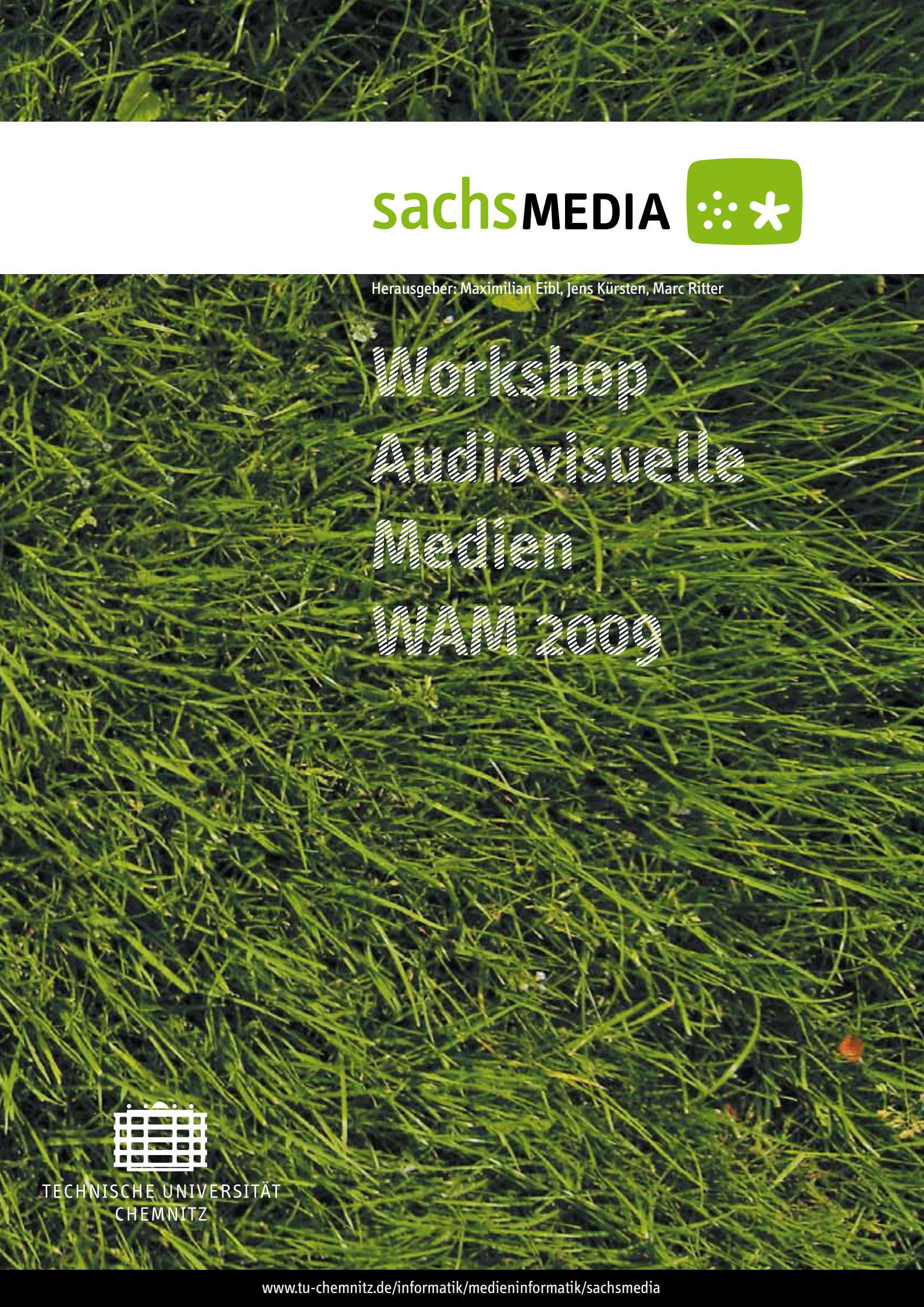


sachsMEDIA

Herausgeber: Maximilian Eibl, Jens Kürsten, Marc Ritter

Workshop Audiovisuelle Medien WAM 2009

TECHNISCHE UNIVERSITÄT CHEMNITZ

www.tu-chemnitz.de/informatik/medieninformatik/sachsmedia

Workshop Audiovisuelle Medien
WAM 2009: Archivierung

Maximilian Eibl, Jens Kürsten, Marc Ritter (Hrsg.)

TECHNISCHE UNIVERSITÄT
CHEMNITZ

CSR-09-04

Aus der Reihe

Chemnitzer Informatik-Berichte

Maximilian Eibl
Jens Kürsten
Marc Ritter

Technische Universität Chemnitz
Professur Medieninformatik
Nachwuchsforschergruppe sachsMedia
Straße der Nationen 62
09111 Chemnitz
E-Mail: vorname.name@informatik.tu-chemnitz.de

Chemnitz 2009

ISBN 9-78300-278587

ISSN 0947-5125

Inhaltsverzeichnis

Special Issues in Multimedia Archiving

Multimedia Analysis and Retrieval

Distribution Aspects

Multimedia Archives – Music

Vorwort

Audiovisuelle Medien stellen Archive vor zunehmende Probleme. Ein stark wachsender (Web-)TV-Markt mit Sendematerial oder Rohmaterial, zunehmender Einsatz von medial aufbereitetem Lehrmaterial in Schulen, Hochschulen und Firmen, die Verbreitung der Videoanalyse als Forschungs- und Lehrmethode z.B. empirische Bildungsforschung, Lehrerausbildung in der Erwachsenenbildung, die Ausbreitung von Überwachungskameras sowie die immer günstigeren Produktionsbedingungen vom Professionellen Produzenten bis zum Heimvideo sind nur einige Stichworte um die neuen quantitativen Dimensionen zu umreißen. Die aktuellen archivarischen und dokumentarischen Werkzeuge sind heute mit dieser Situation überfordert.

Der Workshop versucht hier Probleme und Lösungsmöglichkeiten zu umreißen und beschäftigt sich mit den technologischen Fragestellungen rund um die Archivierung audiovisueller Medien, seien es analoge, digitalisierte oder digitale Medien. Dabei werden zum einen die technologischen Probleme angesprochen, die zum Aufbau eines Archivs bewältigt werden müssen. Zum anderen wird der praktische Einsatz von der Gestaltung der Benutzungsoberfläche bis zur Frage des Umgangs mit kritischem Material diskutiert. Der vorliegende Tagungsband enthält 22 auf dem Workshop vorgestellte Beiträge.

Der Workshop wurde durch die BMBF-geförderte Nachwuchsforschergruppe „sachsMedia“ organisiert und in Kooperation mit der Fachgruppe Knowledge Media Design der Gesellschaft für Informatik e.V. sowie dem Forschungsschwerpunkt Intelligente Multimediale Systeme (IMS) der Fakultät für Informatik der TU Chemnitz am 4. und 5. Juni 2009 durchgeführt.

Chemnitz, im Mai 2009

Maximilian Eibl, Leiter Professur Medieninformatik
Jens Kürsten, Leiter sachsMedia
Marc Ritter

MedioVis 2.0 - A novel User Interface for Seeking Audio-Visual Media Libraries

Harald Reiterer, Mathias Heilig and Sebastian Rexhausen

Universität Konstanz
Fachbereich Informatik und Informationswissenschaft
Arbeitsgruppe Mensch-Computer Interaktion

{harald.reiterer,mathias.heilig,sebastian.rexhausen}@uni-konstanz.de

Abstract: Knowledge work is a demanding activity caused on the one hand by the increasing complexity of today's information spaces. On the other hand, knowledge workers are acting correspondingly to an individual creative workflow, which involves multifaceted characteristics like diverse activities, locations, environments and social contexts. Although it is important to find solutions to specific aspects of knowledge work (information-seeking, information-management, media-warehousing, etc.) our design approach – MedioVis 2.0 – tries to support the entire workflow in one coalescing Knowledge Media Workbench, showcased in the context of digital libraries. To achieve this goal, we apply the concept of zoomable user interfaces, different visualization techniques and investigate additional considerations to provide a satisfying user experience.

Keywords: Zoomable User Interface, User Experience, Semantic Zooming, Information Landscape, Information Visualization.

1 Introduction

Nowadays, accessing digital information spaces such as personal data, online databases or audio-visual media libraries is a daily activity of nearly every individual. However, information work like "writing a scientific paper" or "investigating for a news article" is a very demanding task. One reason for this is the continuously growing complexity of information spaces, resulting from the increasing quantity and heterogeneity of information objects and relations between them. Another cause is the difficulty in execution within a multifaceted individual creative workflow [KUH04] [ADA05] within today's digital information-systems. Such a workflow contains diverse activities like information seeking, information management or archiving of information objects. The majority of tools focus on assisting in single aspects of such a workflow, e.g. the very important task of information seeking (e.g. our visual information seeking system MedioVis 1.0, which is in use in the Library of the University of Konstanz[1] since 2005). Nevertheless, most of them are isolated applications that are hard to integrate into a creative workflow of a knowledge worker. Content and functionalities are

[1] http://hci.uni-konstanz.de/research/projects/MedioVis Funded by the DFG LIS 4 GZ: BIB45-INST 15176/1-1

scattered over dozens of applications, websites, storage formats, interaction models and devices – challenging the user's cognitive skills respectively. This often leads to the necessity for workarounds, resulting in a destructive degree of complexity and "information fragmentation" [KAR06].

2 System Design

Based on these requirements for creative knowledge work and inspired by the framework for mega-creativity from Shneiderman [SHN02], we designed a “Knowledge Media Workbench” [EIB06] that supports the entire workflow in one unifying workspace. Our new system, called *MedioVis 2.0*, tries to offer comprehensive visual support for all activities of creative work with digital audio-visual media libraries such as searching and browsing different information spaces (e.g. audio-visual media libraries) or keeping and managing of information objects and knowledge artefacts for later use. As an example data source we use the media specific part of the library of the University of Konstanz, consisting primarily of DVDs or VHS tapes. Additionally, we augmented this database with different online services like Google Maps[2] or the IMDb[3].

MedioVis 2.0 relies primarily on the paradigms of zoomable user interfaces and object-orientation. In consequence, no windows, menus, files or dialogs are used. To accommodate the various activities of information workers, we integrated different techniques to search and explore the information space via different visualizations. *MedioVis 2.0* offers furthermore personalization functionalities, to keep and manage gathered information objects. Finally, the concept has been designed to work on various devices like PCs, smart phones or multitouch tables, presented with one consistent user experience.

2.1 Zoomable User Interface

The fundamental visualization and interaction paradigm of *MedioVis 2.0* is the idea of a Zoomable Object-Oriented Information Landscape (ZOIL) [JET08]. Within this paradigm, an information landscape of virtually infinite size serves as basic starting point for exploration of the information space (see Fig. 1a). *MedioVis 2.0* arranges each media object corresponding to its primary genre on the landscape. Users are able to navigate in this landscape with zooming and panning operations [LIN05]. This navigation technique takes advantage of the human abilities of visual-spatial orientation and remembering visual “landmarks” [PER93]. By employing this concept, users are

[2] http://maps.google.com/

[3] http://www.imdb.com/

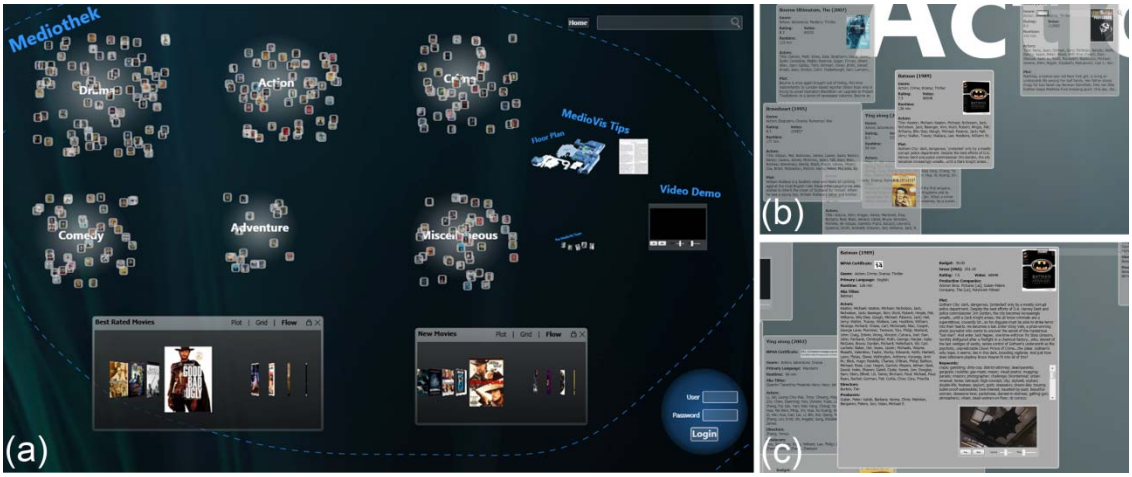


Figure 1: (a) Initial screen of MedioVis 2.0 (b) Zoomed into the action region, movie objects are represented in the second semantic zoom level (c) Most detailed semantic zoom level of a movie object, with access to a full digitalized movie.

able to utilize natural and intuitive operations as search strategy in media collections (see Fig. 1a-c).

The deeper the user zooms into the content, a “semantic zoom” reveals the more details and functionalities (see Fig. 1a-c), following the visual information-seeking mantra of Shneiderman [SHN96]: “Overview first, zoom and filter, then details-on-demand”. Thus, the available functionalities such as playing a video (see Fig. 1c) or accessing a website (see Fig. 3a) are always coupled with the information object itself, as it is proposed by object-oriented user interfaces [COL94].

2.2 Search, Filter and Explore

As another way to formulate information needs, analytical search methods are supported by MedioVis 2.0. Users are able to enter text queries into a search field on the upper right corner of the screen (see Fig. 2). With each key press, the visual representation of matching objects expands. We applied the concept of “Dynamic Queries” [AHL92] and “Sensitivity” [TWE94] for direct highlighting of objects, which still match the current query instead of removing all non-matching objects. With this technique, the attention of the user is automatically directed towards media objects of current interest. By this approach the representation of the search results is directly integrated into the landscape, instead of displaying them in an isolated visualization like a list view.

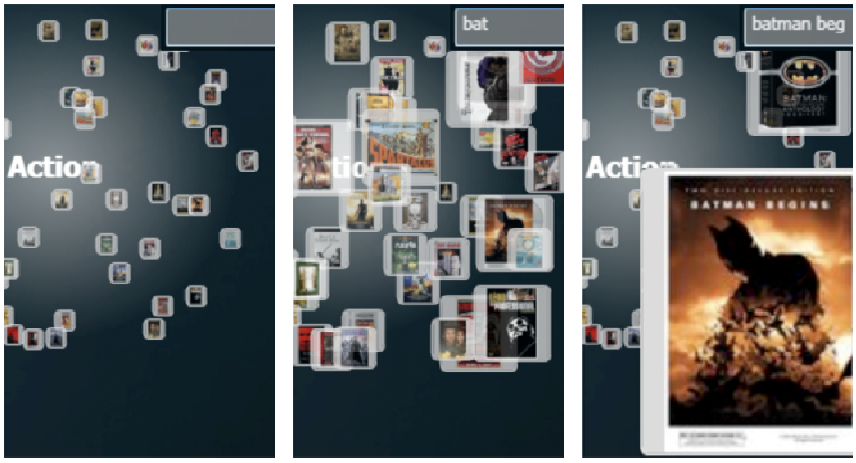


Figure 2: Query “batman begins” entered into the search field on the upper right. The size of the matching movie objects increases with every key press.

2.3 Portals and Visualizations

Portals [PER93] provide a supplementary way of exploration. By selecting an arbitrary region of the information landscape via a lasso gesture, the user creates a portal, providing a special view on the enclosed media objects. Within these portals, *MedioVis 2.0* offers multiple visualization techniques – for understanding, filtering and querying – ranging from a rapid serial visual presentation similar to a cover flow view (see Fig. 3c) [DEB00] over a scatter plot visualization called HyperScatter (see Fig. 3b) [GER08] to a table-based visualization called HyperGrid (see fig. 3a) [JET05].

The HyperGrid is a novel visualization integrating zooming concepts and an Internet browser into the well-known spreadsheet visualization. Every row, representing one media object, can be zoomed to access further information. By integrating the hyperlink-concept and an embedded Internet browser, users can immerge into the information space without losing their context.

Furthermore, portals provide visualization-independent filter mechanisms. These filters are preserved even if the user switches the visualization. By moving and scaling portals in the landscape, *MedioVis 2.0* allows to visually formulating complex queries in a direct-manipulative manner [AHL92] as proposes with the concept of magic lenses [BIE93].

Figure 3: Inside of portals a subset of movie objects can be represented by various visualizations: (a) the HyperGrid, (b) the HyperScatter and (c) a rapid serial visual presentation inspired by Apple's Cover Flow[1].

2.4 Personalization

To retain the state of a portal – with its filters and visualizations – for later use, *MedioVis 2.0* provides the possibility to lock its state and assign a name to it. Furthermore, users can reach their personal region of the information landscape (see Fig. 4) by logging into the system, revealing space to store and manage individual information artefacts. Drag & drop operations allow adding previously "locked" portals or copies of single media objects into this region. A search task is therefore no longer a transient action that is often difficult to repeat but rather a persistent object of a creative work process. Furthermore, the personal region can be organized individually by annotating, labelling and arranging artefacts the way that fits best to the user's needs.

2.5 Multiple Environments and Devices

Creative information work is a complex activity, usually executed in varying physical and social situations and environments. Therefore, a further goal of *MedioVis 2.0* is to develop an interface concept suitable for many different devices, which unifies all kinds of content and functionality with one consistent interaction model, while leaving the user the possibilities to establish own workflows, data structures or views on the information space.

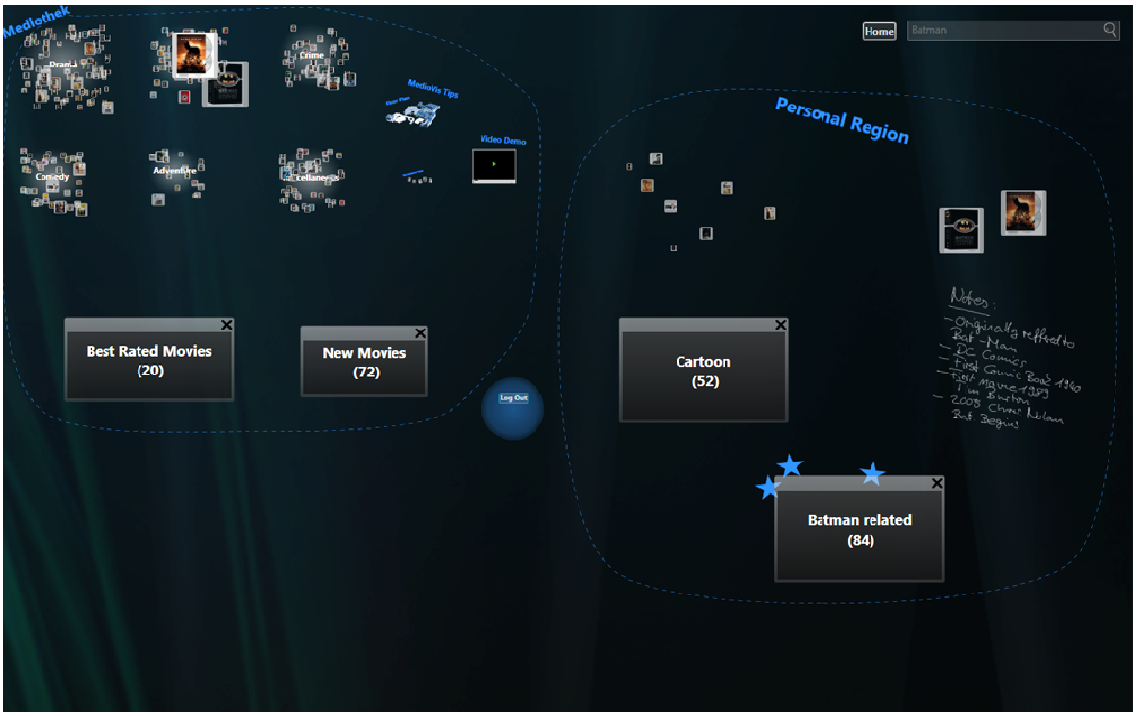


Figure 4: Users reach their personal region of the information landscape by login into the system. Herein, gathered movie objects and portals can be stored, arranged and annotated. The portals are also represented in different semantic zoom levels, depending on the zoom level of the landscape

Due to the nature of zoomable user interfaces, information presentation scales implicitly to different display sizes and is therefore applicable on very different hardware platforms. We currently run *MedioVis 2.0* on standard desktop PCs as single user workstations, on large high resolution displays which are used as public walls to enable the work in groups or project teams and on multitouch tables to further improve simultaneous multiuser interaction (see Fig. 5).

2.6 User Experience

An additional design goal of *MedioVis 2.0* is to unite all techniques and features described above in one consistent and positive user experience [KUN03]. Besides a satisfying usability we also considered several soft factors like joy of use or attractive visual design.

As every visual design communicates associations of values and functionalities [CRO03], we used a semi-transparent background for portals, so that contained objects are still perceptible. We also placed premeditatedly sized halos behind the genre clusters to generate visual landmarks. To improve the joy of use we chose animated zooming as interaction technique, using a sinusoidal instead of linear animation, to imitate a real world movement. The transition between different semantic zoom levels is accomplished by a cross-fading morphing animation. Furthermore, a parallax

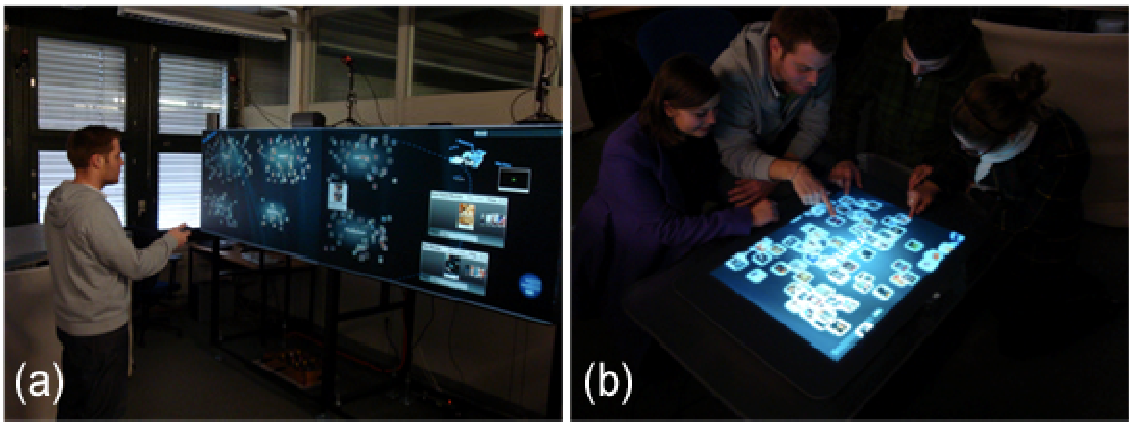


Figure 5: Besides traditionally used devices like desktop and laptop computers, *MedioVis 2.0* is designed to be used on forward-looking devices, e. g. (a) large high resolution displays or (b) multitouch tables like the Microsoft Surface[1].

background layer is placed behind the information landscape that zooms with a smaller factor to arouse the feeling of depth and speed and to enhance the orientation in the landscape.

Eventually, the design of *MedioVis 2.0* encourages the explorative and playful discovery of information objects or novel functionalities during the overall navigation process in the information landscape.

3 Outlook

MedioVis 2.0 represents a novel perspective on how to implement a comprehensive Knowledge Media Workbench through the use of the zoomable user interface paradigm and object-oriented user interfaces. Thereby, knowledge workers are able to accomplish activities within one consistent system. *MedioVis 2.0* provides a unified user experience, not only on a desktop PC, but also on different devices such as multitouch tables and large high-resolution displays. To further evaluate the potential of the concept, we will transfer *MedioVis 2.0* to other complementary devices like mobile gadgets (e.g. smart phones, netbooks), which already play an increasing role in creative workflows of many knowledge workers.

Reality based interaction [JAC08] and the combination or “blending” of real world artefacts with digital information objects will also be of particular importance in our future research. For example the user can put a DVD on the surface and related content is displayed, e.g. the famous actors or actresses, producer, etc (Fig. 6a). This information can now become a starting point to explore the digital collection. Output becomes input, e.g. selecting one or more actors or actresses presents all matching movies on the information landscape. This query by example approach allows the

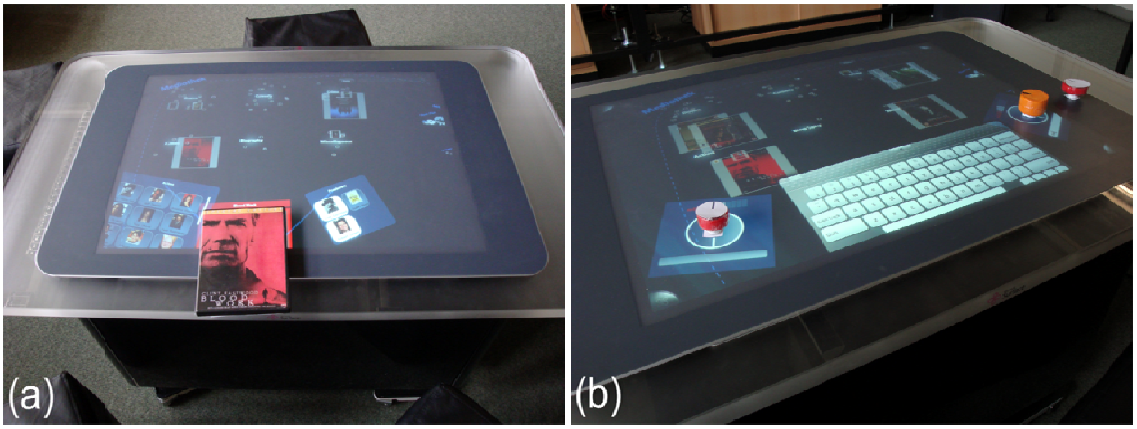


Figure 6: Integrating real world objects into the interaction offers new ways of exploration: (a) A DVD can be used to find related content, by laying it onto the surface and using the provided filter panels; (b) Serch terms can be embodied by rotary knobs, their weights can be adjusted by turning these tokens.

blending of real world objects with the power of digital content and thereby allows a smooth transition between a searching and browsing mode.

Another idea is to develop specific kind of search tokens to manually weight the search terms (Fig. 6b). The user puts the tangible search token embodied by a rotary knob onto the screen and then enters the search term(s) via the automatic launched onscreen keyboard. Turning the knob to the right increases the weight of the term and so the matching objects increase in size, turning to the left decreases the weight. This way the user can combine a variety of search tokens standing for different search terms and each term can get an individual weight.

Despite of open issues regarding the creation of knowledge artefacts or collaboration, which we want to approach in future work, we believe that our Knowledge Media Workbench – *MedioVis 2.0* – offers reasonable possibilities to support individual creative workflow of knowledge workers.

4 References

[ADA05] Adams, A. and Blandford, A. 2005. Digital libraries' support for the user's 'information journey'. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (Denver, CO, USA, June 07 - 11, 2005). JCDL '05. ACM, New York, NY, 160-169.

[AHL92] Ahlberg, C.; Williamson, C.; Shneiderman, B. 1992: Dynamic queries for information exploration: an implementation and evaluation. CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press.

[BIE93] Bier, E. A.; Stone, M. C.; Pier, K.; Buxton,W.; DeRose,Tony D. 1993: Toolglass and magic lenses: the see-through interface. SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques.

[COL94] Collins, D. 1994. Designing Object-Oriented User Interfaces. Benjamin-Cummings Publishing Co., Inc.

[CRO03] Crow, D. 2003. Visible Signs: An Introduction to Semiotics. AVA Publishing SA, Lausanne.

[DEB00] de Bruijn, O. and Spence, R. 2000. Rapid serial visual presentation: a space-time trade-off in information presentation. In Proceedings of the Working Conference on Advanced Visual interfaces.

[EIB06] Eibl, M.; Reiterer, H.; Friedrich, Stephan, P. F.; Thissen, F. 2006: Knowledge Media Design: Theorie, Methodik, Praxis. Oldenbourg; Auflage: 2.

[GER08] Gerken, J.; Demarmels, M.; Dierdorf, S.; Reiterer, H. 2008. HyperScatter – Modellierungs- und Zoomtechniken für Punktdiagramme. M&C 2008, Oldenbourg Verlag, München.

[JAC08] Jacob, R. J. K.; Girouard, A.; Hirshfield, L.M.; Horn, M.S.; Shaer, O.; Solovey, E.T.; Zigelbaum, J. 2008: Reality-Based Interaction: A Framework for Post-WIMP Interfaces. CHI '08. ACM, New York.

[JET05] Jetter, H.-C.; Gerken, J.; König, W.; Grün, C.; Reiterer, H. 2005: HyperGrid - Accessing Complex Information Spaces. People and Computers XIX - The Bigger Picture, HCI 2005, Springer Verlag.

[JET08] Jetter, H.-C.; König, W. A.; Gerken, J.; Reiterer, H. 2008. ZOIL - A Cross-Platform User Interface Paradigm for Personal Information Management. Personal Information Management 2008: The disappearing desktop.

[KAR06] Karger, D. R. and Jones W. 2006. Data unification in personal information management. Commun. ACM, 49(1): 77-82.

[KUH04] Kuhlthau, C. C. 2004. Seeking meaning: a process approach to library and information services, volume 2nd Edition. Libraries Unlimited.

[KUN03] Kuniavsky, M. 2003. Observing the User Experience: A Practitioner's Guide to User Research. Morgan Kaufmann.

[LIN05] Lindell, R. and Larsson, T. 2005. The Data Surface Interaction Paradigm, Theory and Practice in Computer Science, Eurographics Association.

