolle von zusätzlichen filmischen Strukturen, *Konstruktionen* genannt, zum ersten Mal deutlich. Solche Strukturen sind analog zu ähnlichen Begriffen in linguistischen Ansätzen wie der Konstruktionsgrammatik (Goldberg 1995) zu sehen, d.h., als zusätzliche Verankerungsmöglichkeiten für Bedeutung.

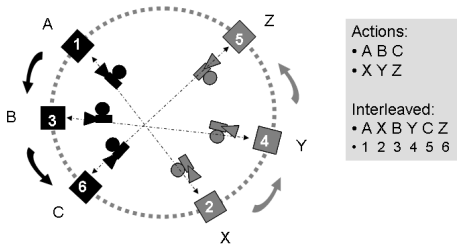
Als ein einfaches Bespiel können wir die für diesen Zweck häufig verwendete Phrase aus dem Englischen nehmen: 'kick the bucket'. Diese Phrase ist als eine 'halb-feste' linguistische Struktur zu analysieren, die eine andere Bedeutung als die durch kompositionelle Zusammenfügung ihrer Teile, besitzt. Genauso finden wir Konstruktionen im Film. Die filmisch bekannte Sequenz *shot/reverse-shot* bedeutet konventionellerweise–wenn mit zwei Sprechern verwendet–dass die Sprecher miteinander reden und wir als Zuschauer den Dialog betrachten. Wichtig ist dabei, dass diese Bedeutung ein Ergebnis der konventionellen Struktur ist und *nicht* eine automatische Inferenz aus den vorhandenen Filmdaten. Andere ähnliche und in der Filmtheorie häufig diskutierte konvenionelle Strukturen sind *point-of-view* Einstellungen (POVs) und *Flashbacks*.

Das Maß, in dem dies ein allgemeiner Bestandteil der Möglichkeiten von Film ist, ist bisher weit unterschätzt worden. Gleichermaßen ist die Tatsache, dass diese Mechanismen auch genauso in verbalen Sprachen vorkommen, nicht berücksichtigt worden. Unsere eigene Vorarbeit zeigt, dass viel mehr solcher Konstruktionen zu finden sind, aber ihr Auffinden erheblich dadurch erschwert wird, dass ein angemessener Zugang zu den Daten nicht verfügbar ist. Ein Beispiel für eine komplexere Konstruktion ist das folgende.

In Alfred Hitchcocks *North by Northwest* (1959) kommt eine Szene vor, wo der entführte Roger O. Thornhill (Cary Grant) erstmals mit seinem Entführer Phillip Vandamm (James Mason) konfrontiert wird. Die lose Klassifikation dieses Fragments laut Metzscher Kategorien ist unproblematisch. Wir können davon ausgehen, dass die Sequenz eine 'Szene' ist, weil der Zeitverlauf kontinuierlich und lückenlos ist, obwohl mehrere Einstellungen und Kamerapositionen vorkommen. Die Einstellungen sind außerdem alle im gleichen Zimmer und haben die Funktion, erst die eine Person, dann die andere, in den Vordergrund zu stellen. Innerhalb der Szene gibt es jedoch weitere distinktive Gruppierungen von Einstellungen, die die Kategorien von Metz nicht erfassen, obwohl sie eine wichtige Rolle für die Bedeutung und Struktur der Filmsequenz spielen. Um diese Gruppierungen

abzudecken, müssen wir den Begriff der Konstruktion heranziehen.

Die erste Konstruktion in diesem Segment verläuft zum Beipiel folgendermaßen. Zuerst tritt der Entführer ins Zimmer. Dann schauen sich Thornhill und Vandamm an. Dann fängt Vandamm an, in dem Zimmer nach links im Kreis herumzugehen. Darauf fängt Thornhill an, auch im Kreis im Zimmer herumzugehen. Die Auswirkung der beiden Vorgänge ist, dass beide Personen etwa in gleichem Abstand von einander bleiben, obwohl beide sich bewegen. Der filmische Gesamtvorgang wird als eine alternierende Sequenz behandelt: zuerst ist Thornhill im Bild, danach Vandamm, dann wieder Thornhill und so weiter. Die Kamera bewegt sich mit jedem der beiden, wenn er weiter geht. Dieses etwas komplexe Arrangement wird im beiliegenden Bild dargestellt.



