VITAlity - ein Framework zur Modellierung und Verwaltung von multimedialen Dokumenten

A. Finger, I. Bruder, T. Ignatova, M. Rust
Graduiertenkolleg Multimedia der Universität Rostock
Email:{af,ilr,temi}@informatik.uni-rostock.de, mrust@rostock.zgdv.de

1 Einführung

Das Projekt VITAlity (Video, Image, Text, Audio) beschäftigt sich mit der Analyse und Verwaltung multimedialer Daten. Motiviert wird diese Arbeit zum einen durch die Erkenntnis, dass bei der Verarbeitung verschiedener Medientypen durchaus ähnliche Arbeitsschritte durchzuführen sind und zum anderen dadurch, dass Querbeziehungen zwischen diesen Schritten bestehen. So kann ein Medium seinen Typ während des Analysevorgangs ändern, ein Bild wird bspw. mittels OCR in einen Text umgewandelt und kann nun durch Methoden der Textanalyse inhaltlich erschlossen werden. Durch die gemeinsame Verarbeitung multimedialer Dokumente innerhalb eines Informationssystems ergeben sich also mehrere Vorteile. Entwickler medienspezifischer Verarbeitungsmethoden haben Zugriff auf eine umfangreichere Wissensbasis und somit erweiterte Analysemöglichkeiten. Anwender profitieren von der Ausnutzung der Beziehungen in Dokumenten und erhalten die Möglichkeit einer medientransparenten Suche. Dieser Artikel stellt erste konzeptionelle Gedanken zur Realisierung eines solchen Frameworks in Form eines Komponentenmodells vor.

2 Ziele, Anforderungen und Abgrenzung

Ein multimediales Informationssystem ist allgemein ein System zur Verarbeitung, Management oder Retrieval von multimedialen Dokumenten. Es existieren verschiedene Vorstellungen zu multimedialen Informationssystemen. Einen sehr weiten Überblick gibt bspw. [3] mit einer sechsschichtigen Architektur bestehend aus Anwendungsschicht, Diensteschicht, Datenmodellschicht, Integrationsschicht, Speicherungssystemschicht und Betriebssystemschicht. Im Projekt VITAlity liegt der Fokus zunächst auf der kombinierten Analyse multimedialer Dokumente und somit innerhalb der Diensteschicht. Es sollen ganzheitliche, multimediale Dokumente als auch enthaltene Medien inhaltlich erschlossen und "verstanden" werden. Es sollen hierbei Querbeziehungen ausgenutzt werden, um unnötigen Aufwand bei der Extraktion von Merkmalen zu vermeiden. Die Ausnutzung von Beziehungen in multimedialen Dokumenten wurde bereits in vorherigen Arbeiten (z.B. [5]) beschrieben. VITAlity beschränkt sich bei der Analyse auf die elementaren Medientypen Video, Image, Text und Audio. Andere Systeme wie etwa MAVA [6], berücksichtigen zusätzlich 3D-Animationen und Spiele als Komponenten multimedialer Dokumente, deren Schwerpunkt liegt jedoch auch mehr auf einer Unterstützung des Authoring-Prozesses. Bei der Definition eines einheitlichen Modells zur Verwaltung multimedialer Dokumente orientieren wir uns an vorhandenen Konzepten. [4] enthält einen guten Überblick zu Dokumentenmodellen, welche anhand aufgestellter Anforderungen analysiert werden. Dabei soll das Dokumentenmodell eine Abbildung räumlicher, zeitlicher als auch interaktionaler Beziehungen ermöglichen und darüber hinaus eine feingranulare Wiederverwendbarkeit, eine Anpassung der Präsentation abhängig vom Anwender und dessen technischem Umfeld sowie eine Repräsentation unabhängig