[PER93] Perlin, K. and Fox, D. 1993. Pad: an alternative approach to the computer interface. In SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques, ACM Press.

[SHN96] Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the IEEE Symposium on Visual Languages, Washington. IEEE Computer Society Press.

[SHN02] Shneiderman, B. 2002. Leonardo’s Laptop, Cambridge MA. MIT Press.

[TWE94] Tweedie, L., Spence, B., Williams, D., Bhogal, R., 1994. The attribute explorer. In: CHI '94: Confer-ence companion on Human factors in computing systems. New York, NY, USA, ACM, April 1994, S. 435-436.

SIVA Suite – Konzeption eines Frameworks zur Erstellung von interaktiven Videos

B. Meixner, B. Siegel, G. Hölbling, H. Kosch, F. Lehner

Universität Passau
Lehrstuhl für Verteilte Informationssysteme
Lehrstuhl für Wirtschaftsinformatik II

{meixner,hoelblin,kosch}@fim.uni-passau.de
{siegel,lehner}@uni-passau.de

Zusammenfassung: Dieser Beitrag stellt ein Framework zur Produktion von interaktiven Videos vor. Mit Hilfe des Systems kann der Autor Videomaterialien editieren und einzelnen Szenen Annotationen wie Bilder, Texte, Hyperlinks oder weitere Videos anfügen. Das Autorentool verfügt über ein Videoschnittwerkzeug zur Definition von Szenen, einem Szenengraphen zur Ablaufkontrolle sowie verschiedenen Editoren zur Erzeugung und Bearbeitung der Annotationen. Besonderer Fokus bei der Entwicklung soll auf einfache Bedien- sowie Erlernbarkeit gelegt werden, um das Tool auch für Laien zugänglich und nutzbar zu machen.

Schlagwörter: Video, Interaktivität, Framework, Autorentool, interaktives Video, interaktives Lernen, E-Learning, Autorensystem, Annotation, Videoannotation

1 Einführung

Die Nutzung von Videos im Web nimmt seit Jahren stark zu, allerdings müssen sich die Nutzer bisher mit dem bloßen passiven Konsum der Medien begnügen. Allerdings existieren Forschungsansätze, Videos mit zusätzlichen interaktiven Features anzureichern. Beispielhaft sei hier YouTube Annotations [YT09] genannt (weitere Beispiele siehe [HR06] und [LF08]). Aufgrund des erkennbaren Trends zu derartigen Veränderungen der Medienlandschaft wurde mit der SIVA Suite (Simple Interactive Video Authoring Suite) ein Beitrag zum Forschungsfeld „Interaktive Videos“ in Angriff genommen. Das System ist in der Lage interaktive Videos zu erstellen, die eine Vielzahl von Vorteilen für den Betrachter bieten, wie die Anzeige von weiterführenden Informationen oder zusätzlichen Steuerungsmöglichkeiten.

Der Nutzen von interaktiven Videos für den Anwender wird im Folgenden an verschiedenen Einsatzszenarien näher erläutert. Im weiteren Beitrag soll ein Überblick über das Konzept der SIVA Suite mit einer detaillierten Darstellung der Elemente Producer und Player gegeben werden. Der Beitrag schließt mit einem Ausblick auf die Weiterentwicklung des Tools.

1.1 Einsatzszenarien für interaktive Videos

Grundsätzlich ist die Integration von interaktiven Videos in den unterschiedlichsten Einsatzszenarien vorstellbar. Am besten lassen sich die Vorteile der Technologie an den Bereichen E-Learning und Tourismus erläutern.

Videomaterialien finden in vielen E-Learning-Applikationen Anwendung, wobei sie bisher mit sehr wenigen Interaktionsmöglichkeiten auskommen. In den meisten Fällen sind das bloße abspielen, stoppen und pausieren des Videos die einzigen Eingriffsmöglichkeiten für den Betrachter. Damit kann man allerdings dem oftmals geforderten Ansatz einer explorativen und problemorientierten Lernumgebung nicht gerecht werden. Tiefer greifende Interaktivität kann den Nutzen eines E-Learning-Videos erheblich steigern. Durch die Aufspaltung des Ursprungsvideos in selbst definierte Szenen kann deren Abfolge beispielsweise an das Vorwissen verschiedener Nutzergruppen angepasst werden. So können unterschiedliche Versionen desselben Materials sowohl für fortgeschrittene Lerner als auch für Anfänger angeboten werden. Das Angebot von zusätzlichem, weiterführendem Inhalt wie Detailansichten, Formeln oder erklärendem Text kann dem Betrachter das Verständnis des dargebotenen Lernstoffes weiter erleichtern. Gleichzeitig verlässt der Betrachter seine passive Rolle, da beispielsweise durch Tests innerhalb der Videodarbietung sofortiges Feedback sowohl an den Lerner als auch an den Dozenten ermöglicht wird.

Im Tourismussektor können interaktive Videos beispielsweise für die Erstellung von virtuellen Stadtrundgängen oder Museumsführungen benutzt werden. Historische Gebäude oder Plätze können mit Zusatzmaterialien und weiteren Videos verlinkt werden. Die Annotation einer Videosequenz könnte aus erklärendem Text oder verschiedenen Ansichten eines Gebäudes im Video bestehen. Verlinkungen können genutzt werden, um zum Beispiel direkten Zugriff auf die Menükarte oder die Öffnungszeiten eines Restaurants zu ermöglichen. Sinnvoll ist hier die Umsetzung von parallelen Handlungssträngen. So kann bei Stadtrundgängen an Wegekreuzungen ausgewählt werden, in welcher Richtung der Nutzer seinen Weg fortsetzen möchte.

Einige Funktionen werden für beide Szenarien interessant sein, beispielsweise die Anzeige von Annotationen, die für die gesamte Dauer des Videos gültig sind. Häufig sind dies Informationen zu Autor und Titel des Videos oder auch Logos oder Homepage-Links. Sprünge zwischen Videos oder Videosequenzen könnten genutzt werden, um Videoinhalte aus einer anderen Perspektive zu zeigen.

1.2 Annotationen

Als Annotation bezeichnet man eine ergänzende Information im Video, die im zugrunde liegenden Basisvideo nicht enthalten ist. Verfügbare Annotationstypen sind Text, Bild, Video, Richtext, Links auf externe Webseiten und Buttons zur Kontrolle des Videofortschritts. Die Markierung von relevanten oder klickbaren Objekten im Video

sowie der Einsatz von Zeitlupe lenken den Fokus des Betrachters auf einzelne Details. Annotationen können für drei unterschiedliche Zeit-Ebenen gelten: sie können für die gesamte Dauer des Videos gültig sein (Video-Annotationen), für die Dauer einer Sequenz (Video-Sequenz-Annotation) oder sich auf ein Objekt im Video beziehen (Video-Objekt-Annotation). Im letzten Fall wäre die Gültigkeit durch eine bestimmte Anzahl an Frames bestimmt.

Annotationen können das Videomaterial auf verschiedene Weise beeinflussen. Durch das Aktivieren eines Zusatzinhalts kann das Video entweder gestoppt oder fortgesetzt werden, eine Sprungmarke kann angesteuert oder ein neues Browserfenster geöffnet werden (hierbei würde das Video in jedem Fall stoppen). Die Aktivierung wiederum kann durch ein Event ausgelöst werden oder durch eine Nutzerinteraktion.

2 Konzept der SIVA Suite

Die SIVA Suite besteht aus drei Bausteinen – dem SIVA Producer, dem SIVA Player und dem SIVA Server. Der SIVA Producer bietet alle Funktionalitäten, die für das Authoring eines interaktiven Videos benötigt werden. Dazu zählen unter Anderem das Schneiden von Videos, das Erstellen eines Szenengraphen und das Erstellen und Bearbeiten von Annotationen. Das im SIVA Producer erstellte interaktive Video-Projekt wird durch eine XML-Datei beschrieben, die die Interaktivität, die Annotationen und die Beziehungen zwischen den Szenen festlegt. Aufgrund der weiten Verbreitung und Unterstützung des Flash-Plugins werden alle Videoinhalte im Flash Video Format (flv) gespeichert. Die Hauptaufgabe des SIVA Players ist es, die XML-Datei auszuwerten und so das interaktive Video-Projekt abzuspielen, das vom SIVA Server bereitgestellt wird.

2.1 SIVA Producer

Der SIVA Producer besteht aus vier Komponenten: einer Projektmanagement-Komponente, der Medienverarbeitung mit Schnittfunktionen, einem Szenengraph-Editor zum Abbilden von Handlungssträngen sowie einem Annotationseditor für die Bearbeitung von Zusatzinformationen. Diese Komponenten werden im Folgenden näher beschrieben und erklärt:

- *Projekt-Management:* Das Projekt-Management bietet Funktionalitäten zum Organisieren von projektspezifischen Daten. Es bietet mittels Repositories eine Übersicht über alle geladenen Ressourcen sowie alle bereits definierten Szenen. Außerdem sind eine Speicher- und Ladefunktion sowie eine Export-Funktion verfügbar. Dabei wird das gesamte Videomaterial des Projekts in flv-Dateien umgewandelt. Die Annotationen zu Szenen und der Ablauf des Videos werden in eine XML-Datei exportiert. Die Ressourcen werden in einer vordefinierten

Ordnerstruktur abgelegt. Das exportierte Projekt kann dann vom Player interpretiert werden.

- *Medienverarbeitung:* Dieser Teil der Software bietet Funktionen die benötigt werden, um auf Medien zuzugreifen, sie zu bearbeiten und anzusehen. Mit Hilfe des Schnittwerkzeuges kann der Autor Szenen selbst definieren. Dazu legt er einen Start- und einen Endzeitpunkt sowie einen Namen fest. Um eine bessere Usability zu erreichen, wird eine Zeitleiste mit Vorschau-Bildern des Videos zur Szenendefinition verwendet (siehe Abbildung 1). Außerdem ist eine Shot-Detection-Funktion in das Schnittwerkzeug integriert.

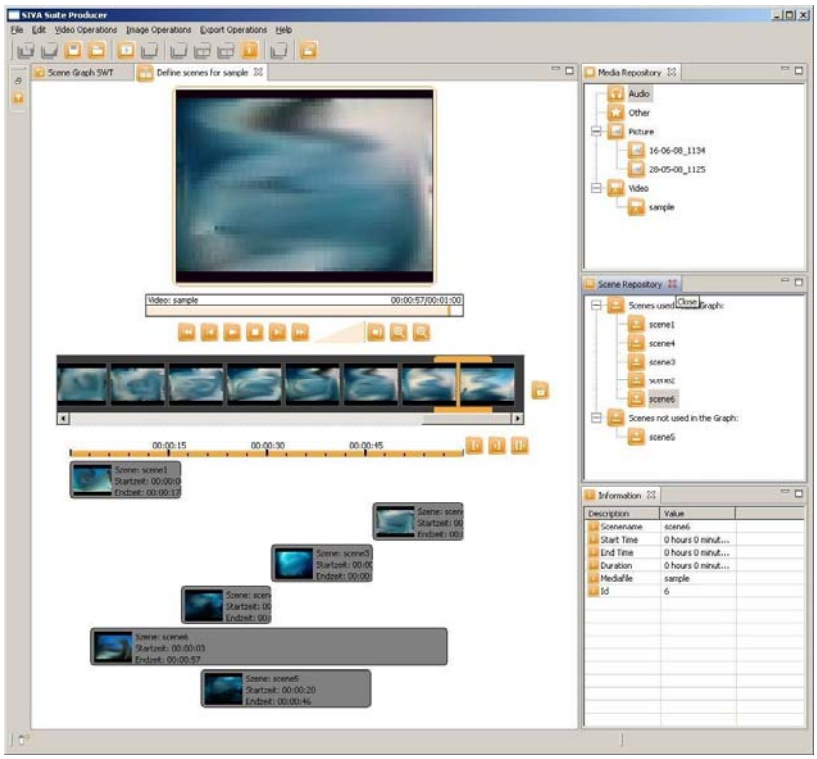


Abbildung 1: Komponente zur Medienverarbeitung

- *Szenengraph-Editor:* Bereits definierte Szenen werden in einem Szenenrepository angeordnet, aus welchem sie via Drag&Drop in den Szenengraphen eingefügt werden können. Dieser Editor erlaubt es dem Autor die Szenen in einem Graphen anzuordnen, wodurch Sprünge im Video und nicht lineare Verläufe visualisiert werden (siehe Abbildung 2). Szenen im Szenengraph können mit Bildern, Videos, Links, Buttons, HTML-Dateien oder Text annotiert werden. Die Toolbar des Szenengraph-Editors bietet Tools zum Löschen und Hinzufügen von Knoten und Kanten.

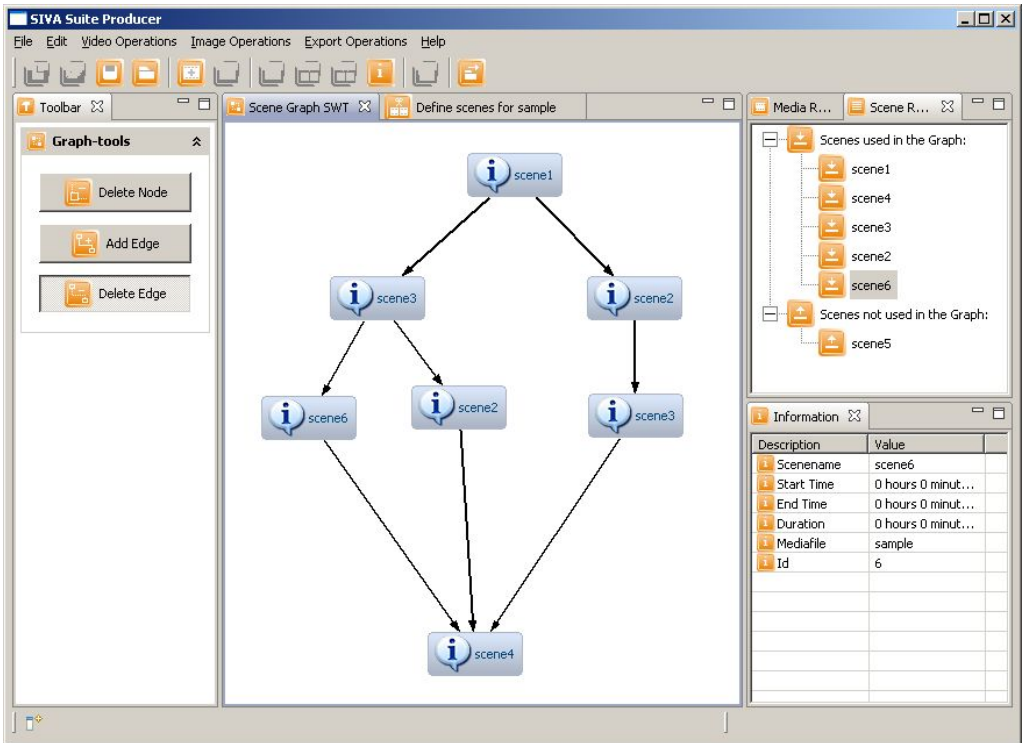


Abbildung 2: Szenengraph-Editor des SIVA Producer

- *Annotationseditoren:* Im SIVA Producer sind drei Annotationseditoren inplementiert. Der Richtext-Editor ist ein WYSIWYG-HTML-Editor, der es ermöglicht, formatierten Text mit Bildern und Links zu gestalten. Die so gestaltete Seite kann in eine HTML-Datei exportiert werden, die dann als Annotation mit einer Szene verknüpft wird. Zudem werden ein Bild-Editor mit grundlegender Funktionalität und ein Plaintext-Editor angeboten. Vordefinierte Inhalte, die bereits ins Projekt geladen wurden und sich im Ressource-Repository befinden, können ebenso als Annotation an eine Szene angehängt werden.

2.2 SIVA Player

Der SIVA Player ist jene Komponente der Suite, welche die Schnittstelle des Videoprojekts zum Endanwender darstellt. Das vom Autor im SIVA Producer erstellte interaktive Video wird über die Einbindung des Players in dessen Webauftritt dem eigentlichen Endkunden präsentiert.

Die Einbindung des Players als Kompontente in bestehende Webseiten war eine grundlegende Anforderung, die direkte Konsequenzen auf die für die Entwicklung verfügbaren Technologien hatte. Der wesentliche Aspekt bestand darin, dass der Endkunde das angebotene Medium problemlos und ohne weitere technische Hürden nutzen können sollte. Die Wahl für die Entwicklungsumgebung des SIVA Players fiel

deshalb auf Adobe Flex. Zwar wird für die Ausführung der in Flex erstellten Anwendungen ebenfalls ein Plugin benötigt – allerdings handelt es sich hierbei um das Flash Plugin. Dieses ist in Version 10 laut Adobe bereits auf 75,3% (in Version 9 sogar 98,6%, Stand März 2009) der Rechner in Europa verfügbar [ADB09], sodass hier der befürchtete Zusatzaufwand zu vernachlässigen ist. Wäre dies das alleinige Entscheidungskriterium, so hätte auch Adobe Flash zur Entwicklung verwendet werden können – jedoch gilt Flex aufgrund seiner speziellen Eignung für Rich Internet Applications (RIA) für den geplanten Zweck als besser geeignet.

Die Entwicklung des SIVA Players wird in zwei Schritten vorgenommen. An erster Stelle steht die Erstellung der Nutzeroberfläche, die folgende Elemente berücksichtigt:

- *Fenstermodus und Vollbildmodus:* Im Fenstermodus werden Annotationen in Infobereichen angezeigt, die links, rechts, ober- und unterhalb des Videobereichs angeordnet sind. Der Vollbildmodus streckt die Anzeige über das gesamte Browserfenster – er verlässt somit den Rahmen der Website. Die Infobereiche überlagern in diesem Modus das Video zu bestimmten Anteilen.

- *Videosteuerung:* Bekannte Elemente zur Videosteuerung werden um spezifische Steuermöglichkeiten für das interaktive Video, wie Szenensprünge oder Sprachenwahl erweitert.

- *Ein-/Ausblenden der Annotationen:* Dem Nutzer wird die Möglichkeit gewährt, die Zusatzinhalte nur bei Bedarf einzublenden. Ausgenommen hiervon sind Annotationsformen, die den Ablauf des Videos in Abhängigkeit von Nutzerentscheidungen steuern.

Im Anschluss an die Nutzeroberfläche wird die Integration der interaktiven Funktionen vorgenommen. Alle Daten, wie Wiedergabereihenfolge, Events und Ressourcen, die vom interaktiven Videoprojekt benötigt werden, bezieht der SIVA Player über das XML-Dokument, welches vom SIVA Producer generiert wird. Aufgrund der Modularität sowie der schrittweisen Implementation der Funktionen im SIVA Producer sollen im Folgenden nur die Primärfunktionen der Playerkomponente dargestellt werden. Diese sind:

- Anzeige von Annotationen (HTML, Plaintext, Bilder, Untertitel) sowie Zuordnung zum relevanten Infobereich

- Anzeige der Annotationen in der Defaultsprache des Projekts, sowie Umschaltung der Sprache nach Nutzerinteraktion

- Bereithalten der Videosequenzen zur Steuerung des Projektablaufs in Echtzeit

- Vereinfachte Darstellung des Szenengraphen zur erweiterten Steuerung des Projektablaufs

3 Ausblick – Weiterentwicklung und Nutzung

Die bisherige Projektarbeit generierte folgende Forschungsthemen:

- *Metadaten-gesteuertes Modell:* Zur Darstellung von Abläufen und Zusatzinformationen mittels XML-Dateien wird ein von Metadaten gesteuertes Modell umgesetzt.
- *Ablauflogik:* Zur Speicherung des Fortschritts bei vorzeitiger Beendigung des Videos wird eine Ablauflogik in die XML-Struktur integriert.
- *Performanzanalyse:* Die Zugriffe auf XML-Dateien werden auf Performanz untersucht. Aufgrund der gewonnenen Ergebnisse können diese an die Anwendungsszenarien angepasst werden.
- *Objekterkennung und Texterkennung:* Objekte (z. B. Autos) und Texte (z. B. auf Straßenschildern) im Video können automatisch erfasst und für weitere Annotationen verwendet werden.
- *Objektmarkierung:* Interessante Objekte im Video können vom Autor wahlweise mit sichtbaren oder unsichtbaren Markierungen belegt werden. Diese Markierung wird zunächst manuell vorgenommen, allerdings ist auch eine automatische Erkennung anhand von Bildmerkmalen geplant.
- *Objektverfolgung:* Die Markierung über einem sich bewegenden Objekt wird im Verlauf des Videos angepasst. Dies ist zum einen manuell durch die Definition von Schlüsselstellen möglich. Hierbei wird eine Bewegung durch so genanntes „Tweening“ interpoliert. Zum anderen kann – aufbauend auf der automatischen Objektmarkierung – auch die Objektverfolgung automatisch anhand von Bildmerkmalen vorgenommen werden.

Aufgrund der Komplexität der genannten Forschungsthemen wurde die Entwicklung des Frameworks zur Erstellung von interaktiven Videos in zwei Phasen aufgeteilt. Die Entwicklung konzentriert sich zunächst auf eine saubere Konzeption sowie auf die Implementierung der Grundfunktionen. Diese Funktionen beinhalten bereits alle unter Kapitel 2 beschriebenen Eigenschaften wie das Laden und Editieren von Annotationen und das Hinzufügen zu Szenen, die Segmentierung von Videomaterial in Szenen, die Umsetzung von mehrsprachigen Projekten sowie den Export des Projekts in eine XML-Datei.

Den größeren Mehrwert für die geplanten Einsatzbereiche bieten komplexere Funktionen, die sich direkt aus den oben genannten Forschungsfragen ergeben. Diese werden daher in einem zweiten Schritt implementiert. Neben der Implementierung von neuen Funktionen wird auch die Verbesserung von bestehenden Funktionen und der Nutzerfreundlichkeit einen Schwerpunkt der weiteren Entwicklung darstellen.

4 Literaturverzeichnis

[ADB09] Adobe Flash Player Version Penetration. http://www.adobe.com/products/player_census/flashplayer/version_penetration.html 11.05.2009.

[HR06] Hammoud, R. (Hrsg.) Interactive Videos. Algorithms and Technologies. Berlin, Springer 2006

[LF08] Lehner, F., Siegel, B., Müller, C. und Stephan, A. Interaktive Videos und Hypervideos – Entwicklung, Technologien und Konzeption eines Authoring-Tools. Diskussionsbeitrag W-28-08 der Schriftenreihe Wirtschaftsinformatik. Passau, 2008

[YT09] YouTube Annotations. http://www.youtube.com/t/annotations_about 11.05.09

Online-Werbung als digitales Kulturgut: Analyse, Erschließung und Archivierung

Christian Wolff

Universität Regensburg
Institut für Information und Medien, Sprache und Kultur
Professur für Medieninformatik

christian.wolff@computer.org

Zusammenfassung: Der Beitrag charakterisiert Online-Werbung als mittlerweile wichtigen Teil des Werbemixes und stellt die wichtigsten Felder des Online-Marketing vor. Online-Werbung wird als relevantes Handlungsfeld der Medieninformatik eingeordnet und davon ausgehend werden Fragen der Analyse und Erschließung von Online-Werbung knapp eingeführt. Kriterien, die für den Aufbau von Inline-Werbearchiven relevant sind, werden vorgestellt.

Schlagwörter: Online-Werbung, Marketing, Kulturgut, Archive, Erschließung

1 Einleitung

Dass Werbung als Kulturgut betrachtet werden kann, ist vor dem Hintergrund der Diskussion um das Entstehen der Kulturindustrie sicher keine Selbstverständlichkeit. Man darf annehmen, dass die grundsätzliche Kritik an der Kulturindustrie, wie sie in Adorno / Horkheimers *Dialektik der Aufklärung* [ADH69] zum Ausdruck kommt, eine intensivere Auseinandersetzung mit dem Phänomen Werbung gerade in den Geistes- und Kulturwissenschaften in den letzten Jahrzehnten behindert haben mag. Möglicherweise erklärt sich so die relative Zurückhaltung dieser Disziplinen gegenüber dem Thema Werbung im Vergleich mit Disziplinen wie Ökonomie (Marketing) oder Psychologie. Diese Vermutung ist deshalb relevant, weil der Hintergrund für den vorliegenden Beitrag ein Forscherverbund ist, der seinen Schwerpunkt gerade in den angesprochenen Fächergruppen der philosophischen Fakultäten an der Universität Regensburg hat: Der *Regensburger Verbund für Werbeforschung* (RVW, http://www.werbeforschung.org) hat sich 2006 als interdisziplinärer Zusammenschluss einer Vielzahl von Fächern konstituiert (u. a. Psychologie, Germanistik, Amerikanistik, Medien- und Kulturwissenschaft, Musik- und Kunstwissenschaft, Medieninformatik, Werbepraxis). Sein Fokus lag zunächst auf der Analyse von Hörfunkwerbung, da der Aufbau des *historischen Werbefunkarchivs* (HWA) an der Universität Regensburg den Anstoß zum Aufbau dieser Forschergruppe gab. Mittlerweile gehören zum Gegenstandsbereich des RVW alle Erscheinungsformen der Werbung, einschließlich der Online-Werbung.

Für die *Medieninformatik* ist die Auseinandersetzung mit Werbung in mehrfacher Hinsicht ein interessantes Unterfangen:

- Zum einen ist informationstechnologische Kompetenz gefragt, wenn Probleme wie Digitalisierung, Erschließung und Archivierung von (traditionellen) Werbemitteln zu lösen sind – das ist die typische Rolle der Informatik als einer *ancilla scientiae vel philosophiae*.

- Liegen Werbemittel digital vor, kann ggf. mit automatisierten / algorithmischen Verfahren ein Beitrag zur Analyse und Erschließung geleistet werden, der über die Werkzeugunterstützung traditioneller Verfahren hinausgeht. Hier lassen sich Verfahren aus dem Bereich maschinelles Lernen oder Mustererkennung z. B. aus dem Bereich der Bildanalyse auf visuelle Werbeformen übertragen. In dasselbe Paradigma fällt etwa die Anwendung von Evaluationsmethoden mit Blickbewegungsanalysen auf Werbemittel.

- Schließlich ist offenkundig, dass Online-Werbung neue Möglichkeiten der Produktion, Distribution und Adaption bietet („Targeting"), die sich technischer Mittel bedienen. Nicht zuletzt sind interaktive und multimediale Werbeformen sie ist damit ein genuines Arbeitsfeld der Medieninformatik.

Ein weiterer Aspekt ist, dass Werbung als Gegenstand der Wissenschaft eine ähnliche Entwicklung genommen hat wie das (Produkt-)Design, das mit Bezug zu digitalen Medien ein anderes Arbeitsfeld der Medieninformatik konstituiert: Beide Bereiche sind im Zuge der Entwicklung industrieller Fertigungsprozesse und Vermarktungsmöglichkeiten entstanden, sind aber erst schrittweise Gegenstand wissenschaftlicher Betrachtung (außerhalb der Werbedisziplinen Marketing und Psychologie) geworden – für beide – Design wie Werbung – ist ihre „Verwissenschaftlichung" noch nicht abgeschlossen [BÜR05, S.276].

2 Formen und wirtschaftliche Bedeutung der Online-Werbung

Online-Werbung wird üblicherweise auf drei Bereiche aufgeteilt (s.u. Abb. 1, [OVK09, S. 7]):[1]

- Die verschiedenen Formen der *Affiliate-Werbung*, gewissermaßen eine moderne Variante der Vertriebspartnerschaft, bei der auf einer Website die Produkte eines

[1] [LAM06] nimmt eine feinere Untergliederung vor, und sieht E-Mail-Marketing und Suchmaschinenoptimierung als weitere Marketingformen, wobei er sich allerdings dezidiert gegen die Verwendung des Begriffs „Online-Werbung" als Oberbegriff für verschiedene Marketingformen ausspricht [LAM06, S. 122].

Vertriebspartners beworben werden. Am ältesten und bekanntesten sind die Aktivitäten des Online-Buchhändlers Amazon in diesem Bereich [LAM06, S. 23].

- Das *Suchmaschinen-Marketing*, bei dem Firmen dafür bezahlen, dass ihre Anzeigen dann in einer Suchmaschine geschaltet werden, wenn Benutzer bestimmte Wörter als Suchbegriffe eingeben, wodurch sich eine individuell auf die jeweilige Suchanfrage zugeschnittene Konstellation von Anzeigen für den Benutzer ergibt.

- Die Online-Werbung i.e.S., insbesondere durch Banner, wo visuelle Werbeflächen auf Websites in unterschiedlichen Formaten und Kontexten zum Einsatz kommen. Hier hat sich ein reiches Formeninventar entwickelt (z. B. *Skyscraper*, *Wallpaper* oder *Eselsohren*, vgl. [Wol09] und zu einer kommunikationstheoretisch begründeten Typologie der Online-Werbung [MCM07]).

Nachfolgend wird nur die „eigentliche“ Onlinewerbung, also die Verwendung visueller und / oder akustischer Werbemittel auf Websites als für die Analyse und Archivierung relevant erachtet. Es sollte dabei klar sein, dass zu einer ganzheitlichen Betrachtung von Werbung als Kulturgut auch der Einsatzkontext gehört: Wie bei einer Printwerbung der redaktionelle Kontext nicht ohne Relevanz ist (oder der Sendeplatz eines Werbespots), so ist auch für Online-Werbung nicht nur zu beachten, auf welcher Website sie geschaltet wird, sondern aufgrund der in aller Regel gegebenen Interaktivität auch die Gestaltung der Zielwebsite, auf die der Benutzer durch Klick auf ein Banner gelangt.