Wenn diese Verfilmung nur einmal vorkäme, wäre sie nicht besonders signifikant oder interessant für Filmanalyse. Aber wir finden genau die gleiche allgemeine Konstruktion in vielen Filmen, und das lässt vermuten, dass diese Sequenz und ähnliche Variationen tatsächlich als Konstruktionen in linguistischem Sinne zu verstehen sind. Regisseure verwenden regelmäßig diese Art von Sequenz als einen Teil ihrer Filmsprache. Üblicherweise trägt die Konstruktion eine Bedeutung von ‘Konfrontation’, und Nuancen in Framing und Kamerawinkel implizieren weitere Variationen von Macht und Möglichkeiten der Beteiligten.

Diese Konstruktion impliziert mehr als die allgemeine Alternation, die wir im verfilmten Dialog finden. Die Kamera bewegt sich mit einer gleich bleibenden Geschwindigkeit, die Beteiligten auch. Die Beteiligten schauen sich symmetrisch die ganze Zeit an, signalisiert durch korrekte ‘eyeline matches’. Der Winkel zwischen Kamera und Schauspielern und zwischen den Schauspielern untereinander bleibt auch beibehalten. Das Ganze funktioniert genau wie eine linguistische Konstruktion. Sie wird durch gewisse halb-feste Eigenschaften der Sequenz getragen und impliziert.

Ein Hauptziel dieses Teilprojekts wird es sein, einen Katalog solcher Konstruktionen in Film zu erstellen und durch automatische Abfrage ihre Semantik gegen Daten zu validieren/präzisieren. Die automatische Erkennung wird in Kooperation mit Teilprojekt TP2 durchgeführt. Darüber hinaus ist es deutlich geworden, dass Bilder, und Bild-Text-Kombinationen noch mehr, eine ähnliche Abhängigkeit vom Diskurs zeigen. Zum Beispiel ist in der semiotischen Literatur vorgeschlagen worden, dass die konkrete Platzierung von Bildern auf einer Seite ideologische und rhetorische Bedeutung ausdrückt. Bilder, die weiter oben auf der Seite vorkommen, sind von Kress & van Leeuwen (1996) als repräsentativ für ‘idealisierte’ Umstände gedeutet worden, Bilder, die weiter unten auf der Seite vorkommen, als ‘Wirklichkeit’ darstellend. Diese Behauptung hat sich in

mehreren Bereichen der multimodalen Semiotik und Analyse fast zu einer Selbstverständlichkeit entwickelt, obwohl bis jetzt keine detaillierten empirischen Studien dazu durchgeführt worden sind (für eine der ersten, s. Holsanova et al. 2006). In diesem Teilprojekt werden wir zusammen mit den in TP1 entwickeltten Werkzeugen auch zum ersten Mal eine solche empirische Studie für unseren Bildkorpus durchführen. Ergebnisse, egal ob negativ oder positiv, werden einen wichtigen wissenschaftlichen Beitrag für diesen Bereich darstellen.

Die Ergebnisse des Teilprojekts werden dementsprechend in zwei Bereichen validiert werden.

Erstens wird der Wert der Ergebnisse für angereicherte Theorien des filmischen und bildlichen Narrativs untersucht werden. Traditionelle Ansätze von filmischer Bedeutung werden mit den zusätzlichen, aus Konstruktionen gewonnenen Bedeutungen verglichen werden. Auf dieser Basis werden weitere Dialoge eröffnet, die die linguistische Analyse von Diskurs zusammen mit literaturwissenschaftlichen Untersuchungen von Narrativ und Erzähltheorie bringen werden. Dabei wird erwartet, dass uns ein vertieftes Verständnis davon, wie Diskurs und Narrativ in *beiden* Medien funktionieren, gelingen wird sowie ein differenzierteres Modell der Unterschiede, die Narrativ in den beiden Medien aufweist.