von der endgültigen Darstellung zulassen. Nicht zuletzt sollte auch die Integration von Businessmodellen und Verfahren des Digital Rights Management möglich sein, um eine Kommerzialisierung nicht von vornherein auszuschließen. SMIL, HyTime, ZyX oder auch MPEG-21 gehören zu den Dokumentmodellen, die diesen Anforderungen zumindest teilweise entsprechen. Weitere Eigenschaften des Frameworks sollen Verteilung der Funktionalität (Analyseverfahren), die Integration von Analyseverfahren in eine Datenbankumgebung sowie klassische Funktionalität zur Dokumentenverwaltung sein. Für das System ergeben sich somit drei wesentliche Nutzergruppen, nämlich Bereitsteller von Daten, Metadatenextraktoren sowie suchend zugreifende Benutzer. Ein wesentliches Ziel dieses Projekts ist die Bereitstellung erweiterter Suchmöglichkeiten über multimediale Dokumente. Durch die Ausnutzung von Strukturen und somit der Generierung medienübergreifenden Wissens soll dem Anwender eine medientransparente Suche zur Verfügung gestellt werden.

3 Komponentenmodell zur Verwaltung multimedialer Dokumente

Abbildung 1 zeigt VITAlity als abstraktes Komponentenmodell, wobei wesentliche Aspekte in verschiedenen Ebenen dargestellt sind. Die oberste Ebene repräsentiert die Vorverarbeitung des

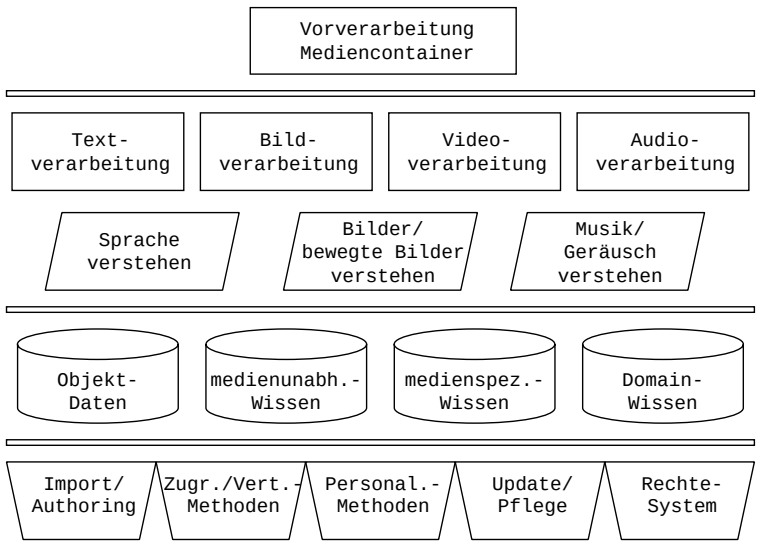

Abbildung 1: Komponentenmodell eines Multimedialen Informationssystems

multimedialen Dokuments, wobei allgemeine Merkmale wie Dokumenttyp (HTML, PDF, SMIL, etc.) des Mediencontainers sowie weitere physische Eigenschaften (z.b. Dateigröße) und Metadaten (Erstellungsdatum oder Autor) aber auch die Struktur, also Komponenten und deren Beziehungen, bestimmt werden. Hierzu benötigtes Wissen setzt sich aus Wissen über den Dokumententyp, dessen Struktur und Methoden zur Verarbeitung des Dokumententyps zusammen. In der zweiten Ebene findet eine medienspezifische Weiterverarbeitung der aus der Vorverarbeitung gewonnenen Dokumentkomponenten statt. Ein Überblick über die Möglichkeiten wird im jeweiligen Abschnitt in Kapitel vier gegeben. Weiterhin sind in dieser Ebene drei "Verstehenskomponenten" abgebildet. Das soll verdeutlichen, dass bei einer kombinierten Analyse der verschiedenen Medientypen medientypunabhängiges Wissen generiert wird. Die extrahierten Merkmale werden dabei auf eine abstraktere Ebene abgebildet, welche an der menschlichen Wahrnehmung orientiert ist. So unterscheiden wir lediglich das Verstehen von Sprache, Bildern bzw. bewegten Bildern und Musik bzw. Geräusche. Dieser Gedanke wird der Tatsache gerecht,

dass wir digitale Medien visuell oder auditiv wahrnehmen können. Sprache kann dabei sowohl im Bild als auch im Schall codiert sein und muss aufgrund der Bedeutung als Kommunikationsmittel gesondert betrachtet werden. Die dritte Ebene des Komponentenmodells kann auch als Wissensebene bezeichnet werden. Hier sind Komponenten zur Repräsentation von globalem also medientypunabhängigem Wissen, medienspezifischem Wissen, Domänwissen als auch Objektwissen, sprich die extrahierten Objektdaten selbst vertreten. Die fünfte Ebene enthält verschiedene Komponenten zur Verwaltung der multimedialen Dokumente und repräsentiert somit Funktionalitäten zur Abbildung des Dokumenten-Life-Cycles sowie Personalisierung oder Rechteverwaltung.