Die mehrmals jährlich publizierten Marktdaten des Online-Vermarkterkreises (OVK) im Bundesverband digitale Wirtschaft (BVDW) belegen den quantitativen Bedeutungszuwachs der Online-Werbung in den vergangenen Jahren eindrucksvoll: Allein von 2007 nach 2008 war ein Wachstum von 25 % zu beobachten, selbst für 2009 wird trotz Wirtschaftskrise noch mit einem Zuwachs um 10 % gerechnet (siehe Abbildung 1). Innerhalb weniger Jahre hat sich die Online-Werbung einen prominenten Platz im Spektrum der unterschiedlichen Werbeformen erobert: Die Marktdaten des OVK zeigen für die Jahre 2005 bis 2008 für den Anteil der Online-Werbung am Werbemarkt eine Steigerung von 4,4 % auf 14,8 %, womit Online-Werbung 2008 bereits auf Platz vier der wichtigsten Werbeträger liegt (Abbildung 2). Man kann argumentieren, dass Online-Werbung damit zu einem der wirtschaftlich sichtbarsten Indikatoren der „Informatisierung des Alltags“ (Friedemann Mattern [MAT07]) geworden ist.

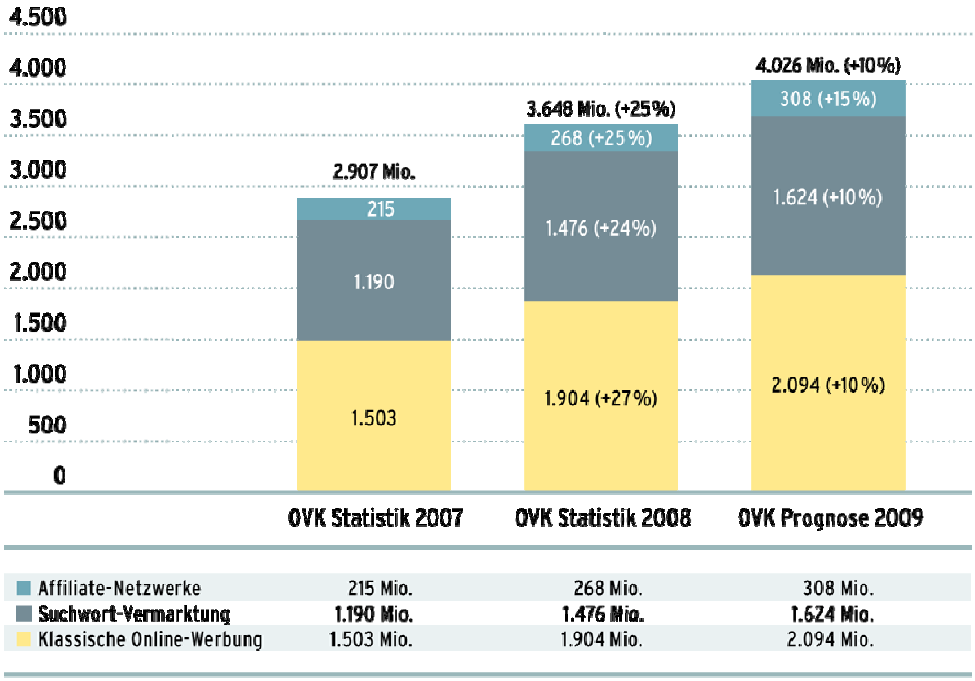


Abbildung 1: Werbestatistik für Onlinewerbung nach Segmenten [OVK09, S. 7]

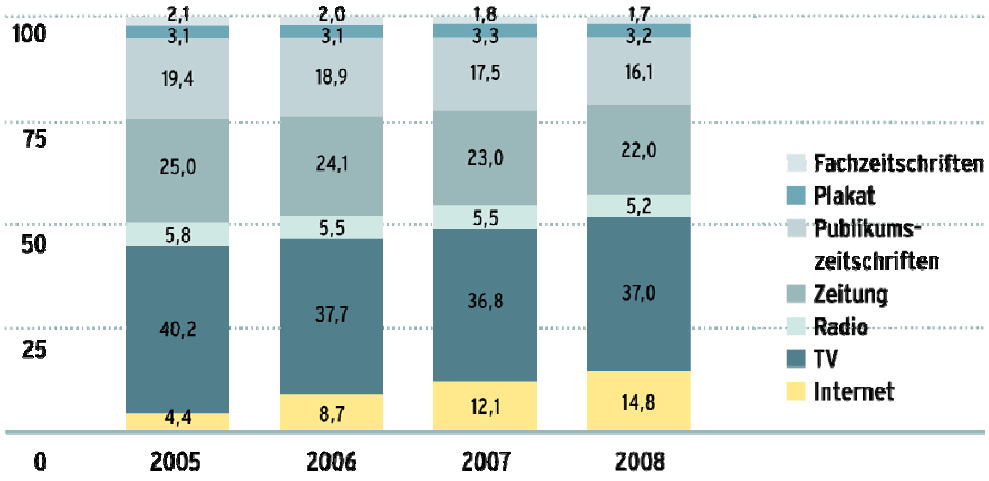


Abbildung 2: Entwicklung des Mediamix im deutschen Werbemarkt 2005 – 2008 [OVK09, S. 9]

Selbst wenn man Bedenken hinsichtlich der kulturellen Wertigkeit von Werbung geltend machen wollte, könnte allein die hohe wirtschaftliche (und damit auch soziale) Bedeutung der Online-Werbung die wissenschaftliche Beschäftigung mit ihr rechtfertigen. Gerade die Auseinandersetzung mit den visuellen Aspekten der Online-Werbung gehört aber zum Arbeitsprogramm einer modernen Bildwissenschaft, die sich nicht auf einen traditionellen Kunstbegriff oder gar das Tafelbild als Gegenstandsbereich reduzieren lassen will [SAH05], [SAH06].

3 Analyse von Online-Werbung

Die Werbeforschung hat unterschiedliche Ansätze zur Analyse von Werbung entwickelt; das Integrationsmodell von Janich [JAN05] führt ältere Ansätze zusammen und sieht für die (sprachwissenschaftliche) Analyse von Werbung mehrere, aufeinander aufbauende Analyseschritte vor, die textexterne Faktoren, Aufbau und Struktur sowie Inhalt und Bedeutung von Teiltexten verbal, paraverbal und visuell untersuchen. Auf die Analyse folgen Syntheseschritte, die zunächst textinterne Faktoren, dann die Wechselwirkung von textinternen und textexternen Faktoren und schließlich die übergreifende Werbewirkung im Wechselspiel von Werbeinhalt und Werbeintention in Betracht ziehen. Auf diesem Modell aufbauend hat Reimann [Rei08] eine Modellerweiterung vorgelegt, die vor allem die Mehrmedialität von Werbung bei komplexen Kampagnen in den Mittelpunkt rückt. Für den Bereich der Online-Werbung hat Grimm [GRI09] für den „klassischen“ Fall der Online-Banner diese Modelle weiterentwickelt: Dabei ist unmittelbar einsichtig, dass medienspezifische Besonderheiten der Online-Werbung, insbesondere Animationen und Interaktivität, in den bisherigen Modellen nicht hinreichend berücksichtigt sind. Mit der Arbeit von Grimm liegt mithin ein erstes und an einem kleinen Corpus von Bannern erprobtes Analysemodell vor, das für andere Formen der Online-Werbung erweitert werden kann. Auch die verstärkte Detail-Betrachtung unterschiedlicher Formen der Interaktivität und des *information behavior* im Umfeld komplexer Online-Werbekampagnen erscheint als fruchtbares Feld für die Erweiterung des Modells [MCM07].

4 Erschließung

Als Vorstufe der Archivierung von Online-Werbung ist zunächst die Frage nach den Mitteln der Erschließung zu stellen. Dabei sollten Beschreibungsmodelle gewählt werden, die eine Erschließung auf verschiedenen Granularitätsstufen möglich macht (Sammlung als Ganzes, Teilbestände, Einzelne Elemente der Sammlung). Automatische Erschließungsverfahren können hier allenfalls Zusatzinformationen liefern, da die wesentlichen Metadaten zu einem Werbebanner typischerweise weder in der Multimediadatei des Banners (Flash, animierte GIF-Dateien etc.) enthalten sind, noch extrahiert werden können. Dies betrifft natürlich auch den eigentlichen „Inhalt“ der Werbung zu, die Werbebotschaft, der nur im Ausnahmefall einer vollständig textuell vorliegenden Anzeige leicht zu erschließen wäre. Dies macht klar, dass für für Online-Werbung eigene Beschreibungsformate entwickelt werden müssen. Dabei sollte ein solches Format sowohl auf generischen als auch auf anwendungsspezifischen Standards aufbauen; dazu sind zu zählen:

- Der Dublin Core-Standard für die bibliographischen Kerndaten,
- MPEG-7 als Beschreibungsformat für multimediale Daten und

- das für die elektronische Beschreibung von Museumsgut entwickelte *CIDOC Conceptual Reference Model* (CIDOC CRM, [HUN02]).

Es ist davon auszugehen, dass keiner dieser Standards bereits alle Anforderungen an die Beschreibung von Online-Werbung erfüllt, da Aspekte wie Interaktivität (*wohin verlinkt ein Banner?*) oder Metadaten zum *Targeting* (*aufgrund welcher Kriterien wurde ein Banner für welche Zielgruppe, zu welcher Zeit und in welcher Region geschaltet?*). Auch wenn offenkundig ist, dass in vielen Fällen derartige Metadaten nicht erhoben werden können, ist ein umfassendes Beschreibungsformat für Online-Werbung sowohl ein Forschungsdesiderat als auch eine wichtige Voraussetzung für den Aufbau von Online-Archiven.

5 Archive für Online-Werbung

Dedizierte Archive für Online-Werbung als grundsätzlich *digitally born data* existieren nach derzeitigem Kenntnisstand mit der Ausnahme kleiner Liebhabersammlungen ausgewählter Bannerwerbung auf verstreuten Websites bisher nicht. „Digitale Werbearchive“ beschränken sich daher auf traditionelle Werbeformen, die digitalisiert wurden oder über das digitale Medium zugänglich sind. Beispiele hier für sind

- Das Regensburger *Historische Werbefunk-Archiv*, in dem mehrere Zehntausend Werbespots aus dem Zeitraum von (ca.) 1945-1990 digitalisiert vorliegen und recherchiert werden können (vgl. *http://www.bibliothek.uni-regensburg.de/mmz/hwa_allgemein.htm*, Zugriff Mai 2009).

- Bestände zu Werbemitteln (insb. Plakate) auf Portalen, die versuchen, integrativ *cultural heritage* im digitalen Medium verfügbar zu machen wie das BAM-Portal („Bibliotheken, Archive, Museen“, *http://www.bam-portal.de*, Zugriff Mai 2009), das vom Bibliotheksservicezentrum Baden-Württemberg betrieben wird, oder auf europäischer Ebene die *Europeana*, *http://www.europeana.eu*, Zugriff Mai 2009), deren Zulieferer insbesondere Kulturgut von allgemeinem Interesse einspeisen, darunter auch Digitalisate von Werbemitteln.

Allgemein kann gesagt werden, dass Medienarchive im Vergleich mit Bibliotheken oder klassischen (dokumentfokussierten) Archiven in Deutschland eine sehr heterogene und zersplitterte Struktur aufweisen, ein zentrales nationales Medienarchiv fehlt [WOL08]. Allerdings ist davon auszugehen, dass – auch im Zuge des geplanten Aufbaus einer „Deutschen Digitalen Bibliothek“ (vgl. dazu den Medien- und Kommunikationsbericht 2008 der Bundesregierung, [BKM08]) – die Menge und inhaltliche Vielfalt digital und online verfügbaren Kulturguts dramatisch ansteigen wird, auch wenn man annehmen kann, dass wir von einer vollständigen digitalen Verfügbarkeit der Inhalte von Bibliotheken, Museen und Archiven noch weit entfernt sind.

Online-Werbung lässt sich im Kontext der Archivierungsbemühungen für das Word Wide Web nachweisen, prominentestes Beispiel hierfür ist sicher das *Internet Archive* (*http://web.archive.org/*) mit seiner *WayBackMachine* als Rechercheoberfläche. So nützlich der Nachweis auch älterer Werbeelemente im Einzelnen ist, so wenig vollständig, zuverlässig und durch Metadaten erschlossen sind diese Bestände und können den gezielten Aufbau von Archiven für Online-Werbung nicht ersetzen.

Vor diesem Hintergrund sollen nachfolgend stichpunktartig Kriterien genannt werden, die für die Entwicklung und den Aufbau eines Archivs für Online Werbung zu beachten sind. Entsprechende Arbeiten sind derzeit an der Universität Regensburg in Entwicklung. Die Liste erhebt aufgrund ihrer Vorläufigkeit keinen Anspruch auf Vollständigkeit, auch sollte deutlich werden, dass es sehr unterschiedliche denkbare Herangehensweisen z. B. bereits bei der Erfassung der Online-Werbung gibt. Aufgrund der Vielfalt möglicher Sammlungscharakteristika ist eine große Heterogenität der Sammlungsprofile vorstellbar.

Sammlungsprofil

- (global – Vollerfassung der Werbung im WWW)
- regionaler und oder sprachlicher Schwerpunkt
- Einschränkung nach Erfassungszeitraum
- Eingrenzung nach Branchen, Produktgruppen, Produkten, Marken (ökonomische Aspekte)
- Werbeformate (Banner, Videos, Interstitials, Wallpaper etc.) der Sammlung
- Sammlung nur ausgewählter (Einzel-)Formate oder Aufnahme komplexer mehr- und crossmedialer Kampagnen
- Einschränkung nach formalen und gestalterischen Aspekten (Z. B. Sprache, Musik, Form, Medienkombination, Animiertheit)
- Plattform bzw. Werbekontext: Eingrenzung nach Platzierungswebsites
- Zielgruppe: Beschränkung auf bestimmte Gruppen (Geschlecht, Alter, Bildungsgrad, Einkommen, grundsätzlich alle Targeting-relevanten Merkmale)
- Offener und dynamisch sich weiterentwickelnder Bestand oder geschlossene Sammlung

Rechtliche Perspektive (Fragen des Urheberrechts, Bestimmung der Rechteinhaber, ggf. auch an Einzelkomponenten eines Werbeformats (Bild, Musik, Text), Zulässigkeit der Speicherung für wissenschaftliche Zwecke, Zugänglichmachen für eine interessierte Öffentlichkeit, *fair* use)

Organisatorische Fragen (Trägerschaft, Kooperation mit Bibliotheken oder Archiven, Finanzierung von Aufbau und Betrieb)

Technische Aspekte der Werbemittel (Realisierungsformen (s.o.) und –technologien der Online Werbung, technischer Aufbau komplexer Banner, Dateiformate)

Erfassungsstrategien

- *Screen Recording*: „Mitschnitt“ als Liveaufnahme einer Website und der auf ihr enthaltenen Werbung
- *Web Crawling*: Speicherung von Websites als Gesamtheit oder nur der auf ihnen enthaltenen Werbemittel
- Sammlungsbereitstellung durch Dritte (Designfirmen, Werbeagenturen, Online-Vermarkter, Host-Plattformen)
- Nutzung externer vollständiger Webspeicher wie des *Internet Archive*

technische Architektur des Archivs

- Speicher- und Archivierungsstrategie (Dateisystem, Datenbank)
- Nutzung dedizierter Software (z. B. Museums-/Archivsoftware [BOR05])
- Unterstützung der Erfassung der Metadaten durch (teilautomatisierte) Tools, generell Unterstützung des Bearbeitungsworkflows
- Berücksichtigung von Standards (z. B. der *Open Archival Information System*-Standard (OAIS = ISO 14721:2003), vgl. [CSD02])
- Beachtung von Empfehlungen und Guidelines (z. B. die Empfehlungen des Kompetenznetzwerks Langzeitarchivierung (Nestor), [COY05])
- Prozesse der Bestandssicherung des Archivguts „Online-Werbemittel“, das oft in proprietären Formaten vorliegt (ggf. erforderliche Konversion, Konvertierung, Migration oder gar Emulation)

Nutzerbezogene Aspekte

- Nutzergruppen (Forschung, allgemeine Öffentlichkeit, Beschränkungsmöglichkeiten für den Zugang)
- Gestaltung der Benutzerschnittstelle des Archivs
- Retrievalfunktionalität
- Darstellungs- und Wiedergabemöglichkeiten für Online-Werbung (statische Bilder, „Abspielen“ animierter Banner, ohne / mit Originalkontext)

Die Vielzahl der oben eingeführten Kriterien sollte deutlich gemacht haben, dass der Aufbau von Archiven für Online-Werbung eine ebenso komplexe wie fruchtbare Aufgabe darstellt.

6 Ausblick

Weitgehend offen ist derzeit noch, wie die Forschungsfragen aussehen, die an ein solches Archiv herangetragen werden können. Das Beispiel des Regensburger Verbundes für Werbeforschung zeigt aber, dass hier vielfältige Perspektiven denkbar sind – für zukünftige Nutzers eines Online-Werbearchivs sind Fragen zu Designentwicklung, sprachlichem, gesellschaftlichen und kulturellen Wandel ebenso denkbar wie das

Nachzeichnen der technischen Entwicklung digitaler Medien oder der Wandel von Interaktionsstrategien. Praktische Fragen nach Trends im Farb- und Formgebrauch („Finden sich die abgerundeten Rechtecke der *social software*-Ära auch in der Online-Werbung wieder?“) sollten ebenso beantwortet werden können wie die Untersuchung der Online-Werbesprache und ihr Zielgruppenbezug.

Hinsichtlich der Erschließung stellt gerade das Finden einer guten Balance aus intellektueller Erschließung und automatischen Analyseverfahren eine interessante Herausforderung für Softwareentwicklung in der Medieninformatik dar.

Grundsätzlich erscheint Eile geboten: Werden nicht recht bald Archive für Online-Werbung aufgebaut, besteht die Gefahr, dass die erste Generation der Online-Werbung zu großen Teilen verloren geht, weil sie in Internetarchiven nur unvollständig enthalten ist, die Werbetriebenden (Produzenten, Gestalter, Auftraggeber, Marketingfirmen, Hostplattformen) über keine dedizierte Archivierungsstrategie verfügen. Diese Überlegung gilt allerdings sicher nicht nur für Online-Werbung, sondern auch für weite Bereiche der digitalen Kommunikationsmedien.

7 Literaturverzeichnis

[ADH69] Adorno, Theodor W. & Horkheimer, Max (1969). Dialektik der Aufklärung. Frankfurt.

[BKM08] Der Beauftragte der Bundesregierung für Kultur und Medien (Hrsg.) Medien- und Kommunikationsbericht der Bundesregierung 2008. Berlin: Bundesregierung, 2008, http://www.bundesregierung.de/Content/DE/__Anlagen/BKM/2009-01-12-medienbericht-teil1-barrierefrei,property=publicationFile.pdf

[BOR05] Borghoff, Uwe M. Vergleich bestehender Archivierungssysteme. München: Universität der Bundeswehr München, Fakultät für Informatik, Institut für Softwaretechnologie, 2005 [= nestor-Materialien, Vol. 3].

[BÜR05] Bürdek, Bernhard E. Design. Geschichte, Theorie und Praxis der Produktgestaltung. 3. Aufl. Basel / Boston / Berlin: Birkhäuser, 2005.

[COY06] Coy, Wolfgang Perspektiven der Langzeitarchivierung multimedialer Objekte. Berlin: Humboldt-Universität zu Berlin, Institut für Informatik 2006 [= nestor-Materialien, Vol. 5].

[CSD02] Consultative Committee for Space Data Systems (CCSDS) (ed.) (2002). Recommendation for Space Data System Standards. Reference Model for an Open Archival Information System (OAIS). CCSDS Document Nr. 650.0-B-1. BLUE BOOK. January 2002. CCSDS Secretariat, Washington/DC: National Aeronautics and Space Administration (NASA), Program Integration Division, online unter http://public.ccsds.org/publications/archive/650x0b1.pdf, Mai 2009.

[GRI09] Grimm, Karin. Entwurf eines Analysemodells für Online-Werbung Universität Regensburg, Institut für Information und Medien, Sprache und Kultur, Magisterarbeit im Fach Informationswissenschaft, März 2009 [wird online über den Dokumentenserver der Universität Regensburg verfügbar gemacht].

[JAN05] Janich, Nina. Werbesprache. Ein Arbeitsbuch. 4., unveränderte Aufl. Tübingen: Narr, 2005.

[LAM06] Lammenett, Erwin. Praxiswissen Online-Marketing. Wiesbaden: Gabler, 2006.

[MAT07] Mattern, Friedemann. Die Informatisierung des Alltags: Leben in smarten Umgebungen. Berlin et al.: Springer, 2007.

[MCM07] McMillan, Sally. Internet Advertising: One Face or Many? In Schumann, David W. & Thorson, Esther (Hrsg.). Internet Advertising. Theory and Research. Mahwah/NJ / London: Lawrence Erlbaum Associates, 2007, S. 15-35.

[OVK09] Online-Vermarkterkreis (OVK). OVK Online-Report 2009/01. Zahlen und Trends im Überblick. Düsseldorf: Online-Vermarkterkreis (OVK). im Bundesverband Digitale Wirtschaft e.V., http://www.ovk.de, Mai 2009.

[REI09] Reimann, Sandra. MEHRmedialität in der werblichen Kommunikation. Synchrone und diachrone Untersuchungen von Werbestrategien. Tübingen: Narr, 2008.

[SAH05] Sachs-Hombach, Klaus (Hrsg.). Bildwissenschaft. Disziplinen, Themen, Methoden. Frankfurt/Main: Suhrkamp, 2005.

[SAH06] Sachs-Hombach, Klaus. Das Bild als kommunikatives Medium. Elemente einer allgemeinen Bildwissenschaft. Köln: Herbert von Halem Verlag, 2006.

[WOL08] Wolff, Christian. Medienarchiv für wissenschaftlichen Film an der TIB und die Medienlandschaft in Deutschland. Expertise im Auftrag der Technischen Informationsbibliothek Hannover. Hannover: Technische Informationsbibliothek (TIB), 2008.

[WOL09] Wolff, Christian. Adwords, Skyscraper und Eselsohren. Erscheinungsformen der Online-Werbung. In Reimann, Sandra & Sauerland, Martin (Hrsg.). Wissenschaft und Werbung. Regensburg: Schnell & Steiner [im Druck].

Beyond Basic Blanks – Vertrauenserhaltende, schrittweise Implementierung neuer Funktionen im Information Retrieval

Arne Berger, Jens Kürsten

Technische Universität Chemnitz

Fakultät für Informatik

Professur Medieninformatik

{arne.berger, jens.kuersten}@informatik.tu-chemnitz.de

Zusammenfassung: Vorgestellt werden zum einen eine Erweiterung der Programmbörse, um plausibel für Anwender schrittweise Funktionalitäten hinzuzufügen. Zum anderen ein Annotations- und Retrievalinterface als zentrales Element des Anwendungskomplexes im Sinne des User-Centered-Designs.

Schlagwörter: Human Computer Interaction, User Centered Design

1 Stand der Dinge – Lokalfernsehsender und Kooperation

Basis für die kooperative Arbeit im Verbund der Lokalfernsehsender in Sachsen bildet die Programmbörse. Die Programmbörse ist eine webbasierte Datenbank, die der Multiplikation und Zweitverwertung eingestellter Beiträge über die Ausstrahlungsgrenzen des jeweiligen Senders hinaus dient. Diese Beiträge repräsentieren nur einen kleinen Teil aus dem wochenaktuellen Programm der Sender. Beiträge lassen sich nur annotiert einstellen und damit der Zweitvermarktung zuführen. Dies ist eine hinreichende Motivation für die Nutzer, die Beiträge zu annotieren.

Da alle Fernsehsender, vor allem aus medienrechtlichen Erfordernissen heraus, ein Sendearchiv führen, existiert in jeder Redaktion ein zweites, meist bandbasiertes Archiv. Dieses ist intellektuell unzureichend annotiert, bietet aber ebenso Potential zur Zweitvermarktung; beispielsweise in Form von Jahreszusammenschnitten für Werbekunden. Vorsichtige Schätzungen, basierend auf einer Bedarfsanalyse in sieben Sendern, lassen hier im Senderverbund der ARiS auf einen jährlichen Gesamtumsatz von 20.000 Euro schließen. Diese Einnahmequelle liegt brach, da dieses Archiv nicht digitalisiert ist und intellektuell mangelhaft annotiert ist. Weitere Möglichkeiten, dieses Archiv gewinnbringend zu nutzen, schließen eine automatische Aktualisierung der Website oder eine automatisierte Kopfstellenvernetzung ein.

Ziel unserer Arbeit ist es auch, Konzepte für Softwareerweiterungen zu schaffen, die es den Sendern erlauben, ohne wesentlichen Mehraufwand all ihre Beiträge umfassender annotiert in einer digitalen Datenbank abzulegen. Es ist utopisch, anzunehmen, dass mit Einführung eines digitalen Archivs in den Fernsehsendern sofort alle Anforderungen an

eine hinreichende intellektuelle Annotation erfüllt werden. Realistisch ist vielmehr eine stufenweise Einführung der technischen Voraussetzungen, wie etwa der Ausbau der Annotationsmöglichkeiten der bestehenden Programmbörse, ein Ausbau der Infrastruktur bis hin zum Senderarchiv mit Möglichkeiten der automatischen Annotation.

Dieser Prozess wird von uns, auf Seiten der Anwender in den Fernsehsendern, als auch auf Seiten der Entwickler im Projekt Sachsmedia, unterstützt. Auf Basis von Nutzeranforderungen und technologischer Machbarkeit wird das Interfacedesign entwickelt. Es werden aber auch Empfehlungen zur Nutzerakzeptanz und -relevanz einzelner Produkteigenschaften bis hin zu ganzen Softwarenutzungskonzepten erarbeitet.

2 Inhaltliche Ausrichtung

In der ersten Phase der Anforderungsanalyse in sieben repräsentativen Fernsehsendern der ARiS wurden Wünsche zur generellen Beschaffenheit eines zukünftigen, digitalen Archivs gesammelt und das Nutzerverhalten analysiert. Zentrale Anforderung ist die gewünschte Verbesserung der Metadatenqualität sowie der Wunsch, den Austausch von Beiträgen via Programmbörse zu erleichtern und zu erweitern. Im Folgenden werden zwei Teilgebiete unserer Forschungs- und Entwicklungsarbeit vorgestellt. Dies ist zum einen ein Konzept zur Erweiterung der Programmbörse in einer Art, um plausibel für Anwender schrittweise Funktionalitäten hinzuzufügen. Zum anderen ein Annotations- und Retrievalinterface, welches das zentrale Element des Anwendungskomplexes darstellt und im Sinne des User-Centered-Design entworfen wurde.

3 Weiterentwicklung der Programmbörse

Im Folgenden sollen denkbare Module für die Weiter-/Neuentwicklung von Services für das Konzept der Programmbörse dargestellt werden. Sinnvoll ist dabei ein genereller Aufbau der Module, beispielsweise Grundfunktionalität und Erweiterungen.

3.1 Automatische Preisanpassung

Die Programmbörse dient in erster Linie dazu, fertig produzierte Beiträge zu günstigen Konditionen von anderen Lokalfernsehsendern zweitverwerten zu lassen. Es liegt in der Natur von Lokalfernsehsendern, dass auf Grund ihrer territorialen Verteilung keine Konkurrenzsituation entsteht. Vielmehr führt der bidirektionale Austausch zur Aufwertung des Programms aller beteiligten Sender. Aktuell sieht die Programmbörse nur Fixpreise vor. Es erscheint plausibel, den Preis von Beiträgen automatisiert von einer Reihe an Faktoren abhängig zu machen. So können Beitragspreise abhängig von

der Anzahl der vom Käufer bereitgestellten Beiträge variieren. Somit können Teilnehmer, die das System der Programmbörse im Sinne aller unterstützen, belohnt werden.

Lösungsszenario: Die Sender legen generelle Minutenpreise für verschiedene Klassen von Beiträgen fest und geben damit Richtpreise vor. Die Sender selbst können diese Preise in festgelegtem Grad nach oben und unten variieren, um der eigenen Einschätzung des Wertes eines Beitrags gerecht zu werden. Der tatsächliche Preis des Beitrags wird jedoch für alle Teilnehmer der Programmbörse dynamisch berechnet. Dabei wird sowohl die allgemeine Produktions-/Verbraucher-Quote des Käufers einbezogen als auch die Statistik der Produktions-/Verbraucher-Quote zwischen dem aktuellen Verkäufer und Käufer. Damit wird die Verbraucherproblematik gelöst, da die Verbraucher weiterhin aktiv sein können, aber reale Marktpreise für Beiträge zahlen müssen, was den Produzenten der Beiträge zu Gute kommt. Zusätzlich wird die Kooperation zwischen den Sendern allgemein sowie speziell zwischen Sendern, die bereits bestehende Beziehungen nutzen, gefördert. Mit der Einführung von festgelegten Minutenpreisen wird einerseits die Kaltstartproblematik umgangen und andererseits ermöglicht sie die flexible Entwicklung von realistischen Marktwerten von Beiträgen.

3.2 Auto-Website-Speisung

In der Programmbörse werden die vom jeweiligen Sender – als bestem Kenner des eigenen Sendegebietes – am relevantesten bewerteten Beiträge eingestellt. Entsprechend sind dies auch die Beiträge, die zur Veröffentlichung auf der sendereigenen Website taugen. Hier bietet sich eine automatisierte Lösung an, die Website automatisch, z. B. durch Video2RSS, mit den aktuellsten Beiträgen zu speisen.

3.3 Auto Saler

Wie schon angedeutet, ist es für viele Sender ein finanziell lohnender Geschäftszweig, für den jeweiligen Endkunde personalisierte Beitragszusammenschnitte auf DVD gesammelt zu verkaufen. Aktuell hemmt der immense Personalaufwand, der für jede einzelne DVD anfällt, den Nutzen enorm. Die Programmbörse erlaubt schon in ihrer jetzigen Form das Zusammenstellen eines Beitragspaketes zum Download. Würde man archivierte Beiträge – möglicherweise nach einer vom Sender bestimmten Sperrfrist – für Endkunden durchsuchbar und zum Kauf per Download verfügbar machen, würde dies den Arbeitsaufwand auf den Endanwender übertragen.