Zweitens wird die Rolle der Konstruktionen als nützliche Einschränkungen für die automatische Bearbeitung von Sequenzen von Einstellungen und innerhalb Bild-Text-Kombinationen evaluiert werden. Wenn das Vorhandensein einer Konstruktion angenommen wird, gibt es daraus-folgend Einschränkungen für die erlaubten Merkmale der vorkommenden Einstellungen. Diese Einschränkungen sollten gezielt verwendet werden, um die Erkennung zu verbessern. Genau dadurch sehen wir auch im Kontext von Film, wie aus der Diskursstruktur (hier Konstruktionen) gewonnene Erwartungen nützliche Einschränkungen für die Bearbeitung bringen können. Dabei ist es möglich, einen uneingeschränkten Rückgriff auf Weltwissen zu vermeiden. Genauso werden Analysen von Diskursstrategien in Bild-Text-Kombinationen, wie von Martinec & Salway (2005) vorgeschlagen, zusammen mit TP3 in dieser Art evaluiert werden.

— Arbeitspakete (TP4) —

Arbeitspaket	*Kurztitel*	*Jahr 1*			**M1**	*Jahr 2*			**M2**	*Jahr 3*			**M3**
TP4.1 a,b	**Erstellung einer initialen Materialbasis**	XX X											
TP4.2	**Analyse der Diskursstrategien und narrativen Struktur**		XXX	X									
TP4.3	**Formalisierung der Diskursstrategien und filmischen Realisierung**			XX	XXX								
TP4.4	**Erweiterterung der Wissensbasis für beide Film- und Bildmaterial**					XX							
TP4.5 a,b	**Formalisierung der Diskursstrategien für Bilder und Bild-Text-Kombinationen**					X	XXX	X XX	X				
TP4.6	**Evaluierung bzgl. der Unterstützung des automatischen Analyseverfahrens**								XX				
TP4.7 a, b	**Definition von kombinierter Film/Bild/Text-Narrativschema**									XX X	X		
TP4.8	**Erschließung von komplexeren filmischen Konstruktionen**										XX	XX	
TP4.9	**Vergleich von Diskursstrategien über Medien und Genre**											X	XXX

TP4.1 Erstellung einer initialen Wissenbasis von Filmmaterial

a) In Kooperation mit TP2 (s. TP2.1) wird relevantes Filmmaterial gesammelt. (2PM)

b) Eine fokussierte weitere Auswahl von Filminhalten bzgl. des Bildmaterials in TP3 wird bereitgestellt, um die darauffolgende Analyse (s. TP4.4) der Gemeinsamkeiten von Film und Bild zu unterstützen. (1PM)

TP4.2 Diskursstrategie-Analyse

Ausgewählte Segmente werden hinsichtlich ihrer Diskursstrategien und narrativen Struktur analysiert unter besonderer Berücksichtigung ihrer Relevanz für die Event-Detektion (vgl. TP2.3). (4PM)

TP4.3 Diskursstrategie-Formalisierung

Formalisierung der Diskursstrategien und ihrer filmischen Realisierungen bzgl. grundlegenden Narrativentwicklungen für Sequenz und Alternanz. Die Ergebnisse werden in TP2.3 exportiert. (5PM)

Meilenstein 1. Systemarchitektur (M12). Diskursstrategien für Sequenz und Alternanz ausformuliert.

TP4.4 Erweiterung der Wissensbasis für beide Film- und Bildmaterial

Weiteres Material wird hier gesammelt und den Teilprojekten zur Verfügung gestellt. (2PM)

TP4.5 Formalisierung der Diskursstrategien für Bilder und Bild-Text-Kombinationen

a) Basisschema werden formell spezifiziert und kompatibel mit TP3.3 gemacht (5PM)

b) Die mit Strategien aus der Filmanalyse erweiterten Schemata werden in TP3.5 exportiert. (3PM)

TP4.6 Evaluierung der Diskursstrategien

Die erweiterten Diskursstrategien werden bzgl. ihrer Unterstützung des automatischen Analyseverfahrens evaluiert und verbessert. (2PM)

Meilenstein 2. Architekturevaluierung (M24).