4 Komponenten zur Analyse und Verwaltung von Medientypen

4.1 Text

Text betrachten wir bezüglich der maschinellen Verarbeitung als nach wie vor geeignetste Repräsentation von Inhalten. Textdaten repräsentieren Sprache in verschiedenen computer-lesbaren Formaten. Es existieren simple Textformate (ASCII, UTF8, Unicode, etc.), proprietäre Binärformate (z.B. Word, PS) oder strukturierte Texte (HTML, XML, SGML). Eine Umwandlung zwischen diesen Formaten ist möglich, wobei sie von binären und srukturierten zu simplen Formaten verlustbehaftet ist. Darüber hinaus kann Sprache, wie bereits erwähnt, auch in anderen Medientypen, etwa Audio oder Bild, codiert sein. Für eine Analyse von Sprache mit Hilfe von Textverarbeitungsmethoden ist also eine Transformation in ein entsprechendes Textformat erforderlich. An diesem Punkt ergeben sich notwendige Schnittstellen zu anderen Komponenten, welche die Bild- bzw. Audiodatenverarbeitung durchführen.
Ein wichtiger Bestandteil der Textverarbeitung stellt die Strukturerkennung dar. Strukturen können inhaltlich (Worte), logisch (Kapitel, Abschnitt, etc.), physisch (z.B. Seiten, Linknavigation) oder anderer Art (z.B. Layoutbezeichnungen) sein, wobei diese Kategorien voneinander unabhängig sind. Zur Verarbeitung von Texten sind vor allem Methoden aus dem Information Retrieval, der Computerlinguistik und dem Text Mining interessant [2]. Das Ziel ist, aus den Texten inhaltliche Feature zu extrahieren und sie mit den Strukturen in Beziehung zu setzen. Von den so normierten Daten lassen sich anschließend auf relativ einfache Art Verbindungen zum domänenspezifischen Wissen herstellen.

4.2 Image

Digitale Bilder treten in vielen Anwendungsgebieten, von Medizin bis zu Anwendungen aus dem Bereich der Erhaltung kulturellen Erbes, auf. Bei der Verarbeitung und der Analyse von Bildern werden Techniken aus verschiedenen Fachgebieten, wie etwa Computer Vision, Information Retrieval, Data Mining, Datenbanken, etc. eingesetzt. Der größte Teil aktueller Forschung beschäftigt sich mit dem inhaltsbasierten Zugriff auf Bilddaten unter Verwendung visueller Charakteristiken, wie Farbverteilung, Texturen oder Formen. Abstraktere Repräsentationen werden genutzt, um semantische Konzepte (z.B. “die Sonne geht auf”) zu extrahieren und zu verstehen. Die Integration eines inhaltsbasierten Zugriffs auf Bilddaten in multimedialen Informationssystemen führt zu erweiterten Möglichkleiten für die Verwaltung struktureller und semantischer Beziehungen zu anderen Medientypen und somit zu einem inhaltsbasierten Zugriff auf Multimedia-Informationen im Allgemeinen [8]. Allgemeine Verarbeitungschritte z.B. bei der Vorverarbeitung oder Segmentierung einzelner Medientypen sind auch auf andere Medien anzuwenden und können somit einer gemeinsamen Wissensbasis entnommen werden. Weiterhin ergeben sich Schnittstellen zwischen der Bildverarbeitung und anderen Verarbeitungskomponenten, wenn Bilddaten bspw. in Text (OCR) oder eine symbolische Repräsentation von Musik (OMR) überführt werden. Die inhaltsbasierte Suche auf Bilddaten in textueller Form (Annotationen),

visueller Form (Query by Example, Query by Sketch) sowie die Suche auf verschiedenen Abstraktionsebenen, sprich low-level (Farbverteilung, Textur, Form) und high-level (Objekte, Szenen, Konzepte) stellen wesentliche Ziele der bildverarbeitenden Komponente in VITAlity dar.