Es ist denkbar, diese Option zum einen nur auf den jeweiligen Sender beschränkt zu implementieren, um eine senderübergreifende Kalkulation und Abrechnung der Verwertungsrechte zu vermeiden. Die alternative Variante, bei der senderübergreifend die kooperative Komponente der Programmbörse bewusst gefördert wird, um

programmbörsenweit Käufe zu ermöglichen, benötigt ein System zum Verwalten dezentral festgelegter Verwertungsklauseln im Sinne der jeweiligen Rechteinhaber.

3.4 Auto-Konvertierung

Idealerweise würde die Programmbörse alle gängigen Videoformate zum Upload akzeptieren und die Beiträge im jeweils gewünschten Videoformat ausliefern. Um Wartezeiten beim Download von häufig genutzten „Standardmaterial“ zu umgehen, sollte Material für den Upload zur Website (siehe 2. Auto-Website-Speisung) direkt nach dem Upload automatisch, und nicht erst bei Abruf, in ein gebräuchliches Format konvertiert werden.

3.5 Kopfstellenvernetzung, Bildschirmtafeln, Videotext2DVB-H

Die Programmbörse böte nach Implementierung obiger Dienste und der antizipierten Akzeptanzsteigerung eine plausible Grundlage, alle Beiträge eines Senders zu archivieren. Damit würde die Programmbörse auch als Datengrundlage für klassische Broadcastinganwendungen dienen.

4 Anwender in den Fernsehsendern

In unseren umfangreichen Site Visits wurden die Arbeitsabläufe der Beitragsproduktion in Lokalfernsehsendern analysiert. Diese ähneln sich sehr stark. In die Produktion eines einzelnen Fernsehbeitrages sind übereinstimmend viele Anwender unterschiedlicher Kompetenzfelder (Redakteure, Kontakter, Cutter, Sprecher, Kamera) involviert, die naturgemäß signifikant unterschiedliches Wissen über den jeweiligen Beitrag haben. Um dieses Wissen in Metadaten umzusetzen ist es der naheliegendste Schluss, einen verantwortlichen Mitarbeiter zu bestimmen, der alle relevanten Metadaten von allen produktionsinvolvierten Kollegen sammelt, was auch in Teilen praktiziert wird. Im Vergleich zu größeren Fernsehsendern mit dediziertem Archivpersonal stehen aber weder ausreichend Zeit noch genügend Arbeitskraft zur Verfügung, weshalb Beiträge nur nach Kenntnis des jeweils verantwortlichen Mitarbeiters annotiert werden. In keinem der besuchten Sender führte diese Arbeit zu einem konsistenten Sammelwerk aller plausiblen Metadaten.

Interessanterweise notieren sich die involvierten Kollegen zusätzlich in ihren eigenen Dokumenten produktionsrelevante Details nach ihrem professionsimmanenten und persönlichen Interesse. Dies ist wiederum allen Mitarbeitern bekannt und wird im Fall einer Suche als mentales, verbal vernetztes Metadatengewerk genutzt. Das folgende Beispiel dieser Wissensverteilung zwischen unterschiedlichen Mitarbeitern ist aus dem Praxisbetrieb übernommen:

Ein Redakteur sucht einen bestimmten Beitrag. Er verbalisiert seine Assoziationen laut im Büro, welches er mit drei Kollegen teilt. Eine Kollege erinnerte sich, dass der Beitrag eine Auftragsarbeit war, was den anwesenden Kundenbetreuer veranlasst, sein Rechnungsprogramm zu durchsuchen, um so das Produktionsdatum einzugrenzen. Da es nur zwei Aufträge vom betreffenden Kunden gab, musste der Redakteur nur an zwei Stellen im zeitlich geordneten Bandarchiv suchen, und dort mit Hilfe der Etiketten auf den Kassetten die Position des Beitrags auf dem Band ermitteln.

Ein Archivsystem, das als Anfragesprache die alltägliche Kommunikation verstünde, wäre ideal, wie an anderer Stelle in diesem Workshopband thematisiert. Bis dieses Unterfangen nicht mehr utopisch ist, soll uns ein adaptives Annotations- und Retrievalinterface helfen.

5 Annotation

Aufbauend auf dem Konzept der Programmbörse mit ihrem adaptiven Rahmen, soll jeder Anwender im Lokalfernsehsender in die Lage versetzt werden, die für ihn relevanten Metadaten zum jeweiligen Beitrag hinzuzufügen. Aus informationswissenschaftlicher Sicht und aus Gründen der Interoperabilität ist es ideal, die Archive aller etwa 70 Lokalfernsehsender mit einem grundlegenden Metadatenschema zur intellektuellen Annotation auszustatten. Wir haben uns hier für ein Subset des „Regelwerk Mediendokumentation“ [RMD 09] der ARD entschieden. Basierend auf diesen Vorgaben und unseren Nutzerstudien, erscheint ein Annotationsworkflow ähnlich dem folgenden plausibel.

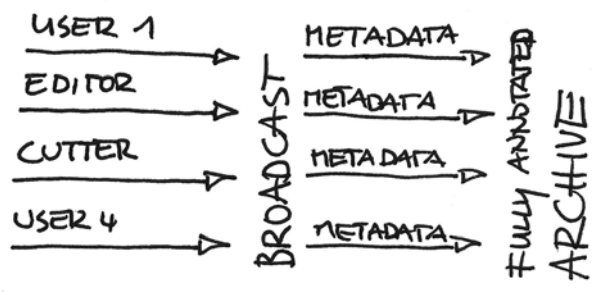


Abbildung 1: Annotationsworkflow

Für unseren ersten Annotations- und Retrievalprototypen haben wir die drei Hauptnutzergruppen der Senderarchive ein Cardsorting der Metadaten des Regelwerks Mediendokumentation durchführen lassen, um a) die Relevanz der Metadaten für das Informationsbedürfnis der jeweiligen Anwender und b) eine inhaltliche Zugehörigkeit der jeweiligen Metadaten in Sets zu ermitteln. Die signifikanten Muster im Informationsbedürfnis und der als relevant ausgewählten Metadaten überrascht nicht

und zeigt deutliche Verbindungen zwischen der Arbeitsaufgabe der Anwender im Produktionsprozess und den ausgewählten Metadaten.

Zusätzlich durchgeführte, strukturierte Interviews haben zudem ergeben, dass Anwender gewillt sind, mehr Metadaten anzugeben, wenn das entsprechende Interface nur die Metadaten eingeben lässt, die für die Aufgaben der jeweiligen Mitarbeiter relevant sind. Zusätzlich haben Anwender weitere Metadaten gewünscht, die über die des Regelwerks Mediendokumentation hinausgehen. Kontakter zum Beispiel wünschten sich die vollständigen Adressinformationen der Produktionsbeteiligten direkt am Beitrag, da sie sich Namen besonders gut merken können. Cutter währenddessen interessieren sich kaum für Sendetermine und wünschen diese folglich nicht zu annotieren. Diesem Schema folgend ermöglicht das Annotationsinterface den Sendern auch, zusätzliche Metadatensets aufzunehmen, sollten Mitarbeiter diese für ihre Arbeit benötigen.

6 Retrieval

Der gerade vorgestellte mental assoziative Suchprozess führt zu einer zweiten Erkenntnis: Nutzer sind ebenso an einem an ihre Informationsbedürfnisse angepassten, flexiblen Suchinterface interessiert. Dieses Bedürfnis liegt in der Regel zwischen einem „basic blank“-Zugang auf der einen Seite und einer Expertensuche mit Suchfeldern für alle ungefähr 70 Metadaten. Mit Papierprototypen haben wir ein Suchinterface getestet, welches ähnlich dem Annotationsinterface strukturiert wurde. Anwender haben zum einen sofort Parallelen wahrgenommen und konnten das Suchinterface entsprechend bedienen. Aktuell wird das adaptive Interface der „individuellen Suche“ gegen eine „Expertensuche“ mit einem Suchgitter als Interface evaluiert.

Abbildung 2: Custom Text Widgets

Basierend auf der Erkenntnis, dass einfache Interface-Widgets, die nur genau eine Suchaufgabe erfüllen, für eine schrittweise Verfeinerung der Suchergebnisse nützlich sind, haben wir diesen Ansatz auch auf multimediale Suchanfragen ausgedehnt, was aus zweierlei Gründen sinnvoll ist. Zum einen benötigen Anwender ständig bessere Suchergebnisse, da intellektuell annotierte Metadaten allein nicht ausreichend sind, Zum anderen ist das Interesse, weniger restriktiv suchen zu können, groß. Die Anwender sind sich der Nachteile automatisch generierter Metadaten durchaus

bewusst, würden aber die Erweiterung der adaptiven textbasierten Suche um multimediale Such-Widgets in jedem Falle vorziehen, was nicht zwangsläufig zu unpräziseren Suchergebnissen führt. Denn für Known-Item-Suchen stehen weiterhin textuelle Such-Widgets zur Verfügung und eine Suche auf rein automatisch annotierten, multimedialen Metadaten kann immer durch textuelle Suchterme präzisiert werden.

Diese hohe Nutzerakzeptanz versetzt uns in die Lage, die Archive der Sender schrittweise um automatisch generierte Metadaten und das Suchinterface um passende grafische, multimediale Such-Widgets zu erweitern, ohne das Vertrauen in die Qualität der Suchergebnisse zu trüben. Die im Projekt Sachsmedia entwickelten, automatischen Verfahren zur Personen-, Text- und Spracherkennung können so, je nach Verfügbarkeit der jeweiligen Metadaten, zusammen mit einem korrespondierenden Such-Widget implementiert werden.

6.1 Parallele Entwicklung in der IR-Community

In der Information-Retrieval-Community werden zahlreiche Konzepte entwickelt, High-Level-Metadaten für die Beantwortung einzelner Suchanfragen heranzuziehen. Ob skizzenbasierte Bildsuchmaschinen http://labs.systemone.at/retrievr oder Werkzeuge, die Bilder nach Farbe durchsuchen lassen http://www.xcavator.net – zur eigenständigen Anwendung taugen sie alle nicht. Derartige Entwicklungen wären aber per Widget gut in einen Suchablauf integrierbar. Wir haben dies in Interviews mit Information-Retrieval-Entwicklern geprüft: Alle befragten Experten fokussieren ihre Arbeit auf genau ein Retrievalkonzept; jede dieser Arbeiten wäre in diesem Widget-Konzept als einzelnes Werkzeug zum Formulieren eines Suchterms geeignet.

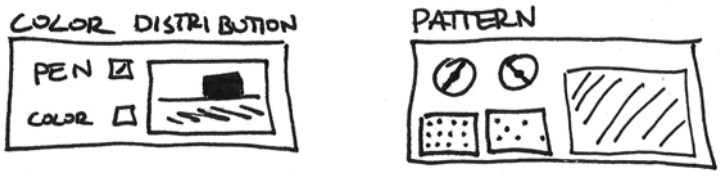


Abbildung 3: grafische Widgets

7 String of Reformulation > Beyond Basic Blanks

Mit den hier vorgestellten Nutzerstudien und des geschilderten Dilemmas der „basic blanks“ als Grundlage, haben wir ein Interface entwickelt, das basierend auf Nutzervorlieben und -wünschen sowie der Abhängigkeit vom Implementierungsgrad von Funktionen, Such-Widgets zur kombinierten Anfrage zur Verfügung stellt.

Wenn Anwender einen bekannten Beitrag suchen, ist eine textbasierte Suche ausreichend und kann mit Hilfe entsprechender Metadatenfelder durchgeführt werden. Die meisten Anfragen jedoch sind derart vage, dass Nutzer versuchen, irgendetwas zu finden, das ihrem mentalen Bild entspricht, weshalb derartige Anfragen mehrmals und substantiell reformuliert werden. Normalerweise beginnen Anwender mit einer kurzen Anfrage und verfeinern nach Durchsicht der Suchergebnisse die Suchanfrage solange, bis eine passende Treffermenge sichtbar wird. [BAT 89] Während dieses Prozesses formen die Anwender eine Reformulierungsfolge in ihren Köpfen. Das wiederholte Editieren textbasierter Anfragen führt dabei zu modalen Brüchen, die den mentalen Reformulierungsprozess behindern, da vorherige Versionen textbasierter Suchanfragen nach dem Editieren nicht mehr zur Verfügung stehen. Mit unserem Suchinterface ist es möglich, diesen Reformulierungsablauf sichtbar und nachvollziehbar zu machen. Es erlaubt Anwendern initiale Suchanfragen iterativ mit weiteren Suchtermen in Form von Widgets zu verfeinern, um die Menge an Suchergebnissen dadurch einzuschränken. Der Suchablauf wird dabei visualisiert und bleibt editierbar.

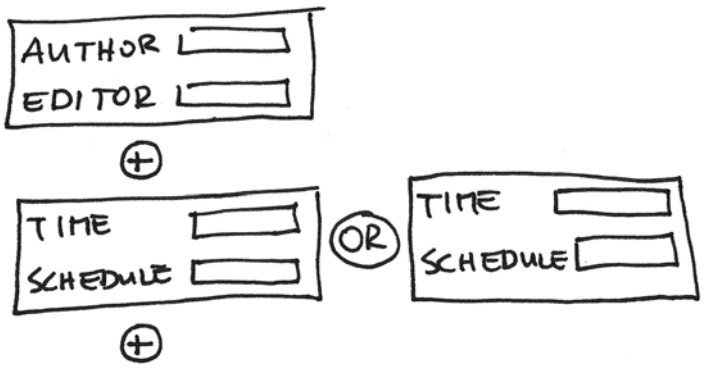


Abbildung 4: Flow Of Reformulation

Auf den Anforderungen der Anwender basierend haben wir Basis-Sets mit textbasierten Widgets erstellt. Zusätzlich sind Interface-Guidelines und ein passendes XML-Schema entwickelt worden, um textbasierte Widgets einfach zu erstellen. Ferner sind grafische Widgets entstanden, die den aktuellen Forschungsstand der automatischen Annotation im Projekt Sachsmedia widerspiegeln.

Anwender können so das am besten passende Such-Widget auswählen, um eine schrittweise formulierte Suchanfrage zu stellen. Auf Grundlage dieser Untersuchungen, den Anwenderanforderungen und unserem User-Centered-Customization Ansatz erweitern wir die klassische Filter/Flow-Metapher [SHN 93] auf offene Datenbanken mit multimedialen Inhalten, automatisch generierte High-Level-Metadaten und vage Anfragen. Shneidermans Metapher, von Wasser, das durch Filter läuft, als visuelle Repräsentanz der Formulierung boolescher Anfragen, schnitt im Test bei Nutzern, die mit boolescher Logik nicht vertraut sind, signifikant besser ab, als die Benutzung von SQL-Syntax [SHN 93]. Datenbanken wachsen seitdem unaufhaltsam und für einen

Großteil der Anwender ist boolesche Logik immer noch ein Praxisrätsel. Das Ziel, Reformulierung zu visualisieren, verfolgt auch Jones mit Venn-Diagrammen [JON 99]. Auch Sentinel [KNE 97] oder InfoCrystal [SPO 93] scheitern als zu schwierig für einen Großteil der Anwender. Interfaces von „Expertensuchen“ fokussieren daher aktuell auf einfachsten Metaphern und schränken die Möglichkeiten, Suchen nutzerfreundlich zu reformulieren, wieder erheblich ein. Der Versatz zwischen diesen beiden Konzepten besteht somit weiter: Nutzer wollen nicht darüber nachdenken, wie sie mit einem System interagieren sollen, wollen sich aber auch nicht auf einen festgelegten Ablauf einlassen, weil das ein Gefühl der eingeschränkten Möglichkeiten vermittelt [WBW 03]. Aus diesem Grund versetzen wir die Nutzer in die Lage, den Suchablauf so präzise oder vage zu formulieren, wie sie für die aktuelle Aufgabe angemessen finden und lassen es damit in der Hand der Nutzer, einfache oder komplexe Anfragen innerhalb des gleichen Begriffsmodells zu stellen. Ausführliches dazu findet sich in „Visual String of Reformulation“ [BER i.E.].

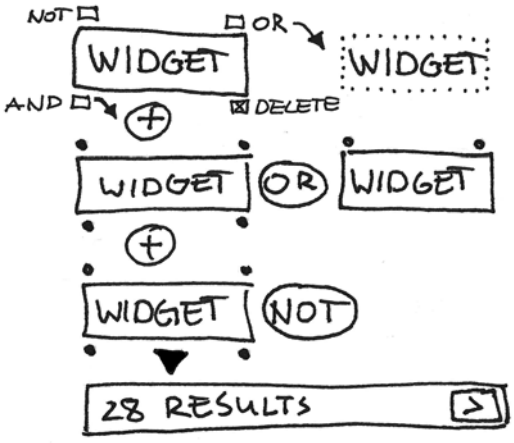


Abbildung 5: Boolesche Verknüpfungen

8 Literaturverzeichnis

[BAT 89] Bates, M. J. The Design of Browsing and Berrypicking Techniques for the online search interface. Graduate School of Library and Information Science. University of California at Los Angeles. 1989

[BER i.E.] Berger, A. Visual String of Reformulation. HCI International. San Diego, CA., USA. 2009 (im Erscheinen)

[JON 99] Jones, S., McInnes, S. A graphical userinterface for Boolean query specification. IntJDigitLibr(1999). pp. 207–223 International Journal on Digital Libraries. 1999

[KNE 97] Knepper, M.M., Killiam, R., Fox, K.L.: Information Retrieval and Visualization using SENTINEL. In: Proceedings of the Trec-7 Conference, pp. 393-397. 1997

[RMD 09] Regelwerk Mediendokumentation. http://rmd.dra.de/arc/php/main.php (zuletzt besucht: 1. Mai 2009)

[SHN 93] Shneiderman, B., Young, D.: A Graphical Filter/Flow Representation of Boolean Queries. Journal of the American Society for Information Science. Vol. 44. pp. 327-339, 1993

[SPO 93] Spoerri, A.: InfoCrystal: a visual tool for information retrieval & management. In: Proceedings of the second international conference on Information and knowledge management, pp. 11-20, Washington, D.C. 1993

[WBW 03] Bauer-Wabnegg, W., Krause, J. Visualisierung und Design - Grundlagen von Softwareergonomie und Mediendesign. Universität Koblenz-Landau. 2003

http://www.cuil.com (zuletzt besucht: 1. Mai 2009)

http://www.sowiport.de (zuletzt besucht: 1. Mai 2009)

http://www.quintura.com (zuletzt besucht: 1. Mai 2009)

http://labs.systemone.at/retrievr (zuletzt besucht: 1. Mai 2009)

http://xcavator.net (zuletzt besucht: 1. Mai 2009)

Beyond Basic Blanks – Akzeptanz adaptiver Annotations- und Rechercheoberflächen

Arne Berger

Technische Universität Chemnitz

Fakultät für Informatik

Professur Medieninformatik

{arne.berger}@informatik.tu-chemnitz.de

Zusammenfassung: User-Centered-Design-Überlegungen zu einer Annäherung von Anwendern und Entwicklern auf beiden Seiten einzeiliger, textbasierter Information Retrievalinterfaces.

Schlagwörter: Human Computer Interaction, User Centered Design

1 Einleitung

„Words intended to represent concepts: that is the questionable foundation, upon which information retrieval is built.“ (MOR 06) Fragwürdig, weil es Worte sind, die Konzepte repräsentieren, Worte als Suchterme, Worte als Metadaten. Worte sind ihrer Natur nach unpräzise, mehrdeutig, undefiniert, gegensätzlich, usw.

2 User Centered Design

Sachsmedia unterstützt die Lokalfernsehsender in ihrem Bestreben, mehr relevante Metadaten zu produzieren, mit automatischen Verfahren des Video Retrieval. Diese bilderkennenden und bildinterpretierenden Verfahren sind in der Implementierung vergleichsweise aufwändig und bilden daher einen größeren Komplex innerhalb des Projektes Sachsmedia. Im Rahmen dessen ist ein Konzept für ein nutzerfreundliches Annotations- und Retrievalinterface entwickelt worden. Durch ein offenes, erweiterbares und nutzeranpassbares Interfacedesign soll versucht werden, auf Seiten der Anwender die Bereitschaft zum intellektuellen Annotieren der Beiträge zu erhöhen. Des Weiteren möchte ich die Akzeptanz unscharfer, automatisch generierter Metadaten erhöhen sowie einen Weg finden, deren schrittweise Implementierung den Anwendern plausibel zu machen. Den Prozess der kontinuierlichen Verbesserung der Metadatenqualität unterstütze ich im Sinne eines ganzheitlichen User Centered Designs auf zwei Seiten.

Interfacedesign steht klassischerweise in der Mitte zwischen Anwendern und Entwicklern. Es ist erklärte Aufgabe, als gemeinsamer Nenner zu dienen, der es Anwendern erlaubt, technisch komplexe Abläufe kompetent zu meistern [SLO 07] und

Entwickler in die Lage versetzt, abstrakte Algorithmik plausibel zu machen. Dies führt zu einem Paradoxon des Information Retrieval, wonach sich die Vorstellung dieser beiden Enden des Interfaces auf höchst ähnliche Weise manifestiert. Ich möchte hier beginnen, aufzeigen, warum das so ist und beide Seiten des Designprozesses betrachten, um ein theoretisches Fundament für eine Lösung zu entwickeln. Dieses Vorhaben orientiert sich am Konzept des User Centered Design. Mittel der Wahl ist die klassische Nutzerpartizipation, jedoch auf beiden Seiten.

3 Idealform des IR-Interfaces aus Sicht der Anwender

Idealerweise würde das Interface eines Archivsystems die verbalisierte, assoziative, natürlichsprachliche Recherche erlauben. Zum einen unterstützte dies die Präferenz der alltäglichen Kommunikation, zum anderen gäbe es keinen modalen Bruch durch Umdeuten der komplexen Wortwelt im Kopf in abstrahierte, boolesch verknüpfte Wortgruppen. Dies führt grundsätzlich zu einem steigenden Missverhältnis zwischen der formulierten Anfrage des Nutzers und deren Repräsentanz im Retrievalsystem. In den meisten Fällen führt dies zu Unsicherheit der Anwender und damit zu Zurückhaltung im Formulieren von Suchanfragen. [RUS 07, WEL 07]

Benutzer haben in den allermeisten Fällen ein diffuses Informationsbedürfnis, das nur vage zu formulieren ist. Nicht umsonst sind Bibliothekar oder Archivar eigenständige Berufe und Legionen an spezialisierten Informatikern beschäftigen sich mit der Optimierung von Information-Retrieval-Systemen. Denn den Suchmaschinen fehlt viel Wissen über den semantisch präzisen Inhalt der Dokumente oder sie können für einen Menschen inhärent logische Aufgaben der Kontextualisierung wie die Unterscheidung homographer und synonymer Worte, nicht lösen. Was Nutzer über die inneren Abläufe eines Retrievalsystems denken, bleibt den Entwicklern genauso verborgen wie dem Nutzer die Kenntnis über die algorithmischen Abläufe. Inwieweit Techniken wie Normalformenreduktion (Stemming), Thesauri und Wörterbücher, Entfernen von Stoppworten oder unscharfes Retrieval eingesetzt werden, um die Suchanfrage zu optimieren, bleibt dem Anwender in der Regel verborgen. Nutzer wissen auch nicht, ob das, was sie suchen, indiziert wurde, ob dieser Index vollständig ist oder ob dort ihre Suchintention als genauso relevant abgebildet wurde wie in ihrem Kopf. [RUS 07] Die Selbsteinschätzung, eine Suchmaschine „gut“ zu beherrschen, ist zudem in den allermeisten Fällen naiv. 93 % aller Teilnehmer einer Studie von Google gaben an, sehr gute Kenntnisse im Umgang mit Google zu haben, während 66 % aller Teilnehmer weniger als einmal täglich die Suchmaschine nutzen. [RUS 07]

Diese gefühlte Kompetenz wird durch den algorithmisch-intelligenten Umgang mit Worten, den Suchmaschinen wie Google pflegen, nur gestärkt. Die Relevanz von Inhalten für die Anwender kann nicht allein mit der Suche nach nur zwei Worten [SIL 99, JAN 00] im Inhalt bestimmt werden. So wird nicht nur die Suche im Inhalt, sondern

auch die Popularität von Inhalten durch Popularitätsalgorithmen [PAG 99] zur Relevanzbewertung hinzugezogen. Diese helfen dann auch, ein kontrolliertes Wörterbuch populärer Metadaten zu erstellen, das einen semantischen Bezug herzustellen versucht. Googles Siegeszug als beliebteste Suchmaschine im Web lässt sich mit einem offensichtlich von der Nutzerpopulation gemeinhin akzeptierten, weil populären Umgang mit den Worten, erklären. Google beantwortet eine Suche nach „Sachsmedia“ mit einer Handvoll Ergebnissen, die allen Nutzern regelmäßig ausreicht, denn diese Ergebnisse thematisieren den Suchterm ausreichend genau und enthalten ihn nicht nur einfach häufig. Die von den Anwendern gefühlte Relevanz der ersten Ergebnisse der Trefferliste ist wichtiger als Precision und Recall des ganzen Ergebnisrests. [RUS 07]

Nimmt ein Nutzer automatische Suchtermveränderungen wahr, führt ihn das entweder dazu, die Sprache des Systems zu erlernen oder bekräftigt ihn darin, das System weiterhin natürlichsprachlich zu befragen, da die Suchtermbehandlung ein offenkundiger Hinweis darauf ist, dass das System in der Lage ist, die Suchanfrage zu interpretieren. Der erste Fall kann als Hinweis gesehen werden, Interfaces für Information-Retrieval-Systeme nicht zu verbessern, weil die Anfragesprache grundsätzlich erlernbar ist. Der zweite Fall zeigt die andere Seite, der einzig mit Mitteln des Information Retrieval nicht überbrückbaren sprachlichen Kluft zwischen Anwender und System. Deshalb müssen sich Nutzer weiterhin mit dem vagen Ergebnis einer optimierten Suchanfrage zufrieden geben. Diesem Modell der Wahrnehmung der Nutzerbedürfnisse folgend werden im populären Interfacedesign „basic blanks“, also formularbasierte, einzeilige Retrievalinterfaces, als naheliegend erachtet und implementiert.

Auch in Fällen, in denen klassische Metadaten mit Tags angereichert werden oder Suchergebnisse in semantischen Clustern visualisiert werden, hat dieses Interface immer noch die Nase vorn. Auch technisch fortgeschrittene Videoretrievalsysteme wie http://xcavator.net, http://www.quintura.com, http://www.seeqpod.com oder http://www.veoh.com bilden hier keine Ausnahme.

Um spezifische Metadaten besser durchsuchbar zu machen und um algorithmisch abstrakt ausdrückbare Suchtermverknüpfungen halbwegs nutzerfreundlich formulieren zu können, werden oftmals „Expertensuchen“ angeboten. Diese, beispielsweise aus Formulargittern bestehenden Interfaces, sind eigentlich softwareergonomische Mindestanforderung. Sie machen routinemäßige Suchen für professionelle Rechercheure halbwegs ergonomisch und bieten allen anderen Anwendern überhaupt einen Weg, komplexere Suchanfragen zu formulieren, ohne die künstliche Sprache eines Retrievalsystems zu erlernen.

4 Idealform des IR-Interfaces aus Sicht der Entwickler

Ironischerweise besteht das Dilemma der „basic blanks“ als Retrievalinterface auch auf Seiten der Information-Retrieval-Entwickler. Die IR-Community hält zwar reichhaltige Entwicklungen bereit, mit denen High-Level-Metadaten bei der Beantwortung einzelner Suchaufgaben hilfreich sein können und zeigt, welche Möglichkeiten die Entwicklung der nächsten Zeit bringen wird. Beispiele dieser Entwicklung sind skizzenbasierte Bild-Suchmaschinen wie http://labs.systemone.at/retrievr oder Entwicklungen wie http://www.cuil.com oder http://www.quintura.com, die sich auf Drilldowns oder Inhaltscluster konzentrieren, oder Werkzeuge wie http://www.xcavator.net, die Bilder nach ihrer Farbverteilung bzw. nach Farbähnlichkeit durchsuchen lassen. Allein für die Anwendung im professionellen Arbeitsumfeld reicht dies nicht aus, sie sind vielmehr nur als browsing-unterstützend zu sehen.

Ein Großteil der Entwickler konzentriert sich zudem nur auf jeweils ein Gebiet des Information Retrieval, weshalb auch dort einfachste, meist textbasierte Interfaces implementiert werden. Ein weiterer Grund für diesen Versatz zwischen den multimedialen Möglichkeiten des Information Retrieval und ihren viel zu abstrakten Zugängen ist die Absenz von zeitgemäßem Interfacedesign. Hier gibt es rühmliche Ausnahmen, wie zum Beispiel http://www.sowiport.de.

Ein entscheidendes Problem ist zudem der Fokus der Entwicklercommunity auf dem Paradigma, dass Nutzer sich wenig mit dem Suchen beschäftigen wollen und mehr Zeit mit dem Browsen verbringen. Dies ist passenderweise gleich die Paradeausrede, um frustrierten Anwendern mit einzeiligen, algorithmisch geprägten Suchinterfaces zu begegnen. Die Ergebnisse der Nutzerforschung unterstützen dieses Bild zudem unfreiwillig: Nutzer verwenden pro Anfrage nur 2,35 Suchterme und verändern, je nach Quelle, die Suchanfrage nur zwei bis drei Mal. [JAN 00, SIL 99] Dies ist mühelos als Desinteresse des Anwenders, eine Anfrage präzise zu formulieren, interpretierbar. Es motiviert Entwickler dazu, nutzerunterstützende Algorithmen zu implementieren und so dem angenommenen Modell der suchfaulen Anwender zu entsprechen. Die Möglichkeit, dass die algorithmische Interpretation der Suchanfrage dem mentalen Modell der Anwender so wenig entspricht, dass diese frustriert auf komplizierte Suchanfragen verzichten, wird kaum in Betracht gezogen.