TP4.7 Definition und Erprobung von kombinierter Film/Bild/Text-Narrativschema

a) Bildbezogene Erprobung wird zusammen mit TP1.7 durchgeführt. (2PM)

b) Filmbezogene Erprobung wird zusammen mit TP2.7 durchgeführt. (2PM)

TP4.8 Erweiterung der Formalisierung

Die Formalisierung der Diskursstrategien wird für Bilder und Bild-Text-Kombinationen erweitert und die Erschließung von immer komplexer werdenden filmischen Konstruktionen wird angestrebt. (4PM)

TP4.9 Diskursstrategien: ein Vergleich

Ein systematische Vergleich von Diskursstrategien (a) über Film und Text und (b) über Spielfilm und Nachrichten/Dokumentarfilme wird durchgeführt und veröffentlicht. Die Generalisierbarkeit des Ansatzes für weiteres, noch nicht betrachtetes Material wird dabei untersucht und dokumentiert. (4PM)

Meilenstein 3 und Projektende (M36). Vollständiges integrietes System: die Diskursstrategien von Film, Bild- und Bild-Text-Kombinationen.

Erwartetes Ergebnis und angestrebte Ergebnisverwertung bis hin zur Implementierung bzw. akademischen Anwendung

Völlig neu ist die Zusammenarbeit und die zuammenführende Art der Arbeit. Obwohl das Problem fehlender Informationsquellen und entsprechender Methoden für Lösungen auf beiden Seiten der Natur- und Geisteswissenschaften immer deutlicher wird, wird selten eine grundsätzliche Lösung und Auflösung der grenzbezogenen Problematik angestrebt. Mit der Zusammensetzung des beantragten Verbunds wird ein wichtiger und nachhaltiger Schritt in dieser Richtung möglich gemacht.

Ergebnisse des Gesamtvorhabens sind in folgenden Bereichen zu verorten: (i) ein erhöhtes theoretisches Verständnis der Mechanismen von Narrativ in Film und Text, (ii) ein erhöhtes theoretisches Verständnis dafür, wie Wissen und Diskursstrukturen die Interpretation von Bild und Film erleichtern können, (iii) eine Verbesserung in der Erfassung von visuellen Daten (a) bezüglich der qualitativen Klassifikation von Daten und (b) bezüglich einer gemeinsamen Klassifikation, die statische und bewegte Bilder im selben Rahmen behandelt. Erweiterungen von Standards für Bilddaten und Bild/Filmkorpora und von Methoden für ihre Erstellung sowie ihre Evaluierung sind auch zu erwarten. Die erzielte Verbesserung der Leistung von Film- sowie Bildinterpretation wird auf ihre kommerzielle Anwendbarkeiten überprüft.

Darüber hinaus ist längerfristige und nachhaltige gemeinsame Forschung zu erwarten, sowie eine Reihe neuer Impulse für die Analyse dieses besonders aktuellen Kommunikationsmodus in Bereichen wie automatische Erkennung, theoretische Betrachtung, Einsatz in der Ausbildung und Lehre, kritischer Umgang mit Design, politisch-ideologische Untersuchungen von Darstellungen

von Kultur und Kulturereignissen, und vieles mehr. Alle diese Forschungsrichtungen vertiefen die Möglichkeiten für Wechselwirkungen zwischen den Natur- und Geisteswissenschaften, setzen aber voraus, dass eine Grundlage für die Kommunikation zwischen beiden Bereichen vorhanden ist. Genau diese wird das aktuelle Vorhaben liefern.

Die in diesem Projekt geführten Verwertungsstrategien für diese Ergebnisse sind im getrennten Anhang detailliert beschrieben.

Kriterien für den Projekterfolg

Die Antragsteller werden einzeln und gemeinsam Veröffentlichungen in internationalen, referierten Zeitschriften, sowie Beiträge in angemessenen Sammelbänden anstreben. In den beteiligten Fächern wird eine enge Kooperation mit laufenden und neuen Dissertationsprojekten sowie mit anderen Forschergruppen verfolgt, um die Ergebnisse schnell in weitere Arbeiten zu überführen. Projekterfolg wird an nachweisbaren Verbesserungen in der Performanz der Bild/Filmbearbeitungssysteme sowie durch internationale Veröffentlichungen gemessen. Regelmässige Projektberichte und interne arbeitsgruppenübergreifende Treffen werden die Arbeit begleiten und die Durchführung des Gesamtprojekts dokumentieren.