4.3 Audio

Bei Audiodaten handelt es sich um eine digitale Repräsentation von Schallereignissen. Dabei muss zwischen mehreren Formen von Schallereignissen, nämlich Sprache, Geräusche und Musik, unterschieden werden. Sprache kann mittels Verfahren der automatischen Spracherkennung in Text umgewandelt und somit an eine textverarbeitende Komponente weitergereicht werden. Wenn Geräusche (unperiodische Schwingungen ohne exakte Tonhöhe) isoliert auftauchen, können sie in Sound-Datenbanken [9] auf Basis ihrer Klangeigenschaften analysiert und verwaltet werden. Eine Anwendung hierfür ist bspw. der Einsatz bei der Vertonung von Filmen. Geräusche, wie sie z.B. Schlaginstrumente erzeugen, können jedoch auch Bestandteil von Musik sein. Systeme zur Verwaltung von Musik existieren meist in Form digitaler Musikbibliotheken wie [1]. In solchen Systemen hat der Nutzer meist die Möglichkeit eine Melodie per Mikrophon (Query by Humming) oder mittels Texteingabe (auf Text abgebildete Notensymbole) anzufragen. Weiterhin ist eine Suche nach bibliographischen Merkmalen wie Komponist, Sänger, etc. möglich. Darüber hinaus existieren jedoch weitere musikalische Merkmale (Rhythmus, Harmonie oder Tempo), die zumindest in bisherigen digitalen Musikbibliotheken oder Informationssystemen nicht zugreifbar sind. Einen Sonderfall stellen symbolische Repräsentationen (z.B. MIDI) von Musik dar. Sie liegen meist im Textformat vor und können auch mit Hilfe von Textverarbeitungsmethoden analysiert werden, benötigen hierzu allerdings auch musikalisches Wissen. Hier ergibt sich wiederum eine Schnittstelle zwischen den Analysekomponenten des multimedialen Informationssystems.

4.4 Video

Mit zunehmend breitbandigeren Internetverbindungen wächst auch die Bedeutung des Medientyps Video in verschiedensten Anwendungsfeldern. Videos können in unterschiedlichen Formaten codiert sein. So existeren für den internetbasierten Zugriff eine Reihe von verlustbehafteten Streaming-Formaten (MPEG-4, Real, Microsoft, Quicktime). Für die Archivierung von Videos in hoher Qualität werden hingegen Standards wie MPEG-2 oder hochqualitative Varianten der angesprochenen Streaming-Formate verwendet. Videos bestehen aus Sequenzen von Bildern, denen eine oder mehrere Audiospuren synchron zugeordnet sind. Grundlage vieler videoverarbeitender Methoden bildet deshalb eine Videosegmentierung [7] und eventuell eine darauf aufbauende zeitliche Strukturerkennung, die eventuell vorhandenes Strukturwissen verwenden kann. Weitere Methoden widmen sich der Erkennung von Situationen und Objekten im Video mit ihren jeweiligen zeitlichen, räumlichen oder semantischen Beziehungen. Zusammen mit der Strukturerkennung ist es so beispielsweise möglich sinnvolle Zusammenfassungen von Videos zu definieren. Ein Ziel des vorgestellten Frameworks VITAlity besteht in der Kombination von Verarbeitungsmethoden für Videos mit Analysemethoden für die Medientypen Bild, Audio und Text. So können im Anschluss an die Strukturerkennung für ausgewählte Keyframes Bildmerkmale identifiziert, Beschriftungstexte extrahiert und verarbeitet oder gesprochene Kommentare erkannt und durch Textmethoden weitergehend analysiert werden.