Dass die natürlichsprachliche Anfrage gerade auf Wissenschaftler und Entwickler einen großen Reiz ausübt, ist natürlich unumstritten. Wolfram et. al. hat ein Information-Retrieval-System versprochen, das natürlichsprachliche Anfragen an http://www.wolframalpha.com zufriedenstellend beantworten kann. [WOL 09a, WOL 09b] Mathematica als Grundlage der „computational knowledge engine“ legt eine – wenn auch vermutlich besonders komplex-differenzierte – mathematisch-algorithmische Behandlung mit all ihren Unzulänglichkeiten nahe.

Bis Information-Retrieval-Systeme zuverlässig in der Lage sind, Sinn und Intention der Anfrage zu interpretieren, sollten die Möglichkeiten des klassischen GUI-Designs genutzt werden, um Nutzer zu motivieren, Suchanfragen zu präzisieren. Dies kann durch Design erreicht werden, das Suchanfragen erlaubt, die besser dem mentalen Modell der Nutzer entsprechen. Ebenso bedarf es Information-Retrieval-Systemen, die ihre Fähigkeiten aktiv den Nutzern kommunizieren und diesen die Möglichkeit geben, das Interface ihren Ansprüchen anzupassen. Idealerweise sollten auch Entwickler in der Lage sein, Funktionalitäten, die auf neuen Indizes oder Metadaten basieren, einfach hinzuzufügen.

Aus diesem Grund sind Interviews im Sinne des User Centered Design auch auf Gespräche mit Retrievalexperten auszudehnen. Dies hilft, Interfaces die Kompetenzabstufungen repräsentieren zu lassen, die den zu Grunde liegenden Algorithmen entsprechen. Dies sind wichtige Mittel, um die Wortwelt im Kopf der Nutzer und die unscharf-unbekannten Suchtermoptimierungen in der Blackbox Retrievalsystem besser zu synchronisieren.

5 Literaturverzeichnis

[JON 99] Jones, S., McInnes, S. A graphical userinterface for Boolean query specification. IntJDigitLibr(1999). pp. 207–223 International Journal on Digital Libraries. 1999

[JAN 00] Jansen , B. J., Spink, A., Saracevic, T. Real life, real users, and real needs: a study and analysis of user queries on the web. In Inf. Process. Manage, 207-227.

[MOR 06] Morville, P. Ambient Findability. Sebastopol, CA, USA: O Reilly Media, Inc. 2006

[PAG 99] Page et.al. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab. 1999

[RUS 07] Russell, D.M. What are they thinking? Searching for the mind of the searcher. Invited presentation: JCDL. 2007

[SIL 99] Silverstein, C., Marais, H., Henzinger. Analysis of a very large web search engine query log. In SIGIR Forum, 33 (1), 6-12. 1999

[SLO 07] Sloterdijk, P. Der ästhetische Imperativ. Hamburg: Philo & Philo Fine Arts | EVA Europäische Verlangsanstalt. 2007

[WEL 07] Wells, J., Truran, M., Goulding, J. Search Habits of the Computer Literate. In Proceedings of the eighteenth conference on Hypertext and hypermedia. ACM, New York, NY, USA, Pages: 37 – 38, 2007

[WOL 09a] http://blog.wolfram.com/2009/03/05/wolframalpha-is-coming/ (zuletzt besucht: 8. Mai 2009)

[WOL 09b] http://www.youtube.com/watch?v=5TIOH80Qg7Q (zuletzt besucht: 8. Mai 2009)

http://xcavator.net (zuletzt besucht: 1. Mai 2009)

http://www.cuil.com (zuletzt besucht: 1. Mai 2009)

http://www.quintura.com (zuletzt besucht: 1. Mai 2009)

http://www.sowiport.de (zuletzt besucht: 1. Mai 2009)

http://labs.systemone.at/retrievr (zuletzt besucht: 1. Mai 2009)

Nutzung von Mediatheken öffentlich-rechtlicher Fernsehsender

Sven Pagel, Carina Bischoff, Sebastian Goldstein und Alexander Jürgens

Fachhochschule Düsseldorf
Fachbereich Wirtschaft
Professur für BWL, insbes. Kommunikation und Multimedia

{sven.pagel,carina.bischoff,sebastian.goldstein,alexander.juergens}@fh-duesseldorf.de

Zusammenfassung: Angebot und Nutzung redaktioneller und werblicher Bewegtbildinhalte im Netz steigen in den letzten Jahren kontinuierlich. Mit Blick auf redaktionelle Inhalte spielen Mediatheken von Fernsehsendern eine wichtige Rolle. Mittels eines explorativen Eyetracking-Tests identifiziert dieser Beitrag erste Ansätze zu Nutzungsschemata der User von Mediatheken. Darüber hinaus werden Nutzungsprobleme aus Usability-Sicht aufgezeigt.

Schlagwörter: Mediatheken, Bewegtbildkommunikation, Usability

1 Einleitung

Sowohl das Angebot als auch die Nutzung von Bewegtbildinhalten im Web steigt seit einigen Jahren kontinuierlich. Beispiele für Angebote reichen von Videoportalen wie YouTube bis zu einzelnen Videos in Online-Shops oder Online-Zeitungen. Der Abruf von Videodateien zumindest gelegentlich ist von 25 Prozent der Onlinenutzer im Jahr 2005 auf 55 Prozent 2008 gestiegen [VA08]. Medienunternehmen bieten ihre Bewegtbildinhalte zunehmend in sog. Mediatheken an. Hierbei handelt es sich um Datenbanken mit audiovisuellen Medien. Angebotsseitig wurde der technische Entwicklungsprozess von Mediatheken in einem ersten Schritt anhand von Experteninterviews untersucht [PA08a]. Im hier vorgestellten Beitrag wird nun nutzungsseitig anhand einer Blickregistrierungsstudie mit Fragebogen die Wahrnehmung und Nutzung von öffentlich-rechtlichen Mediatheken in einer qualitativen explorativen Studie eingehend betrachtet. Die Untersuchung fügt sich ein in eine Forschungsreihe zur Bewegtbildkommunikation im Web.

nach Altersgruppen, in %

nutze das Internet …	Gesamt 2007	Gesamt 2008	14–19 J. 2007	14–19 J. 2008	20–29 J. 2007	20–29 J. 2008	30–49 J. 2007	30–49 J. 2008	ab 50 J. 2007	ab 50 J. 2008
überwiegend zur Unterhaltung	14	19	47	58	17	30	8	13	6	8
überwiegend um Informationen zu erhalten	72	62	32	18	58	42	80	65	85	83
sowohl als auch	14	18	21	24	25	28	11	22	10	10

Basis: Onlinenutzer ab 14 Jahren in Deutschland (2008: n=1186, 2007: n=1142)

Quelle: ARD/ZDF-Onlinestudien 2007–2008.

Abbildung 1: Internetnutzung zur Unterhaltung bzw. zur Information 2007 und 2008 (Quelle: [VA08])

2 Mediatheken als Instrument der Bewegtbildkommunikation

Bewegtbildkommunikation kann aus technischer, grafischer und inhaltlicher Dimension betrachtet werden. Aus *technischer* Sicht lassen sich Bewegtbilder u.a. in Flash-Formate, Video-Formate wie H.264 oder Windows Media Video unterteilen. Videos können per Streaming oder Download abgerufen werden. Hinsichtlich der *grafischen* Einbindung von Videos sind insbesondere Fragestellungen der Informationsarchitektur von Relevanz. In Abhängigkeit des jeweiligen Anbieters lassen sich in einer *inhaltlichen* Betrachtung redaktionelle, werbliche und nutzergenerierte Videos differenzieren, die informativen oder unterhaltenden Inhalt haben können.

Redaktionelle Video-Clips finden sich beispielsweise auf den Websites von Fernsehsendern oder Onlinezeitungen und lassen sich als Medienkommunikation subsummieren. Werbliche Clips werden z.B. über Portale wie www.autofernsehen.de zur Verfügung gestellt, als Pre-Roll-Spot in andere Videos auf beliebigen Websites eingebunden oder durch virale Formen verbreitet (Marketingkommunikation). Nutzergenerierte Clips machen einen Teil des Angebots von YouTube und anderen Video-Hostern bzw. Video-Communities aus und sind teilweise der Medienkommunikation zuzuordnen, da hier Laien „parajournalistisch“ tätig sind [NE00].

Hinsichtlich der Einbindung in die Websites lässt sich eine weitere Unterscheidung identifizieren. So sind originäre Bewegtbilder der eigentliche Zweck einer Website wie z.B. bei YouTube, während additive Bewegtbilder multimediale Zusatzinformation zu einer Textanzeige beispielsweise in einem Stellenportal darstellen [PA08b].

Im vorliegenden Beispiel sollen Online-Mediatheken untersucht werden. Hierbei handelt es sich somit um originäre Bewegtbildinhalte aus der Medienkommunikation. Beispielhaft herausgegriffen wurden drei öffentlich-rechtliche Angebote.

Die Funktionsweise von Mediatheken wird in Abbildung 2 grafisch dargestellt und wie folgt beschrieben. Online-Mediatheken sind als „Content-Application“ [ME08] zu bezeichnen. Entweder nutzen die Mediatheken dasselbe zugrunde liegende Content Management System wie die zugehörige Website des Fernsehsenders oder es wird ein spezifisches Video Content Management System aufgebaut. Neben dieser Software-Umgebung wird das Backend durch Hardware wie Transcodierungs-Plattformen und Ausspiel-Server beschrieben. Das Frontend umfasst die oftmals Flash- und Ajax-basierte Web-Präsentation auf dem Endgerät des Nutzers. Bei der Frontend-Navigation z.B. der ZDF-Mediathek „steht das Bild im Mittelpunkt und die Navigationselemente werden weitgehend dynamisch nur dann eingeblendet, wenn der Nutzer sie auch benötigt“ [SC08]. Komplexe Anforderungen wie zeitnahe Videokonvertierung, Metadatenverwaltung und Streaming Management müssen bei der Entwicklung und dem Betrieb von Mediatheken berücksichtigt werden.

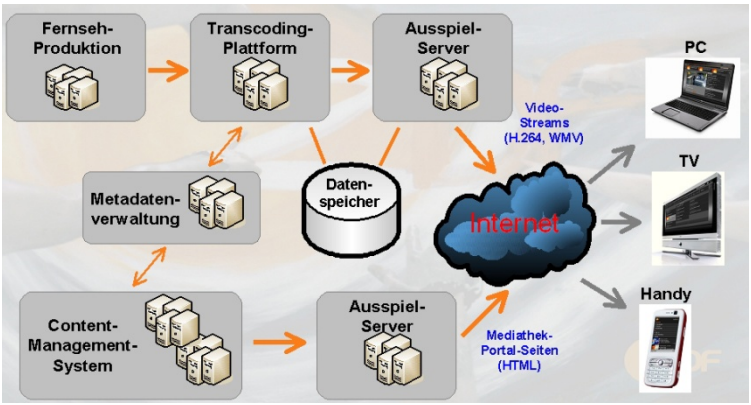


Abbildung 2: Systemverbund zur Produktion der ZDF-Mediathek (Quelle:[SC08])

Medienproduktion und Mediennutzung sind die beiden maßgeblichen Schritte der Wertschöpfungskette in den Medien, die auch für Bewegtbildinhalte gilt. Es wurde bereits erwähnt, dass in einem ersten Schritt durch Experteninterviews mit den Verantwortlichen für Mediatheken bei Fernsehsendern überprüft wurde, ob Usability-Erkenntnisse bei der Entwicklung von Mediatheken derzeit Anwendung finden (vgl. hierzu [PA08a]).

Die dabei identifizierte Verwendung von nutzergenerierten Werkzeugen der Softwareentwicklung wird in der folgenden Tabelle dargestellt. Test-driven Development und User Personas finden in der Entwicklung der Mediatheken überhaupt keine Verwendung. RTL arbeitet nur in der Bewerbung seiner Mediathek mit derartigen Stereotypen (Max, Mona, Mia). Die Differenzierung zwischen User Stories und Use Cases, die eher aus Entwicklungsparadigmen als methodischen Unterschieden resultiert, hat sich als zu feinteilig erwiesen. Immerhin die Hälfte der Sender hat mit einer textuellen Formulierung von Benutzungsszenarien gearbeitet. Die Ergebnisse zum Werkzeug ‚User im Team' werden in der Tabelle differenziert dargestellt: externe Onlineuser waren in keinem Fall eingebunden, interne Nutzer wie Redakteure durchaus. Diese Einbindung fand aber vor allem in Meetings oder kleinen Tests des Backends und nicht in Form gemeinsamer Entwicklung statt. Usability-Tests sind zwar bei mehreren Sendern geplant, durchgeführt wurden sie bisher nur bei zwei Anbietern.

	ARD	ARTE	ProSieben	RTL	ZDF
Test-driven Development	Nein	Nein	Nein	Nein	Nein
User Stories/ Use Cases	Ja	Nein	Ja	Nein	Ja
User Personas	Nein	Nein	Nein	Im Marketing	Nein
User im Team - intern	Nein	Ja	Ja	Ja	Ja
User im Team - extern	Nein	Nein	Nein	Nein	Nein
Usability Tests	Geplant	Geplant	Ja	Nein	Ja

Tabelle 1: Einsatz der nutzerbezogenen Instrumente (Quelle: [PA08a])

3 Untersuchungsdesign des Tests

Der Fokus dieser Untersuchung liegt auf einem systematischen Vergleich mehrerer Mediatheken von öffentlich-rechtlichen Fernsehsendern. Als relevante Untersuchungsobjekte wurden die folgenden Websites herangezogen:

- mediathek.zdf.de (ZDF)
- mediathek.daserste.de (ARD)
- www.wdr.de/mediathek (WDR)

3.1 Testvorbereitung

Bei der Probandenakquise wurde auf soziale Netzwerke der Studierenden einer Lehrveranstaltung im Master-Studiengang „Kommunikations-, Multimedia- und Marktmanagement“ zurückgegriffen. Für eine explorative Studie, die keinen Anspruch auf Repräsentativität stellt, erscheint diese Vorgehensweise angemessen. Anhand eines Screening-Fragebogens wurde geprüft, ob bzw. wie die Probanden die folgenden Merkmale erfüllen. Angestrebt war eine möglichst gute Gleichverteilung, die allerdings nicht bei allen Merkmalen erfüllt werden konnte. Als Merkmale wurden das Alter und Geschlecht, der Familienstand sowie die Anzahl der Kinder, der Beruf, der Schulabschluss und die Internetnutzung abgefragt.

3.2 Testdurchführung

In einem Eyetracking-Test am 17. und 18. Februar 2009 wurden die drei genannten Mediatheken 32 Probanden präsentiert. Jeder Proband musste Aufgaben in allen drei Mediatheken bearbeiten. Durch die Rotation der Aufgaben wurde sichergestellt, dass Reihenfolgeneffekte durch eine wechselnde Sortierung der Mediatheken ausgeglichen wurden. Die Testdurchführung erfolgte in drei Phasen: Pre-Test (zum Screening der Probanden), der Datenerhebung (mit 3x3 Aufgaben) und der Post-Test-Befragung (mit ca. 25 Fragen).

Die Aufgaben stellten sich (am Beispiel der Mediathek von DasErste) wie folgt dar. Für die anderen Mediatheken wurden ebenfalls passende Aufgabeninhalte entwickelt.

1. Sie haben am Sonntag die Sendung „Anne Will“ im Ersten verpasst. Suchen Sie bitte das Video zur Sendung. (Ausgangspunkt: Webseite des TV-Senders).
2. Gleich öffnet sich auf dem Bildschirm die Mediathek eines Fernsehsenders. Schauen Sie sich diese zwei Minuten lang an und bewegen Sie sich frei auf der Webseite. (Ausgangspunkt: Mediathek des TV-Senders).
3. Stellen Sie sich vor Sie haben ein Aquarium und möchten sich über die richtige Reinigung informieren. In der Sendung „ARD Buffet“ vom 12.02.2009 wurde dieses Thema behandelt. Informieren Sie sich bitte in der nachfolgenden Mediathek über die Reinigung von Aquarien. (Ausgangspunkt: Mediathek des TV-Senders).

3.3 Testauswertung

Die Auswertung wurde getrennt für Eyetracking und Befragung vorgenommen. Die Eyetracking-Daten wurden direkt mittels Tobii Studio, dem Auswertungstool des verwendeten Eyetracking-Systems Tobii T60, bzw. mit Hilfe von Excel analysiert. Die Befragung wurde mit dem Open Source Tool Limesurvey durchgeführt und ebenfalls in Excel ausgewertet.

4 Nutzung von Mediatheken

4.1 Optimale Nutzungsprozesse auf Mediatheken

Es wurden für die untersuchten Mediatheken jeweils optimale Navigationspfade entsprechend der möglichen Navigationsstrategien identifiziert. Als allgemeine Strategieansätze können die Nutzung der seiteninternen Suche oder die Verwendung der globalen Navigation angesehen werden. Die Festlegung der Navigationspfade erfolgt dabei entsprechend den Aufgabenstellungen für zwei Ausgangslagen: die Startseite der öffentlich-rechtlichen TV-Sender (Aufgabe 1) und die Startseite der entsprechenden Mediatheken (Aufgabe 3). Aufgabe 2 diente ausschließlich dem sog. Freien Surfen in der jeweiligen Mediathek.

4.1.1 Festlegung von Navigationspfaden zur Mediathek (Aufgabe 1)

Die Abbildung 3 zeigt die optimalen Navigationspfade zu den einzelnen Mediatheken, die im Vorfeld identifizierten wurden. Den Ausgangspunkt stellen hier die Senderseiten von „Das Erste“, „ZDF“ und „WDR“ dar. Unterschieden werden drei Strategieansätze zur Lösung der gestellten Testaufgabe:

1. Die globale Navigation auf den Webseiten der TV-Sender,
2. Die Nutzung der Suchfunktion auf den Webseiten der TV-Sender,
3. Aufruf der Mediatheken und Bedienung selbiger.

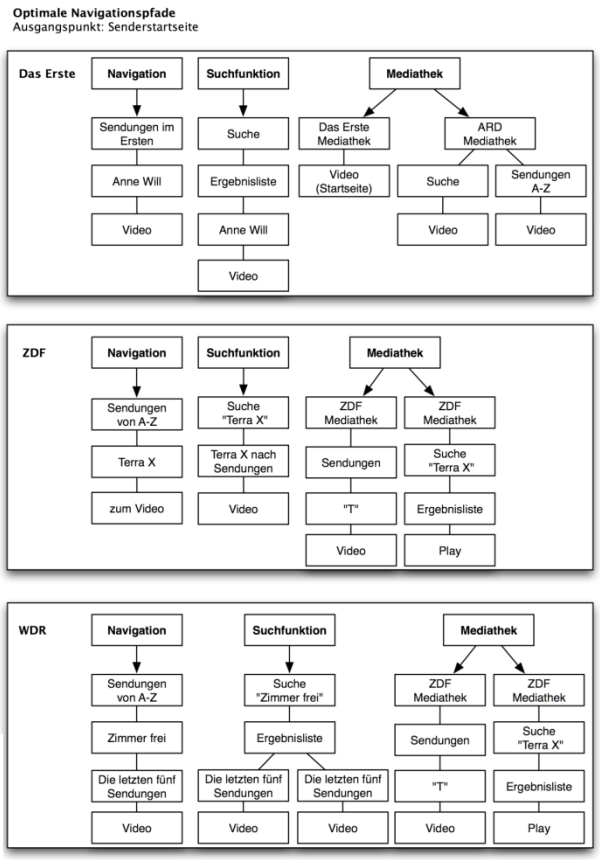


Abbildung 3: Optimale Navigationspfade Aufgabe 1 (Quelle: Eigene Darstellung)

Eine Besonderheit weist die Webseite von „Das Erste“ auf, da die Startseite sowohl über die „ARD Mediathek“ als auch die Mediathek von „Das Erste“ als einzelne Navigationspunkte/-buttons verfügt.

Die obige Abbildung visualisiert somit für die drei untersuchten Sender die für einen User mindestens erforderlichen Klicks um einen der drei Wege zum Ziel zu beschreiten. Verfolgte beispielsweise ein User bei der ZDF-Sender-Webseite die Navigationsstrategie „Suchfunktion“, so konnte das gesuchte Video frühestens nach drei Klicks aufgerufen werden.

4.1.2 Festlegung von Navigationspfaden in Mediathek (Aufgabe 3)

Analog zu den in Kapitel 4.1.1 aufgeführten optimalen Navigationspfaden für die Sender-Webseiten wurden diese für die dritte Aufgabe erstellt. Wie bereits weiter oben erwähnt, war der Ausgangspunkt dieser Aufgabe die Startseite der Mediatheken. Folglich entfiel die im vorigen Kapitel aufgeführte Strategie „Aufruf der Mediathek“. Abbildung 4 stellt die optimalen Navigationspfade grafisch dar.

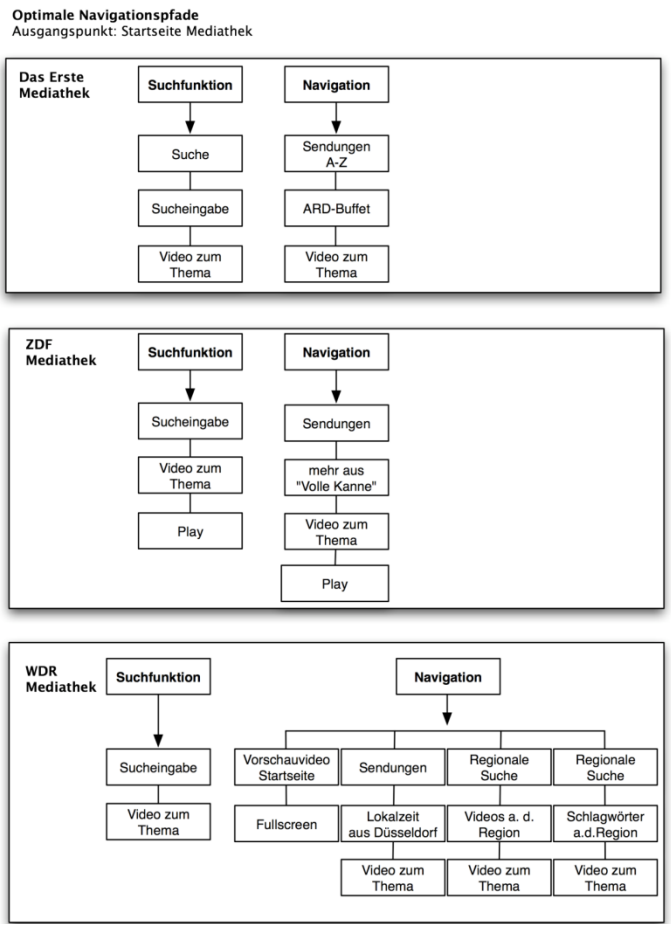


Abbildung 4: Optimale Navigationspfade Aufgabe 3 (Quelle: Eigene Darstellung)

4.2 Nutzungsstrategien

In den nachfolgenden Tabellen wird dargestellt, welche grundsätzlichen Navigationsstrategien von den Probanden verfolgt wurden. So konnte beobachtet werden, dass die Mehrheit der Probanden die globale Navigation als erfolgsversprechende Strategie ansahen, um das gesuchte Video zu finden. Es konnte zudem festgestellt werden, dass der Aufruf der gesuchten Inhalte über eine Mediathek nur von wenigen Nutzern als SuchansatzVerwendung fand.Bei Aufgabe 3 konnte bei den Mediatheken vom „ZDF“ und vom „WDR“ eine annährende Gleichverteilung bei der Strategiewahl festgestellt werden. Bei der Mediathek „Das Erste“ konnte hingegen eine

deutliche Diskrepanz zugunsten der Navigation beobachtet werden. Dies kann möglicherweise dem Umstand geschuldet sein, dass die Mediathek über kein seitenintegriertes Suchfeld verfügt. Um die Suche nutzen zu können, musste zunächst der Navigationspunkt „Suche“ aufgerufen werden.

Aufgabe 1

Strategie	Das Erste	ZDF	WDR
Navigation	52%	63%	81%
Suchfunktion	30%	11%	15%
Mediathek	7% Das Erste 11% ARD	26%	4%

Tabelle 2: Wahl der Navigationsstrategie bei Aufgabe 1 (Quelle: Eigene Auflistung)

Wird ein qualitativer Vergleich zwischen den Suchstrategien von Aufgabe 1 und Aufgabe 3 durchgeführt, so können Verschiebungen bei der Wahl der Nutzungsstrategien festgestellt werden. Bei der Mediathek vom WDR nutzten die Probanden deutlich häufiger die seiteninterne Suche als dies bei der Sender-Webseite der Fall war (44 Prozent versus 15 Prozent).

An dieser Stelle sei noch einmal erwähnt, dass aufgrund der geringen Fallzahl der Einsatz von Signifikanztests nicht zielführend ist. Somit haben die hier aufgeführten vergleichenden Aussagen einen explorativen Charakter. Diese Erkenntnisse bieten Ansatzpunkte für nachgelagerte, quantitative Studien, bieten jedoch nicht die Basis für allgemeingültige Aussagen.

Aufgabe 3

Strategie	Das Erste Mediathek	ZDF Mediathek	WDR Mediathek
Navigation	78%	66%	71%
Suchfunktion	22%	33%	30%

Tabelle 3: Wahl der Navigationsstrategie bei Aufgabe 3 (Quelle: Eigene Auflistung)

Neben der Betrachtung der allgemeinen Navigationsstrategien in dem ersten Teil dieses Kapitels wird nachfolgend eine weitergehende Differenzierung der Navigationsoptionen für Aufgabe 3 (Mediatheken) vorgenommen. Tabelle 4 stellt diese zusammen mit den beobachteten Nutzungszahlen grafisch dar.

Die Mediathek von „Das Erste“ besitzt in der horizontalen Hauptnavigationsleiste sechs verschiedene Navigationsoptionen, von denen jedoch nur die Bereiche „Sendungen von A-Z“ (78%) und die Suchfunktion (22%) genutzt werden. In der Mediathek vom „ZDF“ bietet die Hauptnavigationsleiste sechs verschiedene Möglichkeiten der Benutzerführung, die Subnavigationsleiste ist scrollbar und im Content-Bereich befinden sich weitere Direktverweise zu verschiedenen Videoinhalten.

Aufgabe 3

Navigations-optionen	Das Erste Mediathek		ZDF Mediathek		WDR Mediathek	
Navigation	Sendungen A-Z	78%	Sendungen	33%	Sendungen	26%
					Regionale Suche	26%
			Sendung verpasst	30%	Videovorschau	7%
					Regionen	4%
			Redaktionstipps Startseite	3%	30-Tage-Archiv	4%
					Schlagwörter a. d. Region	4%
Sucheingabe	22%		33%		30%	

Tabelle 4: Detaillierte Aufstellung der Navigationswahl bei Aufgabe 3 (Quelle: Eigene Auflistung)

4.3 Effizienz und Effektivität der Suchstrategie

Neben der Wahl der präferierten Navigationsstrategie wird nachfolgend die Effizienz und Effektivität dieser Auswahl näher betrachtet. Die Zielgrößen Effizienz und Effektivität sind dabei gängige Indikatoren in der Usability-Forschung. Eine Definition dieser Kenngrößen ist in der ISO-Norm 9241-11 aufgeführt [DI98]. So handelt es sich bei der Effektivität um den Präzisionsgrad und die Vollständigkeit, mit der ein Nutzer seine individuellen Ziele mit dem Angebot, in diesem Fall folglich den Mediatheken, erreichen kann. Die Effizienz hingegen beschreibt darüber hinausgehend die Relation zwischen Präzisionsgrad und Vollständigkeit der Aufgabenlösung einerseits und Ressourcen zur Erzielung der Aufgabenlösung (z.B. Zeiteinsatz) andererseits.

Wie Tabelle 5 zu entnehmen ist, liegt der durchschnittliche Erfüllungsgrad der beiden Aufgaben bei den Mediatheken von „DasErste“ und „WDR“ bei um die 90 Prozent. Der Grad der Zielerreichung bei der Mediathekvom „ZDF“ liegt jedoch auffallend dahinter. Bei der ZDF-Mediathek waren nur ca. drei Viertel der Probanden in der Lage, die jeweilige Aufgabe zu erfüllen. Eine Replay-Analyse der Datenaufzeichnung ergab, dass insbesondere Orientierungs-Probleme bei der Nutzung der horizontalen Navigation festzustellen waren. Zudem lieferte die Suche oftmals wenig zielführende Ergebnisse – vielfach keine.

Effektivität

	Das Erste	ZDF	WDR
Aufgabe 1.	93%	74%	89%
Aufgabe 3.	96%	81%	93%

Tabelle 5: Vergleich der Effektivität (Quelle: Eigene Auflistung)

Für die Bewertung der Effizienz wurden in dieser Studie zwei Determinanten zugrunde gelegt, der Zeitaufwand für die Zielerreichung und das Ausmaß der Abweichung vom optimalen Navigationspfad (Zusatzschritte). Als Zusatzschritt wird nach dem Verständnis der Autoren jeder zusätzliche Navigationsschritt angesehen, der unnötigerweise benötigt wird, um die gestellte Aufgabe zu lösen. Werden für die gewählte Strategie „Navigation“ beispielsweise sieben statt der im Optimum vorgesehenen vier Navigationsschritte gebraucht, so benötigt der Proband für die Lösung seiner Aufgabe drei Zusatzschritte.