Öffentlichkeitsarbeit

Die Antragsteller werden einzeln und zusammen an relevanten Tagungen und Workshops teilnehmen. Interaktionen mit anderen Forschungsgruppen werden aktiv initiiert und verfolgt. Es wird nach jedem im Projektplan definierten Meilenstein (s. oben) ein Workshop mit internationaler Beteiligung organisiert, um die Ergebnisse und Methoden bekannt zu machen und weitere Möglichkeiten für Transfer und Kooperationen auszusuchen. Eine aktive Webpräsenz wird aufrechterhalten. Darüber hinaus werden internationale Kooperationen mit folgenden Forschern und Einrichtungen durchgeführt werden: Prof. Theo van Leeuwen (Sydney), Prof. James R. Martin (Sydney), Prof. Kay O'Halloran (National University of Singapore), Prof. Kenneth Holmqvist und Dr. Jana Holsanova (Lund Universität). Alle sind führend in dem Bereich Multimodalität.

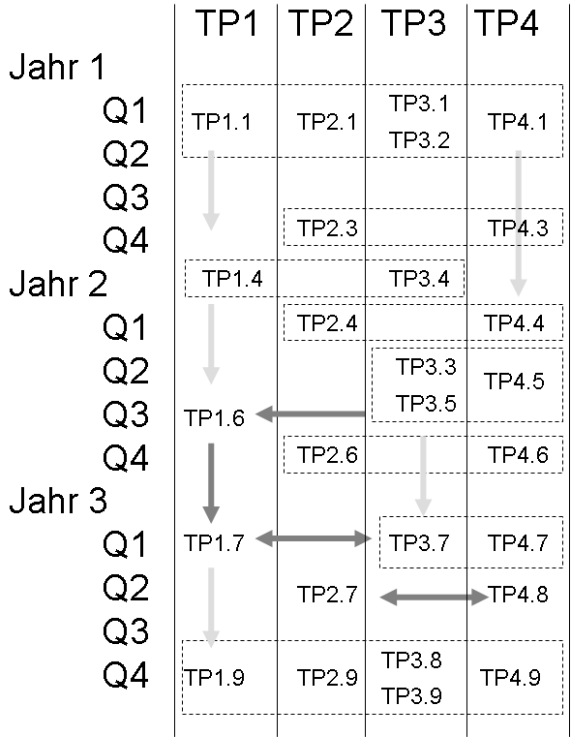
Zeitplanung und Projektablauf

Alle Teilprojekte laufen 3 Jahre. Der Arbeitsverlauf ist durch Meilensteine synchronisiert, die für alle Teilprojekte gelten. In jedem Jahr entsteht eine neue Revision der gesamten Architektur des

Projekts, die evaluiert und in der Öffentlichkeit vorgestellt wird. Die Rolle der Meilensteine ist in der folgenden Tabelle zusammengefasst.

Meilenstein 1	M12	Architekturdefinitionen: Film und Bild
Meilenstein 2	M24	Architekturevaluierung
Meilenstein 3	M36	Erweiterte Architektur: gemeinsame Behandlung von Film/Bild/Text

Die Synchronisierung der Teilprojekte und der Austausch ihrer Hauptzwischenergebnisse ist in der folgenden Grafik dargestellt. Daraus entsteht eine mehrstufige und eng verzahnte Zusammenarbeit, in welcher der gelieferte Systemprototyp sukzessiv weiterentwickelt und evaluiert wird.



Literatur

Aigrain, P., Joly, P. & Longueville, V. (1995), Medium Knowledge-based Macro-Segmenation of Video into Sequences, *in* 'Proc. of IJCAI Workshop on Intelligent Multimedia Information Retrieval'.

Asher, N. & Lascarides, A. (2003), *Logics of conversation*, Cambridge University Press, Cambridge.