5 Komponenten zur wissensbasierten Interpretation multimedialer Inhalte

Für inhaltsbasiertes Retrieval ist Wissen wichtig, um Daten so zu verstehen, dass konkrete Antworten auf vage Anfragen gegeben werden können. Dabei wird Wissen als eine Abstrakti-

on der konkreten Daten behandelt und in unserem Fall werden multimediale Dokumente auf das abstrakte Wissen abgebildet bzw. durch dieses Wissen beschrieben. Neben diesem meist domänenspezifischen Wissen wird auch Wissen für die Anwendung der umfangreichen Analysefunktionalitäten, Wissen über die verschiedenen Repräsentationsformen von Inhalten und Wissen zu den konkreten Instanzen benötigt. Wissen kann als einfaches Lexikon beschrieben werden, das von einer Repräsentationsform auf eine andere (eventuell abstraktere) Form abbildet. Wissen kann auch als semantisches Netz (z.B. Ontologien, Thesauri) mittels Konzepten, Attributen und Beziehungen viel komplexer dargestellt werden. Die gemeinsame Anwendung dieser beiden Wissensformen wird hier und in vielen anderen Projekten verwendet, um eine Interpretation von Daten zu ermöglichen.
Hinsichtlich des Wissens stellt sich auch die Frage nach der Generierung, nach der Anwendung und vor allem auch nach der Anfrageunterstützung. Für die Generierung des Wissens sind Experten auf dem jeweiligen Gebiet notwendig. Der Generierungsvorgang ist i.d.R. manuell, oder teilautomatisiert bspw. durch Data Mining Funktionalität. Die Anwendung muss für die jeweilige Schicht und Anwendungsfunktionalität gesondert definiert werden. Man kann einfache Abfragen in der Wissensbasis oder auch Inferenzberechnungen durchführen. Für die Anfrageunterstützung bietet es sich an, dem Nutzer die Wissensbasis als Navigationswerkzeug zur Verfügung zu stellen. Man kann aber auch direkte Anfragen an die Wissensbasis zulassen, um darüber auf konkrete Informationen weiterzuverweisen.

6 Zusammenfassung und Ausblick

Das hier vorgestellte Framework VITAlity bildet einen Rahmen für die Modellierung, Verwaltung und Analyse multimedialer Daten. Es formt das Umfeld für verschiedene Forschungsprojekte, die sich mit der inhaltlichen Erschließung multimedialer Dokumente befassen. Der Fokus liegt auf der Betrachtung möglicher Gemeinsamkeiten und Querbeziehungen bei der Medienanalyse. Die Ausnutzung dieser Beziehungen führt zu einer Wissensbasis die ein Verstehen des Inhaltes unabhängig vom Medientyp und damit Benutzern medientransparenten Zugriff ermöglicht.

Literatur

[1] D. Bainbridge, C. G. Nevill-Manning, I. H. Witten, L. A. Smith, and R. J. McNab. Towards a Digital Library of Popular Music. In *Proceedings of the fourth ACM conference on Digital libraries*, New York, 1999. ACM Press.

[2] R. Beaza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[3] H. Berthold, F. Binkowski, A. Henrich, S. Hollfelder, W. Lindner, U. Marder, K. Meyer-Wegener, and G. Robbert. Architektur Multimedialer Informationssysteme. *Informatik – Forschung und Entwicklung*, 17(2), 2002.

[4] S. Boll, W. Klas, and U. Westermann. *Multimedia Document Models – Sealed Fate or Setting Out for New Shores?* Kluwer Academic Publishers, 2000.

[5] I. Bruder and T. Ignatova. Utilizing Relations in Multimedia Document Models for Multimedia Information Retrieval. In *Proceedings of the International Conference - Information and Communication Technologies and Programming - Special Session on Multimedia Semantics*, Varna, 2003.

[6] J. Hauser. Component-based extensible multimedia system. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications PDPTA'01*, 2001.

[7] I. Koprinska and S. Carrato. Temporal Video Segmentation: A Survey, 2000.

[8] B. Perry, S.-K. Chang, J. Dinsmore, D. Doermann, A. Rosenfeld, and S. Stevens. *Content-Base Access To Multimedia Information - From Technology Trends To State of the Art.* Kluwer Academic Publishers, Boston, 1999.

[9] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-Based Classification, Search, and Retrieval of Audio. *IEEE Multimedia*, 3, 1996.