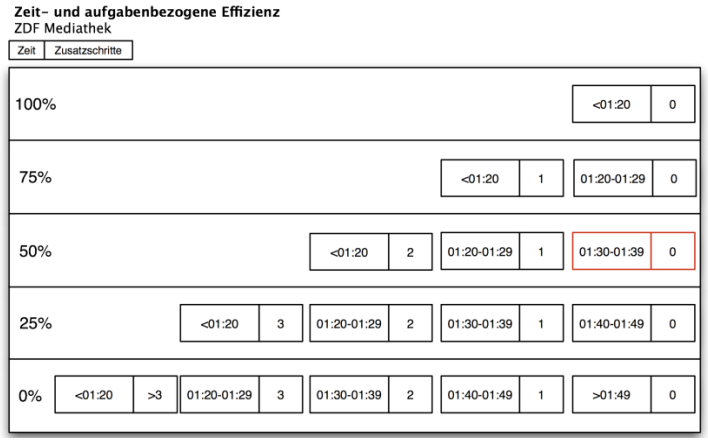


Abbildung 5: Zeit- und aufgabenbezogenen Effizienz bei der Lösung von Aufgabe 3 auf der ZDF-Mediathek (Quelle: Eigene Darstellung)

Als Zeitbezug wurde die durchschnittliche Dauer für die Lösung einer Testaufgabe herangezogen. Auf der Webseite von „Das Erste“ benötigten die Probanden durchschnittlich 52 Sekunden für die Lösung der gestellten Aufgabe, auf der Webseite des „ZDF“ 00:59 Minuten und beim „WDR“ 01:08 Minuten. Bei der dritten Aufgabe betrug der durchschnittliche Zeitaufwand zur Lösung der gestellten Aufgabe bei der Mediathek „Das Erste“ 00:37 Minuten, bei der „ZDF“-Mediathek“ 01:35 Minuten und bei der „WDR“-Mediathek 00:35 Minuten.

Beispielhaft für die ZDF-Mediathek (dritte Aufgabe) wird in der nachfolgenden Abbildung die Zeit- und aufgabenbezogene Effizienz schematisch dargestellt. In den einzelnen Ausprägungen werden jeweils links die Zeiten in Minuten und rechts die benötigten Zusatzschritte ausgewiesen.

Effizienz

	Das Erste	ZDF	WDR
Aufgabe 1.	49%	36%	40%
Aufgabe 3.	60%	27%	44%

Tabelle 6: Darstellung der Effizienz bei den gestellten Aufgaben (Quelle: Eigene Auflistung)

Wird die Effizienz der Suchstrategien betrachtet, so ergibt sich ein deutlich reservierteres Ergebnis bezüglich der Qualität der Nutzerführung auf den Sender-Webseiten und den Mediatheken. Wie in Tabelle 6 dargestellt, lag die Effizienz der Aufgabenerfüllung beim ZDF bei 36 Prozent (Sender-Webseite) respektive 27 Prozent (Mediatheken). Dies bedeutet, dass die Lösung der Aufgaben nur mit relativ hoher Energieleistung möglich war. Im Vergleich der drei Mediatheken konnte die Probanden auf der Mediathek „Das Erste“ am effizientesten

die gestellten Aufgaben lösen. Auch bei Aufgabe 1 erreichte die Webseite von „Das Erste“ die höchste Effizienzrate.

Wird die Effizienz getrennt nach den eingeschlagenen Suchstrategien ausgewertet, so ergeben sich die in Tabelle 7 dargestellten Werte. Aus der Tabelle sticht insbesondere die sehr geringe Effizienz bei der Strategie „Navigation“ der ZDF-Mediathek hervor. Es scheint so, dass die vom ZDF gewählte horizontale Navigationsstruktur wenig zweckmäßig für eine Mediathek ist. Die bereits erwähnten Orientierungsprobleme bei der ZDF-Mediathek mögen hierin ihre Ursache haben. In dem nachfolgenden Kapitel werden einige ausgewählte Nutzungsprobleme dargestellt, die die festgestellten Effizienzzahlen begründen können.

Das Erste			
Aufgaben	**Navigation**	**Suche**	**Mediathek**
Aufgabe 1	51%	31%	(Das Erste) 88% (ARD) 58,3%
Aufgabe 3	56,8%	75%	/
Durchschnitt	**54,3%**	**53%**	**73%**
ZDF			
Aufgaben	**Navigation**	**Suche**	**Mediathek**
Aufgabe 1	29,4%	16,7%	60,7%
Aufgabe 3	16,1%	41,7%	/
Durchschnitt	**22,75%**	**29,2%**	**60,7%**
WDR			
Aufgaben	**Navigation**	**Suche**	**Mediathek**
Aufgabe 1	42%	37,5%	0%
Aufgabe 3	35,4%	50%	/
Durchschnitt	**38,7%**	**44%**	**0%**
Gesamt-durchschnitt (Das Erste, ZDF, WDR)	**38,7%**	**44%**	**0%**

Tabelle 7: Effizienz der Aufgaben differenziert nach Suchstrategien (Quelle: Eigene Darstellung)

5 Nutzungsprobleme

Die Analyse der drei untersuchten Mediatheken hat verschiedene Schwachstellen aufgezeigt. Anhand von Dos und Don’ts konnten zentrale Handlungsempfehlungen abgeleitet werden. Diese umfassen insbesondere die Navigations- und Informationsstruktur sowie die Qualität der Suche und der Präsentation der Suchergebnisse.Aus der Replay-Analyse und den in Tabelle 6 dargestellten Effizienzzahlen sollte eine vertikale Navigationsstrategie insbesondere für die zweite Hierarchieebene

Abbildung 6: Das Erste Mediathek – vertikale Navigation (Quelle: http://mediathek.daserste.de)

einer horizontalen Hierarchie vorgezogen werden (vgl. Abbildung 6).

Auf die in Abbildung 7 dargestellte horizontale Navigation der ZDF-Mediathek sollte hingegen möglichst verzichtet werden, da bei großer Datenmenge die Orientierung für den Nutzer erschwert wird.

Abbildung 7: ZDF Mediathek – horizontale Navigation
(Quelle: http://www.zdf.de/ZDFmediathek/)

Generell konnte bei allen untersuchten Stimuli eine mangelhafte interne Suchfunktion festgestellt werden. So lieferten die Such-Engines oftmals keine oder nur wenig zielführende Suchergebnisse. Insbesondere bei großen Datenbanken wie Mediatheken besteht hier ein dringender Optimierungsbedarf.

6 Fazit

Anhand einer explorativen Studie wurden mögliche Navigationsstrategien in Mediatheken identifiziert. Die Usability-Kriterien der Effektivität und Effizienz wurden für die öffentlich-rechtlichen Angebote überprüft. Beispielhafte Nutzungsprobleme der Mediatheken von DasErste.de, ZDF.de und WDR.de wurden ebenfalls aufgezeigt. Nächste Schritte in der Forschung zur Bewegtbildkommunikation im Web können hierauf aufsetzen.

7 Literaturverzeichnis

[DA06] Dahm, M. Grundlagen der Mensch-Computer-Interaktion. München: Pearson, 2006.

[DI98] DIN EN ISO 9241, Teil 11: Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten. Anforderungen an die Gebrauchstauglichkeit – Leitsätze. Deutsche Fassung EN ISO 9241-11, Berlin: Beuth , 1998.

[FU09] Funk, L. und Pagel, S. Wettbewerbsökonomische Analyse des Internetfernsehens: Ein erster Überblick. In: Krone, J. (Hrsg.) Bleibt Fernsehen Fernsehen? (i.D.).

[GE08] Gerhards, C. und Pagel, S. Webcasting von Video-Content in Online-Zeitungen. Marktanalyse – Kosten – Erlöse. In: Zerfaß, A. und Welker, M. und Schmidt, J. (Hrsg.): Kommunikation, Partizipation und

Wirkungen im Social Web. Bd. 2. Köln: Herbert von Halem Verlag, 2008, S. 154-187.

[GE09] Gerhards, C. und Pagel, S. Internetfernsehen von TV-Sendern und User Generated Content. In: Friedrich-Ebert-Stiftung, Reihe Medien Digital, 2009.

[GO02] Goldstein, B. Wahrnehmungspsychologie, 2. Aufl., Heidelberg: Spektrum, 2002.

[GS 07] Gscheidle, C. und Fisch, M. Onliner 2007: Das ‚Mitmach-Netz' im Breitband-Zeitalter. In: Media Perspektiven, Heft 8/2007, S. 393–405, 2007.

[ME08] Merkel, W. ARTE+7 – die ARTE-Online-Mediathek. In: Fernseh- und Kinotechnik, Heft 4/2008, S. 168-173, 2008.

[NE00] Neuberger, C. Journalismus im Internet. In: Media Perspektiven, 31. Jahrgang, Heft 7/2000, S. 310-318, 2000.

[PA08a] Pagel, S. Partizipative Softwareentwicklung für das Internetfernsehen am Beispiel von Online-Mediatheken. In: Gadatsch, A. und Vossen, G. (Hrsg.): Auswirkungen des Web 2.0 auf Dienste und Prozesse, EMISA 2008, Tagungsband, St. Augustin 2008, S. 25-38, 2008.

[PA08b] Pagel, S. und Goldstein, S. und Jürgens, A. Erste methodische Erkenntnisse zur Usability-Analyse von Video-Inhalten auf Websites mittels Eyetracking. In: Brau, H., Diefenbach, S., Hassenzahl, M., Koller, F. und Peissner, M.und Röse, K. (Hrsg.): Usability Professionals 2008, Stuttgart 2008, S. 177-181, 2008.

[PAG09] Pagel, S. und Goldstein, S. Nutzung und Wirkung von Video-Content in Online-Jobbörsen. Erkenntnisse einer explorativen Studie. In: Forschungsberichte des Fachbereichs Wirtschaft der Fachhochschule Düsseldorf (i.V.).

[SC08] Schmidt, J. Produktionstechnik der ZDFmediathek. In: Fernseh- und Kinotechnik, Heft 4/2008; S. 161-167, 2008.

[VA07] Van Eimeren, B. und Frees, B. Internetnutzung zwischen Pragmatismus und YouTube-Euphorie. In: Media Perspektiven, Heft 8/2007, S. 362-378, 2007.

[VA08] Van Eimeren, B. und Frees, B.Internetverbreitung: Größter Zuwachs bei Silver-Server. In: Media Perspektiven, Heft 7/2009, S. 330-344, 2008.

Video-Tools im Schulunterricht: Psychologisch-pädagogische Forschung zur Nutzung audiovisueller Medien

Carmen Zahn, Karsten Krauskopf und Friedrich W. Hesse

Institut für Wissensmedien, Tübingen

c.zahn@iwm-kmrc.de

Zusammenfassung: Dieser Beitrag beschäftigt sich mit dem praktischen Einsatz von audiovisuellen Medien im schulischen Kontext: Vorgestellt werden zwei psychologische Untersuchungen zur Nutzung von Video-Tools im Deutsch- und Geschichtsunterricht. Die Ergebnisse zeigen, dass neue technologische Entwicklungen Vorteile für den Einsatz audiovisueller Medien in Lernszenarien auf individueller vor allem aber auch auf sozio-kognitiver Ebene erbringen. Konsequenzen für die Gestaltung von Webarchiven werden diskutiert.

Schlagwörter: Audiovisuelle Medien, Video-Tools, digitale Werkzeuge, gestaltendes Lernen, learning through design, Wissensprozesse, Wissensaustausch, Schulunterricht

1 Einleitung

Audiovisuelle Medien, die in Medienarchiven zur Verfügung gestellt werden, sind eine wichtige Ressource für schulisches Lernen. Sie können als Lehrfilme oder Quellen in verschiedenen Unterrichtsfächern verwendet werden. Allerdings besteht ihr Potential für die Schule nicht ausschließlich in der Informations*darstellung*, denn: Audiovisuell dargestellte Inhalte gelten im Alltag oft fälschlicherweise als leicht verständlich und werden deshalb nur oberflächlich rezipiert. Die medienpsychologische Forschung hat dagegen gezeigt, dass reines Betrachten audiovisueller Medien für die genaue Erschließung der enthaltenen Information oft nicht ausreicht [SP07]. Im Gegensatz zu textbasierten Medien sind audiovisuelle Medien zeitabhängig, d.h. „flüchtig“ und stellen besonders hohe Ansprüche an die individuelle kognitive Verarbeitung, bzw. kollaborative Verarbeitung, (d.h. das Zusammenwirken kognitiver Prozesse, die über mehrere Personen verteilt sind). Diese Befunde sind von zentraler Bedeutung, wenn audiovisuelle Medien im Schulunterricht eingesetzt werden sollen.

Im vorliegenden Beitrag beschäftigen wir uns mit so genannten *Video-Tools*. Video-Tools sind digitale Werkzeuge, die mit einer unterschiedlichen Breite an Möglichkeiten (Ausschnitte bilden, Annotieren, Kommentieren, Taggen, Einfügen von Hyperlinks) für die Kontextualisierung, Analyse und weitere Bearbeitung von Videos aus Medienarchiven genutzt werden können und auf diese Weise sozio-kognitive Aktiväten auf Seiten der Nutzer gezielt unterstützen [GO07].

Ein Beispiel ist das Tool DIVER/WebDIVER™ [PEA04], das von der Arbeitsgruppe um Roy Pea an der Stanford University (Stanford Center for Innovations in Learning - SCIL) entwickelt wurde (s. Abbildung 1). Es wurde speziell für das kollaborative Arbeiten mit Videos konzipiert und erlaubt das Ausschneiden einzelner Bildteile aus einem Video mittels einer Art. Hierdurch sollen zentrale visuell-analytische Fähigkeiten, wie das genaue Beobachten realer Szenen und die Identifikation wichtiger Details in einem Videobild, ermöglicht werden. Bei der Arbeit mit WebDIVER™ werden Bildausschnitte nicht nur kognitiv ausgewählt und fokussiert, sondern durch die reale Manipulation am Bild ausgeschnitten. Dabei werden keineswegs nur statische Bildausschnitte generiert, sondern Bildausschnitte mit zeitlicher Erstreckung – so genannte „DIVES“. Beispielsweise kann in der Videoaufnahme ein einzelnes Objekt oder eine einzelne Person ins Visier genommen werden, indem man „heranzoomt“ und das Geschehen über eine gewisse Zeitspanne hinweg mitschneidet. Den anschließend aus dem Gesamtvideo heraus exportierten, isolierten Bildausschnitt kann man dann getrennt vom Ursprungsvideo betrachten, schriftlich kommentieren und über eine Web-Anbindung auch gemeinsam mit anderen analysieren und interpretieren. Erklärtes Ziel der Entwicklung des DIVER/WebDIVER™-Werkzeugs ist die Schulung des genauen Beobachtens in verschiedenen Bereichen. Die zugrunde liegenden Schlüsselkonzepte umfassen gelenktes Wahrnehmen („Guided Noticing“ ™) und Multiperspektivität.

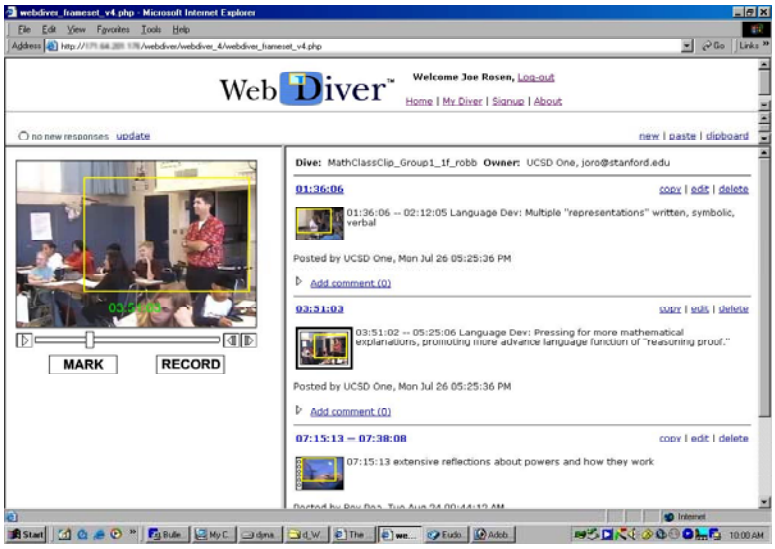


Abbildung 1: Screenshot WebDiver™. Links: Videopräsentation und Funktionen zur Selektion der Bildausschnitte; rechts: Arbeitsbereich für kooperatives Arbeiten an den Bildausschnitten.

Überträgt man die Überlegungen zu DIVER/WebDIVER™ und ähnlichen Video-Tools [GO07] auf schulisches Lernen, ergeben sich neue Chancen und zugleich Herausforderungen für diesen Kontext [ZA05]. Die Chancen bestehen darin, dass solche digitalen Technologie-Impulse das Einsatzspektrum von audiovisuellen Medien im Unterricht erweitern, denn es entstehen bessere Rahmenbedingungen für *aktives*

Lernen. Dies wiederum ist mit nachhaltigen Wissenserwerbs-, Lern- und Erkenntnisprozessen verbunden, die letztendlich auch zur Ausbildung von Kernkompetenzen (z.B. visuelle Medienkompetenz, kritisches Denken) führen. Darüber hinaus eröffnet die Möglichkeit der aktiven Arbeit an im Web archivierten audiovisuellen „Quellen“ neue Zugänge für spätere Nutzer, die so auf Ergebnisse der Reflexionen anderer Lerner zugreifen und diese einbeziehen können. Um die erwünschten Lernziele aber tatsächlich zu erreichen, müssen im Schulunterricht Lerngelegenheiten geschaffen werden, bei denen die Video-Tools in der beabsichtigten Weise einsetzbar sind.

Hierfür scheint das gestaltende Lernen in Designprojekten (*learning through design*, [KA96] besonders geeignet. Designprojekte bedingen aktives, projektorientiertes Lernen mit Medien [BA99]. Sie erhöhen die intrinsische Motivation zu lernen, weil die Lernenden produktiv sein können und sich als kompetent erleben, wenn sie selbstständig etwas für sie Authentisches, Sinnvolles und Bedeutsames leisten [BE94]. Außerdem wird ein besseres Verständnis des jeweiligen Themengebiets erwartet, denn: Während sie „als Designer“ eine Informationsstruktur schaffen, müssen die Lernenden Inhalte auswählen, vergleichen und darüber reflektieren, wodurch sie sich das jeweilige Wissen besonders gut aneignen [BER87]. Und schließlich fördert projektorientiertes Lernen sowohl die Problemlösefähigkeit als auch die Medienkompetenzen, weil die Lernenden Informationsstrukturen planen, kreativ mit verschiedenen Medienformaten arbeiten und diese sinnvoll zur Darstellung der Informationen kombinieren müssen [BE94].

Ob die Bearbeitung gestalterischer Aufgaben mit Video-Tools im Schulunterricht zum erwünschten Ziel führt, ist allerdings noch unklar. Ebenso ist noch offen, inwiefern das Gestalten mit *unterschiedlichen* Video-Tools *unterschiedliche* Wissensprozesse, Gruppendiskussionen und letztlich Designprodukte bewirkt.

2 Empirische Studien

Vor diesem Hintergrund, untersucht ein DFG-Projekt exemplarisch das gestaltende Lernen mit audiovisuellen Medien und Video-Tools im Deutsch- und Geschichtsunterricht. Ziel ist es, die Randbedingungen für effektives Lernen mit audiovisuellen Medien im Schulunterricht herauszuarbeiten, wobei der Fokus auf den funktionalen Eigenschaften von Video-Tools liegt. Hierfür wurde eine prototypische Design-Aufgabe entwickelt, in der Schüler(innen) gebeten werden, in Partnerarbeit eine digitalisierte Original-Kino-Wochenschau aus dem Jahr 1948 zum Thema „Berliner Luftbrücke“ auszugestalten, um das Thema *für andere* Mitschüler auf einer Webseite attraktiv und verständliche Weise zugänglich zu machen. Die Zielvorgabe der Aufgabe ist die Gestaltung eines Entwurfs für die Seite eines virtuellen Geschichtsmuseums (LeMO, http://www.dhm.de/lemo/home.html). Die *Lern*ziele für die Schülerinnen und

Schüler umfassen (im Einklang mit den Bildungsstandards für Realschule/Gymnasium) sowohl inhaltliches Wissen und historisches Verständnis von Medien (hier: Propaganda im Nachkriegsdeutschland) als auch visuelle Medienkompetenz (Filmanalyse, kritische Reflexion der Wochenschau als historische Quelle, Fertigkeiten der Nutzung von Video-Tools). Die praktische Durchführbarkeit dieser Design-Aufgabe und die beteiligten Wissensprozesse werden detailliert in einer Serie von Labor- und Unterrichtsexperimenten untersucht. Die beiden hier vorgestellten Studien adressieren digitale Video-Tools als Unterstützung für effizienten Wissensaustausch und kollaboratives Lernen mit audiovisuellen Medien in dyadischen Lerngruppen, das die Resultate festhält.

2.1 Vorstudie: Führt das Arbeiten mit Video-Tools zu Wissensaustausch, Lernen und zum Erwerb visueller Kompetenz?

Die oben skizzierte kollaborative Design-Aufgabe wurde zunächst in einer experimentellen Laborstudie mit studentischen Versuchspersonen (N=24 Dyaden) eingesetzt, in der die Frage nach spezifischen Tool-Einflüssen untersucht wurde: Konkret wurde das Video-Tool WebDIVER^{TM} als *kollaboratives Werkzeug* mit spezifischen technischen Funktionen für gemeinsame Filmanalyse und Erstellung direkt filmbildbezogener Kommentare mit einer einfacheren technischen Lösung (Video-Player und Texteditor) verglichen. Die zugrunde liegende Fragestellung war, ob sich bei derselben Aufgabenstellung und denselben audiovisuellen Medien beim Arbeiten mit *unterschiedlichen* Bearbeitungswerkzeugen Unterschiede in den kollaborativen Verarbeitungsprozessen der Dyaden zeigen würden. Die Interaktionen der Dyaden während der kollaborativen Design-Aufgabe wurden für spätere Analysen auf Video aufgezeichnet (Screenvideo und Webcambild der Lernenden). Zudem wurde im Anschluss an die Aufgabenbearbeitung auf individueller Ebene der Lernerfolg erfasst (Qualität der Design-Produkte, Wissenstests zu historischen Inhalten, Transferaufgabe zu kritischer Reflexion audiovisueller Inhalte). Die Ergebnisse belegen insgesamt eine große Akzeptanz und Effektivität der Aufgabe: Über 70% der Bearbeitungszeit wurde aufgabenrelevanten Tätigkeiten gewidmet, der Wissenszuwachs zum inhaltlichen Thema (Berliner Luftbrücke) war substantiell. Zudem zeigten sich signifikante Einflüsse der Video-Tools auf die Designprozesse und den Lernerfolg. Die individuellen Leistungen der Probanden im anschließenden thematischen Wissenstest, sowie in der Transferaufgabe (Analyse eines Wahlwerbespots) unterschieden sich signifikant: Die Bedingung mit unterstützender Videotechnologie zeigte sich der Bedingung *ohne* Unterstützungsfunktion hinsichtlich des Wissenserwerbs ($p. < 05$, $d = 0.9$) und hinsichtlich des Transfers ($p < .02$, $d = 1.0$) überlegen. Das WebDIVER^{TM}–Werkzeug förderte außerdem die Zuwendung zu designrelevanter Gruppendiskussion, erhöhte die Qualität der Design-Produkte und hinterließ in Fallbeispielen nachweisbare “Spuren“ in den Diskussionsverläufen der

Dyaden. Hier zeigte sich insbesondere, dass die Verwendung von WebDIVER™ dazu führte, dass die Überlegungen der Zweiergruppen zu Video-Sequenzen oder Kamera-Einstellungen identifizierbar für diese Selektionen festgehalten wurden. Um die Gültigkeit dieser Laborbefunde zu prüfen, wurde anschließend eine Feldstudie mit Schulklassen durchgeführt.

2.2 Hauptstudie: Video-Tools im Unterricht

In der experimentellen Feldstudie bearbeiteten 234 Schülern und Schülerinnen (N = 117 Dyaden) der 10./11. Klasse (Alter 16 Jahre, Gymnasium und Realschule) in einer Doppelstunde die unter 2 beschriebene kollaborative Design-Aufgabe. Vier verschiedene Experimentalbedingungen unterschieden sich zum einen in der Art der verwendeten Video-Tools (Faktor 1) zum anderen in den vorgegebenen Designzielen (Faktor 2). Faktor 1 variierte wie in der Vorstudie (DIVER ™ vs. Video Player + Texteditor), Faktor 2 variierte im Hinblick auf zwei Designziele, die den Schülern anhand verschiedener Metaphern für das zu erstellende Designprodukt vorgegeben waren: In der einen Instruktion wurde den Schülern nahe gelegt ein Produkt zu gestalten, das es erlaubt, im virtuellen Museum „in die Kino-Wochenschau einzutauchen“, in der anderen wurde ihnen gesagt, sie sollten für das virtuelle Museum ein „Video mit Lesezeichen“ gestalten. Die Forschungsfragen bezogen sich auf die Auswirkungen der verschiedenen Bearbeitungswerkzeuge (siehe Vorstudie 2.1), die Auswirkungen der verschiedenen Designziele und auf mögliche Interaktionen der beiden Faktoren. Die Durchführung der Studie erfolgte wie in der Vorstudie, jedoch vor Ort an den Schulen mit Hilfe eines mobilen Notebook-Klassenzimmers. Die Ergebnisse replizieren zu weiten Teilen die Laborergebnisse: Der Lernerfolg war in allen Bedingungen insgesamt gut, unterschied sich jedoch nicht zwischen den Bedingungen. Was den Einfluss der Video-Tools (Faktor 1) betrifft, so zeigte sich, dass die Schülerinnen und Schüler, die mit DIVER™ zusammenarbeiteten, qualitativ hochwertigere Designprodukte erstellten, die signifikant mehr Details enthielten ($p < .01$, Partielles Eta $^2 = 0.9$). Außerdem diskutierten sie mehr über Designfragen und Inhalte unabhängig vom Originalvideo ($p < .05$, Partielles $Eta^2 = .44$) als die Schülerinnen und Schüler in den anderen Bedingungen. Sie baten außerdem tendenziell seltener die Versuchsleiter um Hilfe ($p < .10$, Partielles $Eta^2 = .26$). Insgesamt zeigte sich damit bei den Schülerinnen und Schülern, die mit dem Video-Tool arbeiteten, ein autonomeres Lernverhalten während der kollaborativen Design-Aufgabe. Hinsichtlich der unterschiedlichen Designziele (Faktor 2) zeigten sich entgegen der Erwartungen jedoch keine signifikanten Unterschiede.

2.3 Diskussion

Die Ergebnisse aus den beiden Studien zeigen, dass die funktionalen Eigenschaften von Video-Tools im Rahmen ein und derselben Design-Aufgabe unterschiedliche

Wirkungen auf Wissensprozesse haben können. Diskussionsverläufe, Designprodukte und Lernerfolg unterschieden sich in den Studien je nach Wahl des „Arbeitswerkzeugs“ für den Umgang mit audiovisuellen Medien. Diese Effekte traten nicht nur isoliert im Labor auf, sondern auch in der komplexen Lernsituation „Schulunterricht“. Hier jedoch deutlich weniger ausgeprägt. Die Ergebnisse aus einzelnen experimentellen Studien mit relativ kleinen Stichproben sind nur eingeschränkt gültig und können selbstverständlich nicht vorschnell auf andere Situationen übertragen werden. Deshalb werden weitere Studien nötig sein, um ein klares Bild davon zu bekommen, wie audiovisuelle Medien und Video-Tools für im konstruktivistischen Schulunterricht nutzbar sind.

Sollten sich die Ergebnisse bestätigen, so haben sie wichtige praktische Implikationen für die Gestaltung attraktiver Benutzeroberflächen von Medienarchiven oder Mediatheken: Für die Wissenskommunikation sind Video-Tools nicht nur ein attraktives „Add-on“, sondern essentiell wichtig. Wenn Archivnutzern neben den technischen Funktionen für Informationssuche und –rezeption auch Video-Tools zur Unterstützung spezifischer kognitiver und kollaborativer Funktionen angeboten werden, so wird ihnen die vertiefende Informationsverarbeitung, -bewertung und der Wissenssaustausch mit audiovisuellen Medien erleichtert. Dazu ermöglicht die Erstellung von Kommentaren, Verweisen, etc. auf detaillierter, inhaltlicher Ebene den Zugang zu den einzelnen Quellen sowie die Einbindung in verwandte Strukturen, z.B. Text- und (statische) Bildarchive. Dies beinhaltet eine verbesserte Nutzbarkeit audiovisueller Medien(archive) nicht nur für den Schulunterricht, sondern auch in Lern- und Arbeitssituationen (STA07), bis hin zu informellen Settings (z.B. Museen, Museumsarchiven).

3 Literaturverzeichnis

[BA99] Baake, D. Medienkompetenz als zentrales Operationsfeld von Projekten. In Baake, D., Kornblum, S., Lauffer, J., Mikos, L. und Thiele, G.A. (Eds.), *Handbuch Medien: Medienkompetenz - Modelle und Projekte* (pp. 31-35). Bonn: Bundeszentrale für politische Bildung. (1999)

[BE94] Beichner, R. J. Multimedia editing to promote science learning. *Journal of Educational Multimedia and Hypermedia, 3*(1), (1994). 55-70.

[BER87] Bereiter, C., & Scardamalia, M., *The psychology of written composition.* Hillsdale N.J.: Erlbaum. (1987).

[GO07] Goldman, R., Pea, R., Barron, B. und Derry S. (Hrsg.), *Video research in the learning sciences* Mahwah, NJ: Lawrence Erlbaum Associates. (2007). 93-100.

[KA96] Kafai, Y.B. und Resnick, M.. (Hrsg.). *Constructionism in practice: Designing, Thinking, and Learning in a Digital World.* Mahwah, NJ: Erlbaum Press. (1996).

[PEA04] Pea, R., Mills, M., Rosen, J., Dauber, K., Effelsberg, W., & Hoffert, E. The Diver Project: Interactive Digital Video Repurposing. In *IEEE MultiMedia*: IEEE Computer Society (2007).