Bateman, J. A. (2007), 'Towards a *grande paradigmatique* of film: Christian Metz reloaded', *Semiotica* **167**(1/4), 13–64.

Bateman, J. A., Delin, J. L. & Henschel, R. (2004), Multimodality and empiricism: preparing for a corpus-based approach to the study of multimodal meaning-making, *in* E. Ventola, C. Charles & M. Kaltenbacher, eds, 'Perspectives on Multimodality', John Benjamins, Amsterdam, pp. 65–87.

Bateman, J. A., Delin, J. L. & Henschel, R. (2007), Mapping the multimodal genres of traditional and electronic newspapers, *in* T. D. Royce & W. L. Bowcher, eds, 'New Directions in the Analysis of Multimodal Discourse', Lawrence Erlbaum Associates, pp. 147–172.

Biederman, I. (1987), 'Recognition by components: A theory of human image understanding', *Psychological Rev* pp. 115–147.

Blankert, L., Jacobs, A., Miene, A., Hermes, T., Ioannidis, G. & Herzog, O. (2005), An environment for modelling telecast structures, TZI-Bericht 32, TZI Technologie-Zentrum Informatik, Universität Bremen, Bremen.

Bordwell, D. (1996), Convention, construction and cinematic vision, *in* D. Bordwell & N. Carroll, eds, 'Post-theory: reconstructing film studies', University of Wisconsin Press, Madison, Wisconsin, pp. 87–107.

Christel, M., Hauptmann, A. & Wactlar, H. (2001), 'Improving Access to Digital Video Archives through Informedia Technology', *To Appear in Journal of the Audio Engineering Society* .

de Vries, J. (2008), 'Newspaper design as cultural change', *Visual Communication* **7**(1), 5–25.

Deardorff, E., Little, T., Marskall, J., Venkatesh, D. & Walzer, R. (1994), Video Scene Decomposition with the Motion Picture Parser, *in* 'IS&T/SPIE Symposium on Electronical Imaging Science & Technology (Digital Video Compression and Processing on Personal Computers: Algorithms and Technologies)', Vol. SPIE 2187, San Jose, CA, pp. 44–55.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995), 'Query By Image And Video Content: The QBIC system', *IEEE Computer: Special issue on Content Based Picture Retrieval Systems* .

Fukushima, K. (1975), 'Cognitron: A self-organizing multilayered neural network', *Biological Cybernetics* **20**, 121–136.

Goldberg, A. E. (1995), *Constructions: a construction grammar approach to argument structure*, University of Chicago Press, Chicago.

Hermes, T., Klauck, C. & Herzog, O. (1999), Knowledge-based image retrieval, *in* B. Jähne, H. Haussecker & P. Geissler, eds, 'Handbook of Computer Vision and Application, Vol. 3', Academic Press, chapter 25, pp. 517–532.

Hermes, T., Klauck, C., Kreyß, J. & Zhang, J. (1995), Image Retrieval for Information Systems, *in* 'Proceedings of IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology', San Jose, CA, USA.

Holsanova, J., Rahm, H. & Holmqvist, K. (2006), 'Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements', *Visual Communication* **5**(1), 65–93.

Knox, J. S. (2007), 'Visual-verbal communication on online newspaper home pages', *Visual Communication* **6**(1), 19–53.

Kress, G. & van Leeuwen, T. (1996), *Reading Images: the grammar of visual design*, Routledge, London and New York.

LeCun, Y., Matan, O., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., , Jackel, L. D. & Baird, H. (1990), Handwritten zip code recognition with multilayer networks, *in* 'Proceedings of the 10th International Conference on Pattern Recognition', IEEE Computer Science Press, Los Alamitos, CA.

Lienhart, R. (1999), Comparison of Automatic Shot Boundary Detection Algorithms, *in* 'Proc. SPIE Vol. 3656 Storage and Retrieval for Image and Video Databases VII', San Jose, CA, USA, pp. 290–301.

Mann, S. & Picard, R. (1995), Video orbits of the projective group: a new perspective on image mosaicing, Technical Report 338, MIT Technical Report.