[SP07] Spiro, R., Collins, B. P. und Ramchandran, A. Reflections on a Post-Guttenberg epistemology for video use in ill-structured domains: Fostering complex learning and cognitive flexibility. In Goldman, R., Pea, R., Barron, B. und Derry S. (Hrsg.), *Video research in the learning sciences* Mahwah, NJ: Lawrence Erlbaum Associates. (2007). 93-100.

[ST07] Stahl, E., Zahn, C., & Seidel, T. Videobasierte Lernsoftware zur Förderung kommunikativer Kompetenzen. In Kanning, U.P. (Ed.), Förderung sozialer Kompetenzen in der Personalentwicklung. Göttingen: Hogrefe. (2007). 39-63.

[ZA05] Zahn, C., Pea, R., Hesse, F. W., Mills, M., Finke, M., und Rosen, J.. Advanced video technologies to support collaborative learning in school education and beyond. In T. Koschmann, T., Suthers, D. und Chan, T.W. (Hrsg.): *Computer supported collaborative learning 2005: The next 10 years.* Mahwah, NJ: Lawrence Erlbaum Associates. (2005). 737-742.

Einsatz Pixelbasierter Datenfusion zur Objektklassifikation

Jan Thomanek, Holger Lietz, Basel Fardi, Gerd Wanielik

Technische Universität Chemnitz
Fakultät für Elektrotechnik / Informationstechnik
Professur für Nachrichtentechnik

{jtho,holi,fardi,wanielik}@hrz.tu-chemnitz.de

Zusammenfassung: Dieser Beitrag beschreibt den Einsatz der Pixelbasierten Datenfusion zur Verbesserung der Objektklassifikation am Beispiel eines Fußgängererkennungssystems. Dabei werden zunächst sowohl die verwendeten Fusionstechniken, als auch die Grundlagen der Objekterkennung erläutert. Im Weiteren wird ein System zur Erkennung von Fußgängern beschrieben, welches auf der Fusion der Bilder einer Infrarot- und einer visuellen Kamera basiert. Es werden die Ergebnisse erläutert und verglichen.

Schlagwörter: Datenfusion, Objekterkennung, Klassifikation

1 Einleitung

Die rasante Entwicklung neuer leistungsstarker Sensoren, wie zum Beispiel im Bereich der Fahrzeugumfelderkennung, erfordert zunehmend intelligente Technologien zur effizienten Verarbeitung der gewonnenen Informationen. Dabei erfahren Systeme der Kombination von Daten unterschiedlicher Sensoren eine wachsende Bedeutung. Eine solche Fusion kann bei bildgebenden Sensoren bereits auf der Pixelebene erfolgen. Ein fusioniertes Bild bietet aufgrund des größeren Informationsgehaltes entscheidende Vorteile für eine weiterführende Bildverarbeitung, wie z.B. bei der Objekterkennung und Objektklassifikation. Unter anderem bildet hierbei auch die Fußgängererkennung für die unterschiedlichsten Anwendungen einen wichtigen Forschungsschwerpunkt vieler Unternehmen und staatlicher Institutionen. Allerdings bedeutet die Entwicklung einer rechnergestützten Fußgängererkennung insbesondere im Hinblick auf die verschiedenen Variationen beim Erscheinen eines Fußgängers oder die Vielzahl möglicher Licht- und Umweltbedingungen eine große Herausforderung. Mit einem einzigen Sensor ist es unmöglich, eine Umgebung komplett zu erfassen. So wie auch der Mensch seine verschiedenen Sinnesorgane gleichzeitig benutzt, ermöglicht erst die Kombination unterschiedlicher Sensoren eine vollständigere Erfassung und Interpretation des eigenen Umfeldes.

Dieser Beitrag beschreibt die Verfahren und vergleicht die Ergebnisse eines Fußgängererkennungssystems unter Anwendung einer pixelbasierten Datenfusion von einer *Far-Infrared*-Kamera (FIR) und einer visuellen Kamera. Dabei gibt Kapitel 2 einen Überblick über die verwendeten Fusionstechniken. Im Abschnitt 3 werden die

Algorithmen für Objekterkennung und Klassifikation erläutert. Kapitel 4 beschreibt die konkrete Anwendung der pixelbasierten Fusion in der Fußgängererkennung und diskutiert die gewonnenen Ergebnisse.

2 Pixelbasierte Datenfusion

2.1 Überblick

Generell kann man Daten bzw. Information von zwei oder mehreren Sensoren auf verschiedenen Ebenen der Abstraktion fusionieren. Bei der Fusion auf Signalebene erfolgt eine direkte Kombination der Sensorsignale. Die Fusion auf Bildebene wird durchgeführt, indem jedes Pixel bzw. Pixelgruppen der Eingangsbilder bewertet und schließlich zu einem fusionierten Bild kombiniert werden. Im Gegensatz dazu werden bei der Fusion auf Merkmalsebene zunächst die Merkmale aus den jeweiligen Eingangsbildern extrahiert und diese anschließend kombiniert. Schließlich können Information auch nach ihrer Klassifikation auf Symbolebene fusioniert werden.

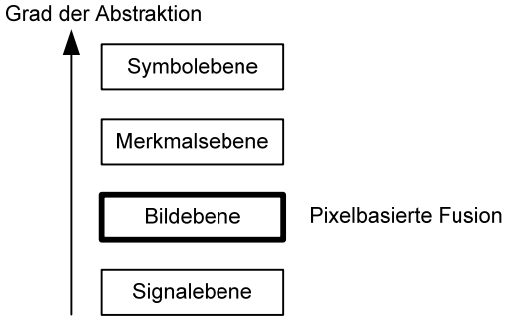


Abbildung 1: Abstraktionsebenen der Datenfusion

Gegenstand der pixelbasierten Fusion ist es nun, durch Kombination der Eingangsbilder ein Mischbild zu erzeugen, welches einerseits für die weiterführende rechnergestützte Bildverarbeitung, wie z.B. Segmentierung, Feature-Extraktion oder Zielerkennung, aber auch andererseits für eine menschliche Erfassung der Bildinformation, eine genauere Beschreibung der Szene darstellt als ein einzelnes Eingangsbild [BLU06].

2.2 Bild-Registration

Eine wichtige Voraussetzung für die pixelbasierte Fusion ist, dass alle Eingangsbilder räumlich korrekt aufeinander abgestimmt sind, d.h. entsprechende Pixelpositionen in den Eingangsbildern repräsentieren denselben Punkt in der Welt (engl. *registration*). Aufgrund verschiedener Blickwinkel der Sensoren auf die Szene und möglicher unterschiedlicher intrinsischer Eigenschaften der Kameras entstehen Verzerrungen, die

durch geeignete geometrische Transformationen der Bildebenen kompensiert bzw. minimiert werden müssen [GOS05]. Allerdings berücksichtigen diese Verfahren nur die zweidimensionale Beziehung zwischen den Bildern, da eine solche Transformation jeweils nur für eine Ebene im Raum definiert ist. Von daher wird im vorliegenden Beitrag hinsichtlich der Anwendung ein neuer Weg beschritten, indem nicht die Bilder, sondern die Kamerasysteme transformiert werden. Hierbei werden beide Kameras in ein gemeinsames Stereo-Kamerasystem überführt, wo u. a. die optischen Achsen parallel ausgerichtet sind. Ein solches Verfahren wird *Rektifikation* genannt. Dabei müssen die Eingangsbilder in eine gemeinsame Bildebene transformiert werden, wo die Skalierung und die Disparität zwischen den Bildern in einer Dimension Null sind. Um auch in der anderen Richtung zumindest für Objekte ab einem entsprechenden Mindestabstand von den Kameras eine Disparität von Null zu erreichen, werden beide Sensoren so eng wie möglich zueinander positioniert.

Es gibt viele Möglichkeiten die Rektifikation zu bestimmen. Das vorgestellte Verfahren lehnt sich an die Methode von Fusiello [FUS06] an, wo die Rektifikation auf der Basis von Punktekorrespondenzen und ohne Kenntnis der Kalibrierung ermittelt wird. Dabei werden die Eingangsbilder durch geeignete Rotationen in die gemeinsame Ebene transformiert.

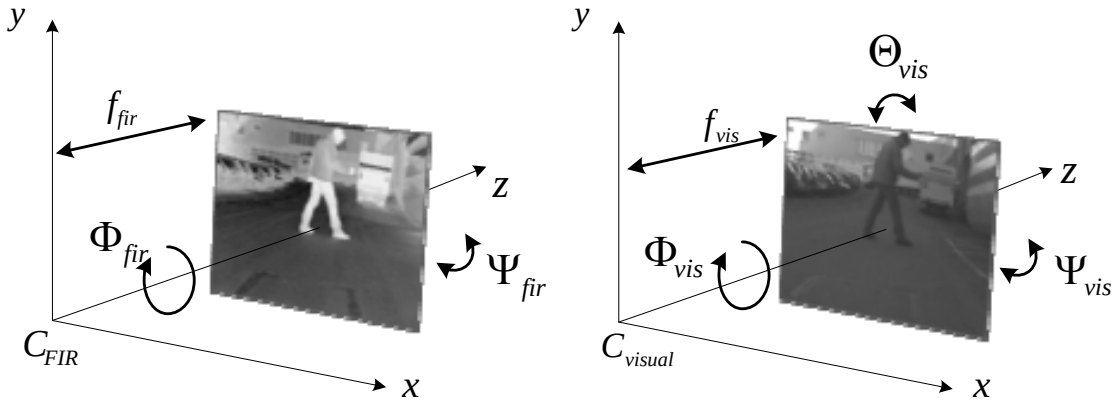


Abbildung 2: Freiheitsgrade bei der Rektifikation

Die Parameter für die Transformation der Kamerabilder werden in einem Kalibrierungsschritt offline ermittelt, wobei diese basierend auf den Punkte-Korrespondenzen $\{m_f^j, m_v^j\}$ durch Minimierung des geometrischen Re-Projektionsfehlers mittels eines nichtlinearen iterativen Optimierungsverfahren geschätzt werden [HAR03].

$$\sum_j \frac{\left(\mathbf{m}_v^{j\,T} F \mathbf{m}_f^j\right)^2}{\left(F\mathbf{m}_f^j\right)_x^2 + \left(F\mathbf{m}_f^j\right)_y^2 + \left(F^T\mathbf{m}_v^j\right)_x^2 + \left(F^T\mathbf{m}_v^j\right)_y^2} \rightarrow min \tag{1}$$

Dabei stellt die Fundamental-Matrix F die Beziehung der Bilder zueinander dar und berechnet sich im gegebenen Fall wie folgt:

$$F = K_v^{-T} R_v^T [u]_\times R_f K_f^{-1} \tag{2}$$

Die Rotationsmatrizen R_v, R_f und die intrinsischen Matrizen K_v, K_f werden schließlich aus den zu schätzenden Parametern (Rotationswinkeln, Brennweite, etc.) gebildet.

2.3 Bildfusion

In der Literatur sind eine Vielzahl von Verfahren und Algorithmen zur pixelbasierten Datenfusion zu finden. Für die beschriebene Fußgängererkennung wurde aufgrund der relativ einfachen Implementierung und dennoch effizienten Algorithmen die Fusion über die so genannte Multiskalenzerlegung gewählt. Dabei werden die Eingangsbilder mittels einer geeigneten Transformation Ψ in Komponenten unterschiedlicher räumlicher Auflösungen zerlegt. Diese Komponenten werden dann separat analysiert und gemäß bestimmter Regeln ϕ zu einer gemeinsamen Multiskalenrepräsentation fusioniert. Eine entsprechend umgekehrte Transformation Ψ^{-1} erzeugt daraus das fusionierte Bild.

$$I_{fus} = \Psi^{-1}\left(\phi\left(\Psi(I_{fir}), \Psi(I_{vis})\right)\right) \tag{3}$$

Das verwendete Fusionschema [PIE02] für die pixelbasierte Fusion ist im folgenden Bild dargestellt:

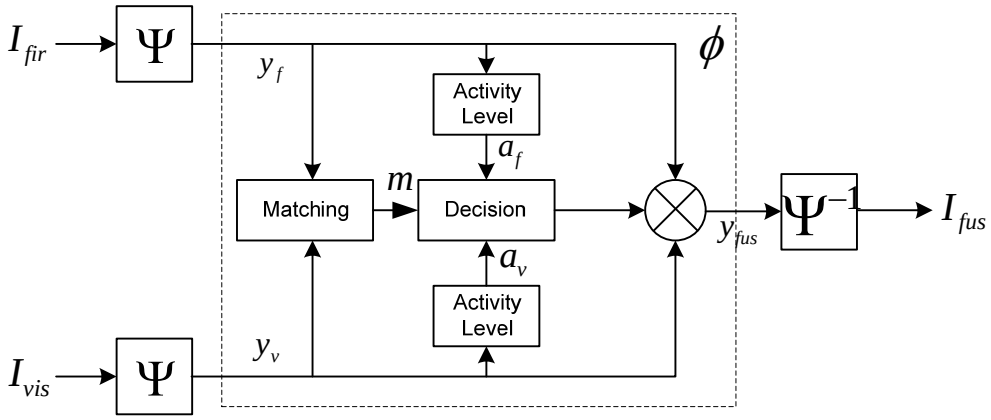


Abbildung 3: Fusionsschema für die pixelbasierte Fusion

Für die Zerlegung der Eingangsbilder in die Multiskalenrepräsentation wird die Diskrete Wavelet Transformation (DWT) verwendet. Die Multiskalendarstellung besteht aus den hochfrequenten Anteilen (Detail-Bildern) und den niederfrequenten Informationen (Approximationsbild). Die Zerlegung erfolgt nach der rekursiven Methode von Mallat [MAL89], die sich effektiv mit einer digitalen Filterbank implementieren lässt. Durch Anwendung orthogonaler Tief- und Hochpassfilter wird dabei das Bild zunächst zeilen- und anschließend spaltenweise gefiltert. Dadurch

entstehen pro Skale jeweils 3 Detail-Bilder (Sub-Bänder) mit Kanteninformationen in horizontaler, vertikaler und diagonaler Richtung, sowie ein Approximationsbild, was als Ausgangsbild für die nächste Zerlegung dient.

Die Wavelet-Koeffizienten y in den einzelnen Sub-Bändern der gefilterten Eingangsbilder werden nun über geeignete Fusionsregeln zu einer Multiskalen-Repräsentation kombiniert. Dabei dient der *Activity-Level* a als ein Maß zur Beurteilung der Qualität eines bestimmten Teils eines jeden Eingangsbildes. Er ist der Grad eines jeden Koeffizienten, wie relevant er für die aktuelle Aufgabe ist. Der *Activity-Level* berechnet sich an der Position **x** wie folgt

$$a(\mathbf{x}) = \sum_{\Delta\mathbf{x}\in W} |y(\mathbf{x}+\Delta\mathbf{x})|^2 \tag{4}$$

Um die Ähnlichkeit zwischen den Eingangsquellen ausdrücken zu können, ist die Definition eines so genannten *Matchwertes* sinnvoll. Abhängig von diesem *Matchwert* werden die Eingangsbilder miteinander kombiniert. Ein kleiner *Matchwert* bedeutet zum Beispiel eine geringe Ähnlichkeit der Bilder an dieser Stelle.

Den Kern des Kombinationsalgorithmus bildet der *Decision-Block*, welcher die tatsächliche Kombination der Koeffizienten der Eingangsbilder in Abhängigkeit vom *Matchwert* und *Activity-Level* regelt. Das bedeutet also, an Stellen, wo beide Bilder generell unterschiedlich sind (kleiner *Matchwert*), selektiert der Kombinationsprozess den Koeffizienten mit dem größeren *Activity-Level*, was einem höheren Informationsgehalt (z.B. Kanteninformation) entspricht. Hingegen wird das Ergebnis an Stellen, wo die Bilder sich ähnlich sind, aus den Eingangsquellen gemittelt.

$$y_{fus}(\mathbf{x}) = \begin{cases} y_f(\mathbf{x}) & ,m(\mathbf{x}) \le T \; und \; a_f > a_v \\ y_v(\mathbf{x}) & ,m(\mathbf{x}) \le T \; und \; a_f \le a_v \\ \dfrac{y_f(\mathbf{x}) + y_v(\mathbf{x})}{2} & ,sonst \end{cases} \tag{5}$$

Schließlich entsteht das fusionierte Bild durch Anwendung der Inversen Diskreten Wavelet-Transformation (Multiskalensynthese).

3 Detektion von Objekten in Bildern

3.1 Anforderungen an Detektionsalgorithmen im Fahrzeugumfeld

Die Objektdetektion und Objektklassifikation aus Kamerabildern bilden einen Schwerpunkt in der automatischen Bildverarbeitung. Insbesondere im Fahrzeugumfeld stellen sie eine große Herausforderung dar, da sich sowohl die Objekte (wie z.B. Fußgänger oder Fahrzeuge) als auch die Kamera bewegen können. Weiterhin muss ein Objekterkennungssystem echtzeitfähig sein, um innerhalb eines definierten Zeitfensters (für Kameras in der Regel 33-40ms) Detektionen liefern zu können. Zusätzlich sollte das System zum einen verlässlich arbeiten, möglichst viele Objekte korrekt erkennen und möglichst wenig Fehldetektionen liefern. Zum anderen muss es sich selbständig an Veränderungen in der Umwelt, insbesondere an Helligkeitsänderungen und veränderte Wetterbedingungen, anpassen können.

3.2 Prozesskette

Da auf Grund der Bewegung von Kamera und Objekten die meisten Standard-Erkennungsverfahren für statische Kameras (wie z.B. Hintergrundextraktion) nicht anwendbar sind, hat sich die folgende Vorgehensweise zur robusten Objektdetektion und -verfolgung herausgebildet.

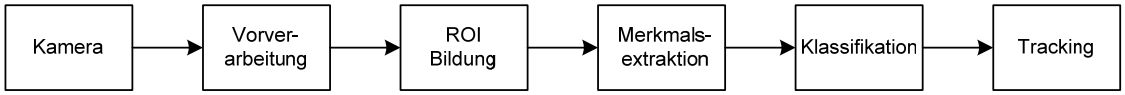


Abbildung 4: Prozesskette Objektdetektion

Im ersten Schritt wird das Kamerabild so vorverarbeitet, dass relevante Informationen für die Bildanalyse verstärkt werden und irrelevante möglichst verschwinden. Zur Anwendung kommen häufig Verfahren wie Rauschunterdrückung, Kontrastverstärkung oder Segmentierung. Im zweiten Schritt werden eine oder mehrere Regionen im Bild festgelegt, die sogenannten *Regions of Interest* (ROIs), in der die gesuchten Objekte überhaupt auftreten können. Alle folgenden Algorithmen werden nur auf die eingeteilten Regionen angewendet. Durch diesen Schritt können Detektionsaufgaben deutlich beschleunigt werden, da häufig nicht das gesamte Bild, sondern nur ein Teil davon durchsucht werden muss. Im Schritt *Merkmalsextraktion* werden signifikante Merkmale einer ROI berechnet, was zum einen Pixelwerte, aber auch Informationen über Kanten im Bild oder aber Histogramme sein können. Diese bilden im folgenden Schritt *Klassifikation* den Input eines Klassifikationsalgorithmus, der letztendlich entscheidet, ob die vorliegende ROI eines der zu detektierenden Objekte beinhaltet oder nicht. Im letzten Schritt *Tracking* werden die gefundenen Objekte über einen längeren Zeitraum beobachtet und Schätzungen über Position, Bewegung und weiterer Eigenschaften vorgenommen.

3.3 Objektdetektion am Beispiel der Fußgängererkennung

Am Beispiel eines Fußgängerschutzsystems, das im EU-Projekt WATCH-OVER [AND07] entwickelt wurde, soll in diesem Abschnitt die Detektion von Objekten in Bildern erläutert werden. Bei diesem System wurde eine Kaskade aus AdaBoost-Klassifikatoren [FS95] mit einer entsprechend großen Menge an Fußgänger- und Nicht-Fußgängerbildern trainiert. Als Basis-Klassifikatoren dienten Entscheidungsbäume. Bei diesem von Viola & Jones [VJ01] beschriebenen Ansatz werden im Bild überlappende ROIs in unterschiedlicher Größe, aber mit konstantem Seitenverhältnis definiert. Aus jeder ROI wird der Merkmalsvektor erzeugt. Im Gegensatz zu [VJ01], wo mit Haar-Wavelet Merkmalen gearbeitet wird, verwendet das vorgestellte System die *Histograms of Oriented Gradients* (HOGs), die von D.G.Lowe in [LO04] erstmalig beschrieben wurden. Sie haben den Vorteil, dass sie unempfindlich gegenüber Beleuchtungsschwankungen, aber auch relativ tolerant gegenüber kleineren Verschiebungen sind. Einerseits durch die Verschiebungsinvarianz der Merkmale, die es ermöglicht, die ROIs in einem deutlich vergröberten Raster einzuteilen, und andererseits durch den Einsatz von Integralbildern zur schnellen HOG-Berechnung [SCH08], konnte Echtzeitverarbeitungsfähigkeit in der Fußgängererkennung erreicht werden. Allerdings führt die Verschiebungstoleranz häufig zu mehrfachen Detektionen für ein und dasselbe Objekt (siehe Abbildung 5), woraufhin ein ROI Clustering durchgeführt wird. Dabei werden mehrere eng beieinanderliegende ROIs einem Objekt zugeordnet. Die Prozesskette wird mit einem Kalman-Filter basierten Tracking System [FAR08] abgeschlossen.

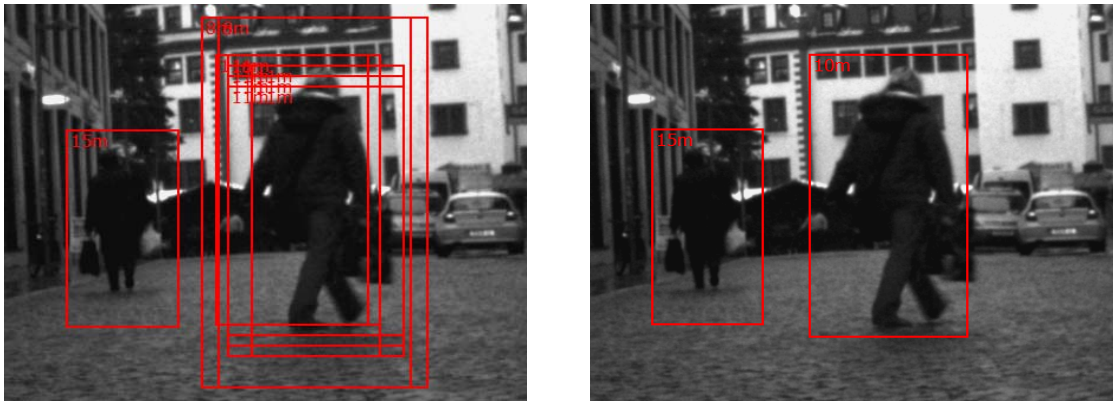


Abbildung 5: a) Mehrfache Detektionen b) Mittels Clustering fusionierte Detektionen

4 Einsatz der Pixelbasierten Datenfusion in der Fußgängererkennung

4.1 Motivation

Wie bereits eingangs erwähnt, gewinnt die automatische Erkennung von Personen und Hindernissen unter anderem im Hinblick der Reduzierung der Verkehrsunfälle zunehmend an Bedeutung. Ein solches System erfordert eine große Bandbreite der Sensoren, um in möglichst vielen Umweltbedingungen einwandfrei zu funktionieren. Eine gewöhnliche Kamera, basierend auf dem sichtbaren Spektrum, spielt hierbei sicherlich aufgrund ihrer gewohnten Abbildung bei der Fußgängererkennung eine große Rolle. Aufgrund fallender Preise werden jedoch auch Infrarot-Kameras (FIR) für die Umfelderkennung im Fahrzeug interessant. Sie haben den Vorteil, dass sich der für gewöhnlich wärmere Mensch gegenüber seinem Hintergrund abhebt und somit eine Detektion im Bild erleichtert. Jedoch kann die FIR-Kamera an sonnigen und heißen Tagen bezüglich ihrer Detektionsleistung versagen [BER09]. Die Kombination des FIR-Bildes mit dem Grauwertbild einer visuellen Kamera soll ein Mischbild schaffen, welches aufgrund eines höheren Informationsgehaltes hinsichtlich der Anwendung die Extraktion aussagekräftiger Features fördert und somit eine robuste Erkennung von Fußgängern ermöglicht.

4.2 Umsetzung

Das unter 3.3 beschriebene Fußgängererkennungssystem wurde schließlich um die beiden Prozessschritte Rektifikation und Bildfusion erweitert. Damit lässt sich das System grob die folgenden Bestandteile gliedern:

Rektifikation:
Die Rohbilder der FIR- und der visuellen Kamera werden zunächst mit den Transformationsmatrizen rektifiziert. Die Transformationsvorschrift wurde gemäß der in Abschnitt 2.1 beschriebenen Methode zur Rektifikation offline ermittelt. Um starke perspektivische Verzerrungen zu vermeiden, wurden beide Sensoren so nah wie möglich zueinander am Fahrzeug montiert.

Bildfusion:
Nachfolgend werden die rektifizierten Frames basierend auf einer Multiskalenzerlegung fusioniert. In einem Vorverarbeitungsschritt wird der Kontrast der Eingangsbilder erhöht und das FIR-Bild zusätzlich invertiert.

Feature-Extraktion:
Das gewonnene Mischbild wird nun mittels der Rechteckmasken (ROIs) in verschiedenen Positionen und Größen überlagert. Innerhalb der ROIs werden die Merkmale (HOGs) aus dem Bild extrahiert. Dabei werden die ROIs nur in den relevanten Bereichen des Bildes angewendet.

Klassifikation:
Der Merkmalsvektor einer jeden ROI wird einer Kaskade von Klassifikatoren (AdaBoost) basierend auf binären Entscheidungsbäumen zugeführt. Ergebnis ist schließlich die Aussage, ob sich im entsprechenden ROI ein Fußgänger befindet oder nicht.

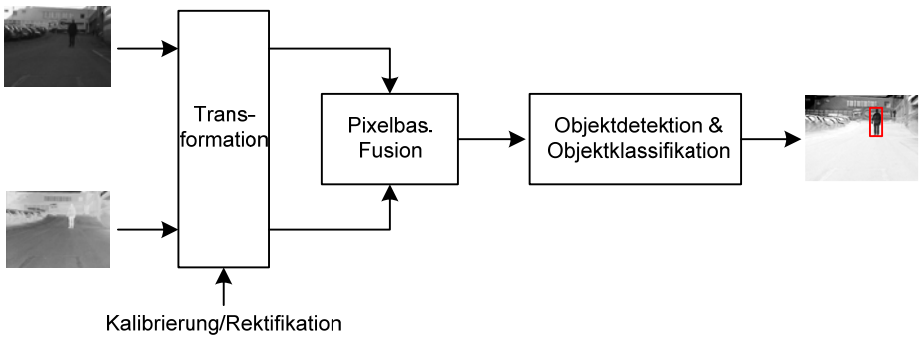


Abbildung 6: System zur Fußgängererkennung

4.3 Ergebnisse

Die Fußgängererkennung wurde an zwei Sequenzen mit 840 bzw. 1000 Frames getestet, die im Winter in der Chemnitzer Innenstadt am bewegten Fahrzeug aufgezeichnet wurden. Beide Szenen beinhalten typische Stadtszenen mit Fußgängern, Fahrzeugen und Gebäuden.

Dabei wurden die Klassifikationsergebnisse, welche mit den fusionierten Streams gewonnen wurden, mit den Ergebnissen bei Verwendung nur eines einzelnen Sensors verglichen. Als Vergleichskriterium dienten zum einen die **Detektionsrate** und zum anderen die **Falschalarmrate**. Die Detektionsrate gibt an, wie viele der möglichen Fußgänger erkannt wurden. Die Falschalarmrate gibt an, wie viele Fußgänger fehlerhaft erkannt wurden.

Die Ergebnisse aus beiden Sequenzen sind in den folgenden Tabellen abgetragen:

TS01		FIR	Visuell	Fusioniert
	Anzahl Frames	842		
	ROIs pro Frame	5763		
	Anzahl Fußgänger	1450		
	Detektionsrate	0.61	0.26	0.47
	Falschalarmrate	0.000106	0.000079	0.000031

Tabelle 1: Ergebnisse aus Testsequenz 1

TS02		FIR	Visuell	Fusioniert
	Anzahl Frames	999		
	ROIs pro Frame	5763		
	Anzahl Fußgänger	1980		
	Detektionsrate	0.67	0.29	0.63
	Falschalarmrate	0.000110	0.000065	0.000058

Tabelle 2: Ergebnisse aus Testsequenz 2

Aus den Ergebnissen in Tabelle 1 und 2 ist ersichtlich, dass der fusionierte Stream die geringste Falschalarmrate produziert. Gegenüber dem FIR-Stream, wo in mehr als jedem zweiten Bild eine Fehldetektion erfolgt, konnte der Anzahl der Falschalarme mehr als halbiert werden. Allerdings ist die Detektionsrate geringfügig geringer, was jedoch durch ein nachfolgendes Trackingverfahren kompensiert werden kann. Die hohe Detektionsrate in den FIR-Bildern resultiert aus der kühlen Witterung während der Aufnahme, wodurch sich insbesondere Personen sehr gut vom kühleren Hintergrund abheben.