Martinec, R. & Salway, A. (2005), 'A system for image-text relations in new (and old) media', *Visual Communication* **4**(3), 337–371.

Metz, C. (1974), *Film language: a semiotics of the cinema*, Oxford University Press and Chicago University Press, Oxford and Chicago. Translated by Michael Taylor.

Miene, A., Dammeyer, A., Hermes, T. & Herzog, O. (2001), Advanced and Adaptive Shot Boundary Detection, *in* 'Proceeding of the ECDL Workshop Generalized Documents: A key challenge in digital library research and development', Darmstadt, Germany.

Miene, A., Hermes, T. & Ioannidis, G. T. (2001), Automatic Video Indexing with the ADViSOR System, *in* 'Proceedings of the CBMI 2001 International Workshop on Content-Based Multimedia Indexing', Brescia, Italy. (to appear).

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

Moncrieff, S., Dorai, C. & Venkatesh, S. (2001), Detecting Indexical Signs in Film Audio for Scene Interpretation, *in* 'Proc. of ICME'01'.

Müller, M. G. (1997), *Politische Bildstrategien im amerikanischen Präsidentschaftswahlkampf 1828–1996*, Akademie Verlag, Berlin.

Müller, M. G. (2003), *Grundlagen der visuellen Kommunikation. Theorieansätze und Methoden*, konstanz edn, UVK, utb.

Müller, M. G. (2006), Die Ikonographie des politischen Händedrucks, *in* S. Appuhn-Radthke & E. Wipfler, eds, 'Freundschaft, Motive und Bedeutungen', Zentralinstitut für Kunstgeschichte, München, pp. 205–215.

Müller, M. G. (2007), 'What is visual communication? Past and future of an emerging field of communication research', *Studies in Communication Sciences* **7**(2), 7–34.

Müller, M. G. & Özcan, E. (2007), 'The political iconography of Muhammad Cartoons: Understanding cultural conflict and political action', *PS: Political Science & Politics* **40**(2), 287–292.

Olshausen, B., Anderson, C. & van Essen, D. (1993), 'A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information', *Journal of Neuroscience* **13**, 4700–4719.

Oram, M. W. & Perrett, D. I. (1994), 'Modeling visual recognition from neurobiological constraints', *Neural Networks* **7**(6/7), 945–972.

Schmidt, K.-H. & Strauch, T. (2002), 'Zur chronologischen Syntagmatik von Bewegtbilddaten', *Kodikas/Code: Ars Semeiotica* **25**(1-2), 65–96.

Smolic, A., Sikora, T. & Ohm, J.-R. (1999), 'Long-Term Global Motion Estimation and its Application for Sprite Coding, Content Description and Segmentation', *IEEE Trans. on CSVT* **9**(8), 1227–1242.

Tanaka, K., Saito, H., Fukada, Y. & Moriya, M. (1991), 'Coding visual images of objects in the inferotemporal cortex of the macaque monkey', *Journal of Neurophysiology* **66**, 170–189.

Teichert, J. & Malaka, R. (2003), An association architecture for the detection of objects with changing topologies, *in* 'Proceedings of the International Joint Conference on Neural Networks (IJCNN 2003)', Portland, OR, pp. 125–130.

Teichert, J. & Malaka, R. (2006), Iterative context compilation for visual object recognition, *in* 'Proceedings of the European Symposium on Neural Networks (ESANN'06)'.

Wallis, G. & Rolls, E. T. (1997), 'Invariant face and object recognition in the visual system', *Progress in Neurobiology* pp. 167–194.

Yusoff, Y., Christmas, W. & Kittler, J. (1998), 'A study on automatic shot change detection', *Lecture Notes in Computer Science* **1425**.

Zhang, H., Tan, S., Smoliar, S. & Hong, G. (1995), 'Automatic Parsing and Indexing of News Videos', *Multimedia Systems* **2**(6), 256–266.

Zhang, T. & Kuo, C.-C. (1999), Audio-guided Audiovisual Data Segmentation, indexing and Retrieval, *in* 'Proc. of SPIE, Vol. 3656', SPIE, pp. 316–327.