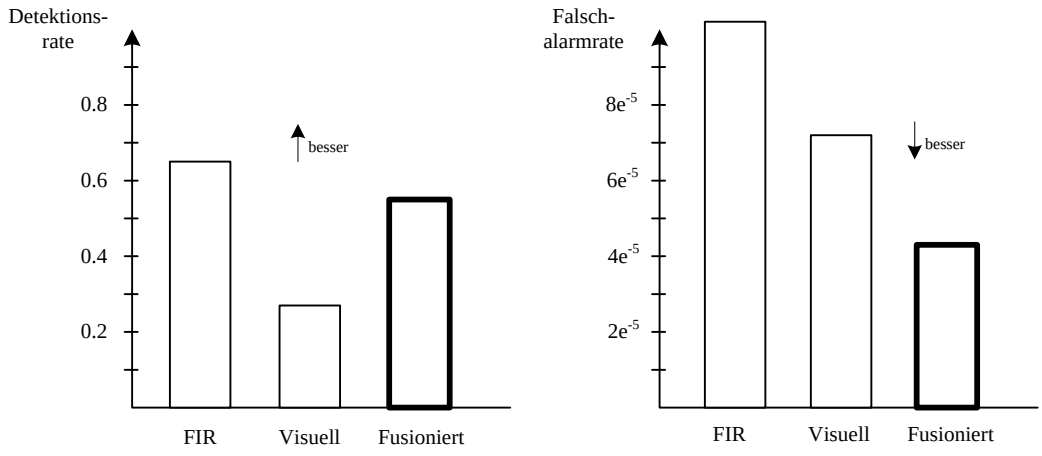


Abbildung 7: Gemittelte Ergebnisse aus beiden Testsequenzen

5 Literaturverzeichnis

[AND07] Andreone, L., Wanielik, G., Vulnerable Road Users Thoroughly Addressed in Accident Prevention: The WATCH-OVER European Project. Beijing: ITS World Congress (Poster), 2007.

[BER09] Bertozzi, M., Broggi A., Felisa, M., Ghidoni, S., Grisleri P., Vezzoni, G., Gómez, C. H., Del Rose, M., Multi Stereo- Based Pedestrian Detection by Daylight and Far-Infrared Cameras, Augmented Vision Perception in Infrared, Advances in Pattern Recognition, p. 371-401, London: Springer-Verlag, 2009

[BLU06] Blum, R. S., Xue, Z., Zhang, Z., An Overview of Image Fusion, Multi-Sensor Image Fusion and Its Applications, Taylor & Francis Group, 2006

[DAU92] Daubechies, I., Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992

[FAR08] Fardi, B., Neubert, U., Giesecke, N., Lietz, H., Wanielik, G., A Fusion Concept of Video and Communication Data for VRU Recognition. In: Proceedings of the 11th International Conference on Information Fusion, 2008.

[FS95] Freund, Y., Schapire, R. E., A Decision-Theoretic Generalization of Online Learning and an Application to Boosting. In: Proceedings of the European Conference on Computational Learning Theory, p. 23-37, 1995.

[FUS06] Fusiello A., Irsara L. Quasi-euclidean uncalibrated epipolar rectification. Research Report RR 43/2006, Dipartimento di Informatica - Università di Verona, 2006.

[GOS05] Goshtasby, A., 2-D and 3-D image registration for medical, remote sensing and industrial applications. John Wiley & Sons, Inc., Hoboken, New Jersey, 2005

[HAR03] Hartley, R., Zisserman, A., Multiple View Geometry in computer vision. Second Edition. Cambridge University Press, 2003

[LO04] Lowe, David G., Distinctive image features from scale-invariant keypoints. In: International Journal of Computer Vision, Vol. 60, No. 2, Springer Verlag, S. 91-110, 2004.

[MAL89] Mallat, S. G., A theory for multiresolution signal decomposition: The wavelet representation, IEEE Trans. Pattern Anal. Machine Intell., 11, 674-693, 1989

[PIE02] Piella G. A general framework for multiresolution image fusion: from pixel to regions, Information Fusion, Vol 4, pp. 259-280, 2003

[SCH08] Schloßhauer, J., Giesecke, N., Fardi, B., Wanielik, G., Fast Implementation of a robust Pedestrian Recognition System In: Proceedings of the 2008 IEEE International Conference on Vehicular Electronics and Safety, 2008.

[VJ01] Viola, P., Jones, M., Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1, S.511, 2001.

Grundlagen für das Retrieval rotationssymmetrischer Gefäße

Stefan Wagner, Christian Hörr, David Brunner und Guido Brunnett

Technische Universität Chemnitz
Fakultät für Informatik
Professur Graphische Datenverarbeitung und Visualisierung
{stefan.wagner,hoerr,brunner,brunnett}@informatik.tu-chemnitz.de

Zusammenfassung: In der wissenschaftlichen Erforschung archäologischer Sachverhalte spielt die präzise Dokumentation der Fundstücke eine äußerst wichtige Rolle. Einer rapide steigenden Anzahl von Fundstücken ist allerdings mit den traditionellen Methoden der Archäologie kaum noch Herr zu werden. Moderne Retrievalsysteme stellen dabei eine Möglichkeit dar, Funddaten über einen längeren Zeitraum systematisch zu speichern und bieten dem Nutzer gleichzeitig passende Werkzeuge für die Abfrage derselben.
Im vorliegenden Beitrag wird der Entwurf eines solches Systems für rotationssymmetrische Keramikgefäße vorgestellt und einzelne Aspekte näher beleuchtet. Insbesondere die automatische Extraktion und Analyse spezifischer Eigenschaften der Gefäße sind dabei wichtig. Sie werden hauptsächlich aus der Analyse der Profillinie gewonnen, deren wichtigste Voraussetzung die Ausrichtung des Gefäßes an seiner Rotationsachse darstellt.

Schlagwörter: Rotationsachse, Keramik, Retrieval, Automatische Extraktion

1 Einleitung

Eine präzise und aussagekräftige Dokumentation von Fundstücken ist eine der Grundlagen für jegliche archäologische Forschung, da sie wichtige Informationen über Eigenschaften und den Fundkontext enthält. Da die Erzeugung dieser Dokumentation insbesondere für mehrere Hunderttausend Fundstücke mit den herkömmlichen Methoden der Archäologie praktisch nicht zu bewältigen ist, gab es in den letzten Jahren viele Projekte, die sich mit der Frage der Automatisierung dieses Vorgangs beschäftigt haben. Dazu gehört auch die an der Technischen Universität Chemnitz in Zusammenarbeit mit dem Sächsischen Landesamt für Archäologie entwickelte Dokumentationssoftware *TroveSketch*, womit Keramikgefäße automatisch untersucht werden können. Dafür wird das

Gefäß zunächst mit einem 3D-Scanner erfasst und die relevanten Eigenschaften anschließend aus dem digitalen Modell desselben abgeleitet. Im wesentlichen wird damit die Erzeugung von Profilskizzen, stilisierten Abbildungen und Maßangaben unterstützt, welche für diese Form von Fundstücken einen wesentlichen Bestandteil der Dokumentation darstellen.

Mit dem Einsatz moderner 3D-Scanner können digitale Kopien der Fundstücke angefertigt werden, die sowohl geometrische (Form) als auch Oberflächeneigenschaften (z.B. Farbe) enthalten. Die damit einhergehende Virtualisierung ermöglicht die Verarbeitung der Fundstück unabhängig von Fund- und Lagerort und schont damit die Originale. Mit einer stetig steigenden Anzahl von Fundstücken stellt sich allerdings sehr schnell die Frage, wie die anfallenden Daten so aufbereitet und gespeichert werden sollen, dass die Wissenschaftler umfassend und schnell darauf zugreifen können.

Retrievalsysteme können dafür zweckdienlich sein, wenn sie entsprechende Funktionalität für Langzeitspeicherung, Verstichwortung (bzw. Kategorisierung) und praxisnahe Suche anbieten. Der interaktive, d.h. vom Nutzer wahrnehmbare Teil des Systems ist dabei die Nutzerschnittstelle, welche verschiedene Werkzeuge zur Verfügung stellen sollte, die diesem in geeigneter Weise helfen. Im Hinblick auf eine Zusammenarbeit vieler verschiedener Forschungseinrichtungen in vielen verschiedenen Länder der Welt ist es darüberhinaus erstrebenswert, wenn das System eine webbasierte Nutzerschnittstelle besitzt. Wir haben es also mit einem System zu tun, das potenziell verschiedene Aspekte der Informatik abdeckt, darunter Datenbanken, Shape Matching dreidimensionaler Objekte, Algorithmen für die Textsuche sowie Ansätze für die automatische Kategorisierung von Fundstücken mithilfe von Methoden der Künstlichen Intelligenz.

Wenngleich sich unsere Forschungstätigkeit auf all diese Bereiche erstreckt, wollen wir uns im vorliegenden Bericht dennoch auf zwei interessante Aspekte beschränken. Einerseits richten wir den Blick auf Keramikgefäße, wie sie in Gräbern aus der Lausitzer Kultur zu finden sind. Dies geschieht insbesondere auch, da solche Gefäße meist rotationssymmetrisch und daher aus geometrischer Sicht noch relativ leicht zu untersuchen sind. Andererseits stellen wir einen Ansatz für eine Referenzsuche vor, bei welcher der Nutzer als Suchanfrage ein digitalisiertes Gefäß eingibt und eine Liste der diesem ähnlichsten Gefäße zurück bekommt. Die bei dieser Suche notwendigen Schritte basieren im Wesentlichen auf der automatischen Extraktion bestimmter Eigenschaften, wie sie bereits in *TroveSketch* zum Einsatz kommen.

Im Folgenden ein kurzer Überblick über den Aufbau dieses Berichts. Im zweiten Abschnitt werden vorhergehende Arbeiten vorgestellt, welche sich mit dem *Information Retrieval* und insbesondere der Suche nach Modellen dreidimensionaler Objekte beschäftigten. Es zeigt sich, dass die Information, nach der spätere Suchen möglich sein sollen, zunächst strukturiert werden muss, auch und insbesondere für dreidimensionale Objekte. Dies geschieht in Form von Deskriptoren. Die Anforderung an solche Deskriptoren für rotationssymmetrische Keramikgefäße werden im Abschnitt 3 diskutiert. Im Abschnitt 4 erläutern wir, wie die Gefäßeigenschaften mittels automatischer Extraktion ermittelt werden und geben im letzten Abschnitt einen kurzen Ausblick auf zukünftig geplante Arbeiten.

2 Grundlagen und vorhergehende Arbeiten

Das *Information Retrieval* ist ein wichtiger Bestandteil innerhalb der IT-Strukturen in Wirtschaft und Verwaltung geworden. Im Unterschied zur herkömmlichen Datenabfrage, z.B. über entsprechende Schnittstellen in Datenbanksystemen, spielt beim Information Retrieval meist ein bestimmter Anteil an Unsicherheit eine Rolle. Im Bezug auf Text-Retrievalsysteme trifft das unter anderem darauf zu, wie der Nutzer seine Anfrage formuliert oder wie das System die Anfrage interpretiert. Eng verbunden mit solchen Formen der Unsicherheit ist die Frage nach der Relevanz der vom System erzeugten Antworten. Da der Nutzer hierbei stets selbst in der Pflicht steht zu definieren, was für ihn eigentlich relevant ist, handelt es sich bei der Beantwortung der Frage nach der Relevanz meistens um eine eher subjektive Einschätzung.

Dieses Problem wird sogar noch komplizierter, wenn es sich nicht um Texte, sondern um andere Medienarten wie Bilder, Musik oder eben dreidimensionale Modelle handelt. Die natürliche Ausdrucksform menschlicher Gedanken ist die Sprache, welche in schriftlicher Form wiedergegeben werden kann und das reine Text-Retrieval damit größtenteils einfach gestaltet. Bei anderen Medienarten fehlt dieser enge Kontakt zum menschlichen Denken dagegen. Folgerichtig müssen andere, also sprachunabhängige Ansätze für die Repräsentation solcher Objekte im Rechner gefunden werden, da textuelle Darstellungen nicht immer verfügbar bzw. mit vertretbarem Aufwand herzustellen sind.

Bei dreidimensionalen Modellen bieten sich deren geometrische Eigenschaften an, die im Mittelpunkt des sogenannten *Shape Matchings* stehen. Dabei werden die geometrischen Eigenschaften benutzt, ein Objekt mehr oder weniger abstrakt darzustellen. Eine solche Darstellung heißt *Deskriptor*. Darüberhinaus ist ein Ähnlichkeitsmaß notwendig,

welches zwei Deskriptorwerten eine reelle Zahl (i.A. im Intervall $[0, 1]$) zuordnet, die beider Ähnlichkeit zueinander ausdrückt.

Nach Hörr [Hör05] können an einen idealen Deskriptor eine Reihe von Anforderungen gestellt werden. Zunächst sollte er selektiv sein, d.h. die durch das Ähnlichkeitsmaß ausgedrückte Ähnlichkeit muss der tatsächlichen Ähnlichkeit zweier Objekte entsprechen. Weiterhin sollte ein Deskriptor eindeutig sein, sodass ein bestimmter Deskriptorwert immer nur genau ein bestimmtes Objekt beschreibt. Eine sehr wichtige Anforderung insbesondere für den Fall dreidimensionaler Modelle ist, dass der Deskriptor transformationsinvariant ist, seine Werte also nicht von der Position, der räumlichen Ausrichtung oder auch der Größe des Objektes abhängen. Letztere Invarianz kann aber auch verworfen werden. Gegenüber Rauschen in der digitalen Repräsentation der Objekte (z.B. als Dreiecksnetze) sollte der Deskriptor robust genug sein. Die Berechnung der Deskriptorwerte sowie der Ähnlichkeiten sollte außerdem effizient ablaufen.

In den meisten Fällen wird aber kein Deskriptor alle diese Anforderungen gleichzeitig erfüllen können. Shilane et al. [SMKF04] haben daher 12 verschiedene Deskriptoren hinsichtlich ihrer Eignung getestet, dreidimensionale Objekte zu klassifizieren. Dabei hat sich keiner als universell einsetzbar erwiesen. Allerdings eignen sich einzelne Deskriptoren sehr gut für spezifische Anwendungen, beispielsweise um natürliche von künstlichen Objekten zu unterscheiden. Eine Konsequenz daraus ist, dass Deskriptoren meistens sehr spezifisch und daher auf einen kleinen Anwendungsbereich zugeschnitten sind. In klar umrissenen Projekten, wie der Verarbeitung rotationssymmetrischer Gefäße, ist das aber kein großes Problem. Bustos et al. [BKS+05] haben eine Reihe bekannter Deskriptoren aufgelistet und eine mögliche Taxonomie für diese vorgeschlagen.

Verschiedene Anfragestrategien für dreidimensionale Modelle wurden von Min et al. untersucht [MKF04]. Ihnen zufolge ist eine ausschließlich textgestützte Suche erfolgreich, wenn die Objekte adäquat annotiert sind. In Disziplinen wie der Archäologie mit mehreren Hunderttausend Fundstücken ist dies allerdings nicht realisierbar. Folglich sind anspruchsvollere Ansätze wünschenswert, beispielsweise die Möglichkeit, das gesuchte Objekt sehr grob zu skizzieren, wobei allerdings auch das Talent des Suchenden über den Erfolg der Suche entscheiden kann.

Ein Retrievalsystem besteht im Allgemeinen aus den in Abbildung 1 dargestellten Komponenten, wobei nur die Nutzerschnittstelle für den Nutzer des Systems sichtbar ist. Falls diese über das Internet bereitgestellt wird, besteht sie normalerweise aus einer Menge von Webseiten, welche verschiedene Suchmasken und zusätzliche Features anbieten, z.B. in

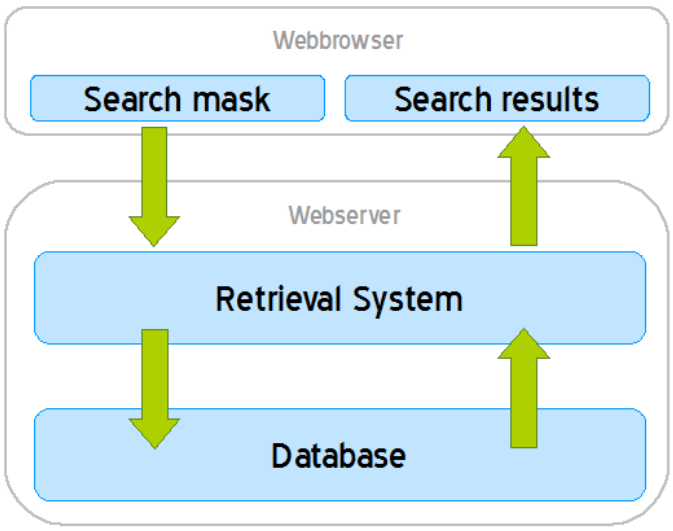


Abbildung 1: Der allgemeine Aufbau eines Retrievalsystems.

der Archäologie die Möglichkeit des Downloads von Modellen oder der zugehörigen Dokumentationen. Die Nutzereingaben werden an die zweite Schicht geleitet, welche die Logik für die Verarbeitung von Anfragen und für Erzeugung von Ergebnisseiten bereitstellt. Die verarbeiteten Daten werden in der dritten Schicht gespeichert, wobei es sich für gewöhnlich um ein Datenbanksystem handelt.

In diesem Bericht wollen wir die Möglichkeit untersuchen, ein dreidimensionales Referenzmodell eines Gefäßes als Suchanfrage zu übergeben. Nach automatischer Analyse dieses Referenzmodells kann es auf Grundlage der so gewonnenen Daten mit den in der Datenbank vorhandenen Objekten verglichen werden, um eine Liste der ähnlichsten Objekte zu erzeugen. Wie bereits erwähnt muss dazu eine Menge sinnvoller Eigenschaften definiert werden, damit diese Suchstrategie tatsächlich funktionieren kann.

3 Formbeschreibung

In diesem Abschnitt beschreiben wir einen Ansatz, wie Gefäße hinsichtlich spezifischer Eigenschaften so beschrieben werden können, dass man sie sehr gut voneinander unterscheiden kann. Wir nehmen dabei Bezug auf Hörr et al. [HB08], welche sich auf die Tatsache stützen, dass die meisten Gefäße Rotationskörper sind, die ggf. noch Anhängsel wie Henkel, Füße oder Ausgüsse besitzen. Die gesamte Form der Gefäße wird aber im Wesentlichen von der Profillinie bzw. ihrer Rotation um die Rotationsachse bestimmt.

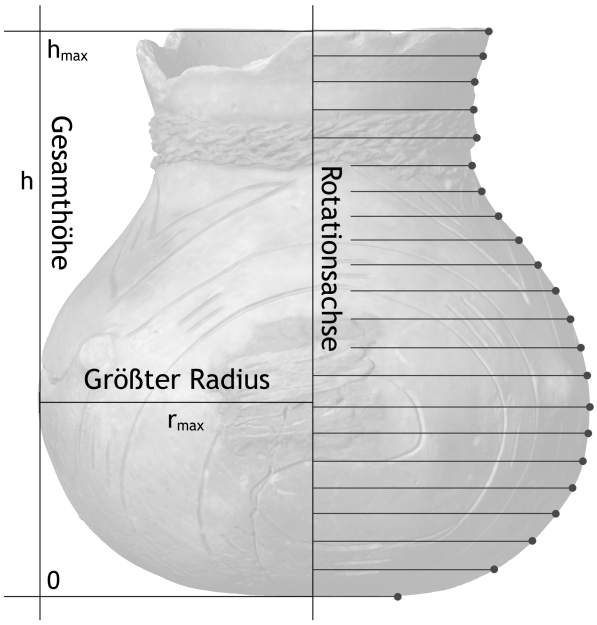


Abbildung 2: Beispiel für die Profillinie einer Flasche.

Bei der Untersuchung dieser Profillinie können charakteristische Eigenschaften des Gefäßes extrahiert werden, die entweder nominal sind oder ein Verhältnis beschreiben. Während mit nominalen Eigenschaften wie der Höhe (h_{max}) oder dem größten Radius (r_{max}) größere von kleineren Objekten unterschieden werden können, eignen sich Verhältnisse besser zur Unterscheidung von Objekten hinsichtlich ihrer Form. Als Beispiel diene folgendes Verhältnis, welches Hauptindex genannt wird.

$$\text{Hauptindex} = \frac{2r_{max}}{h_{max}}. \tag{1}$$

Falls der Hauptindex kleiner als eins ist, dann ist das Gefäß länglich, während ein Wert größer als eins ein flaches Objekt beschreibt. Wir abstrahieren die Form eines Gefäßes, indem wir seine Profillinie als Funktion des Radius r über dem Intervall der Höhe des Gefäßes $[0, h_{max}]$ beschreiben, die für jeden Punkt auf der Profillinie dessen Abstand zur Rotationsachse angibt.

$$r : [0, h_{max}] \rightarrow \mathbb{R}_0^+. \tag{2}$$

Der bereits erwähnte maximale Radius r_{max} ist einer der Werte dieser Funktion. Wir normalisieren die Funktion r, indem wir alle Werte durch diesen Wert teilen.

$$r_{norm} : [0, h_{max}] \rightarrow [0, 1]\,. \tag{3}$$

Weiterhin wird der Definitionsbereich von r durch die Unterteilung in n paarweise disjunkte Teile gleicher Länge diskretisiert. Damit können wir die Form eines Gefäßes als Folge von n reellen Zahlen im Intervall $[0, 1]$ beschreiben. Der zugehörige Deskriptor kann dann wie folgt dargestellt werden.

$$d : [1, n]_{\mathbb{N}} \rightarrow [0, 1]_{\mathbb{R}} \, . \tag{4}$$

Mit diesem Ansatz haben wir auf einfache Weise einen bedeutungsvollen und insbesondere auch skalierungsinvarianten Deskriptor für die Gefäßform gefunden. Weiterhin ist noch ein Ähnlichkeitsmaß nötig, damit Gefäße miteinander verglichen werden können. Der Einfachheit halber verwenden wir die folgende Metrik zur Bestimmung der Ähnlichkeit zweier normalisierter Radien $d_1(i)$ und $d_2(i)$ auf derselben Höhe $i \in [1, n]$.

$$sim_i(d_1, d_2) = \begin{cases} \frac{d_1(i)}{d_2(i)} & \text{if } d_1(i) < d_2(i) \\ \frac{d_2(i)}{d_1(i)} & \text{if } d_2(i) < d_1(i) \\ 1 & \text{sonst} \end{cases} . \tag{5}$$

Da wir genau n solcher Wertpaare vergleichen müssen, addieren wir deren Ähnlichkeiten und dividieren das Ergebnis durch n. Wir haben damit wiederum auf einfache Weise ein Ähnlichkeitsmaß für zwei Profillinien definiert.

$$Sim(d_1, d_2) = \frac{1}{n} \sum_{i=1}^{n} sim_i(d_1, d_2). \tag{6}$$

Zusätzlich zu diesem Profilliniendeskriptor kann der Nutzer weitere geometrische Eigenschaften definieren, welche auf gleiche Weise im Retrievalvorgang berücksichtigt werden sollen, indem ihr Einfluss auf das Ergebnis vom Nutzer explizit festgelegt wird (siehe Abbildung 3).

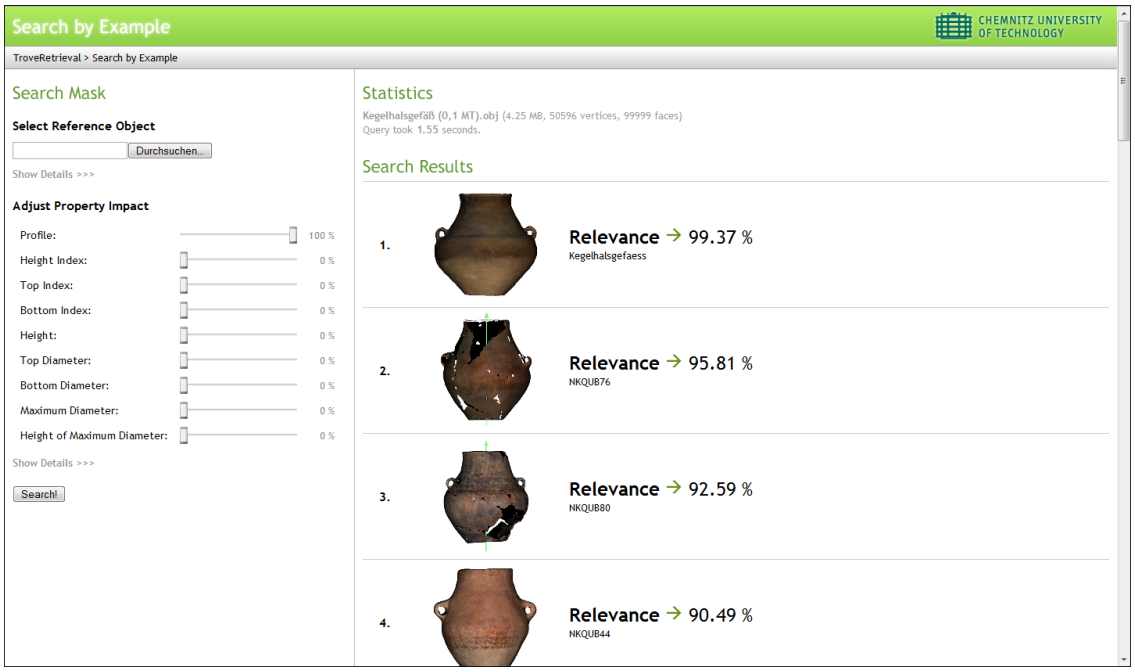


Abbildung 3: Screenshot der Suchmaschine für die Suche nach Keramikgefäßen

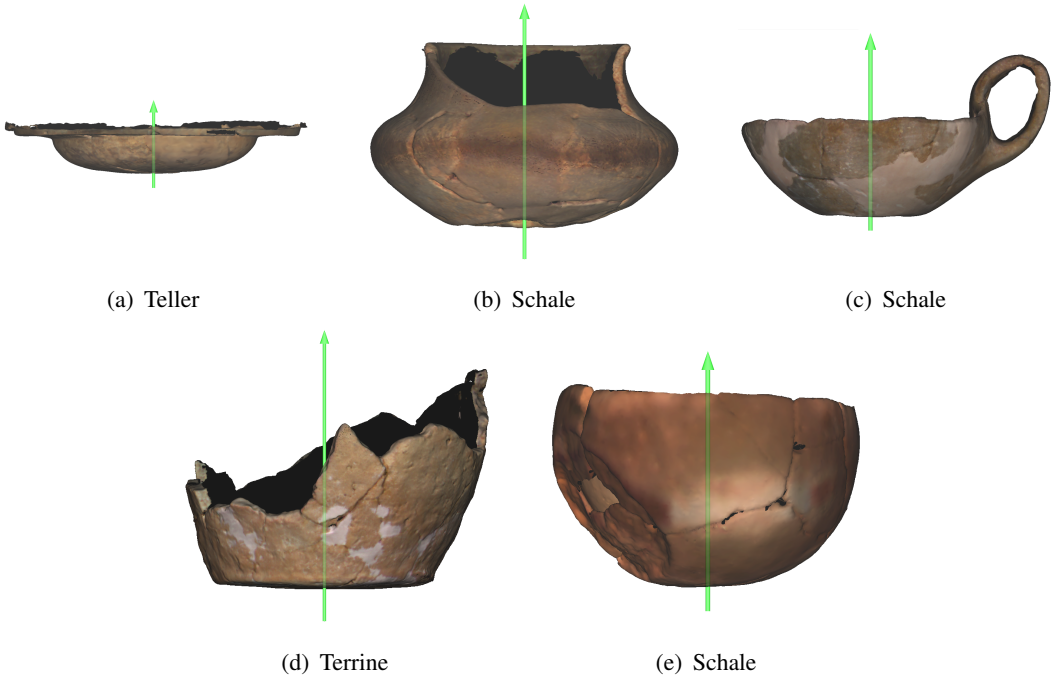
4 Automatische Analyse

In diesem Abschnitt skizzieren wir kurz, wie das vom Nutzer übergebene Referenzmodell automatisch analysiert wird, um mit den so gewonnenen Eigenschaften die Suche durchzuführen. Zuerst muss das Gefäß an seiner Rotationsachse ausgerichtet werden, da nicht davon ausgegangen werden kann, dass dies bereits der Fall ist. Die Kenntnis der Rotationsachse ist eine grundlegende Voraussetzung für alle weiteren Schritte. Daran anschließend können die Profillinie bestimmt und entsprechende Eigenschaften abgeleitet werden.

Rotationsachse Die Berechnung der Rotationsachse eines Objektes, von welchem weder Position noch Ausrichtung bekannt sind, ist nicht unkompliziert. Die Standardverfahren von Halir [Hal99], Kampel [Kam03] und Cao und Mumford [CM02] basieren auf der Annahme, dass es sich bei dem Objekt um einen idealen Rotationskörper handelt, dessen Oberflächennormalen in umgekehrter Richtung exakt auf die Rotationsachse zeigen. Im vorliegenden Anwendungsfall ergeben sich daraus aber zwei wesentliche Probleme. Zum einen ist die Normalenberechnung, insbesondere für sehr fein aufgelöste polygona-

le Netze sehr rechenintensiv und daher für eine Serveranwendung eher ungeeignet, bei welcher es immer auch auf möglichst kurze Antwortzeiten ankommt. Zum anderen sind die Verfahren nicht robust genug bei der Verarbeitung von nicht idealen Rotationskörpern, wie z.B. stark deformierten Gefäßen oder solchen mit großen Henkeln oder starken Verzierungen, da hierbei das genannte Normalenkriterium häufig verletzt wird.

Wir haben uns daher für einen neuen Ansatz entschieden, bei welchem die Stärke von globalen Symmetrien innerhalb des Objektes untersucht wird. Zabrodsky et al. [ZPA95] haben hierzu vorgeschlagen, Symmetrie nicht strikt als vorhanden oder nicht vorhanden zu definieren, sondern als kontinuierliche Eigenschaft, die mehr oder weniger stark ausgeprägt sein kann. Das folgt aus der Beobachtung, dass den Formen vieler natürlicher und künstlicher Objekte Spiegel- oder Rotationssymmetrien innewohnen, die aber bei den allerwenigsten Objekten perfekt ausgebildet sind. Aufbauend auf dieser Überlegung haben Kazhdan und Podolak et al. [KFR04,PSG+06] verschiedene Formdeskriptoren für Spiegel- und Rotationssymmetrien entwickelt. Dies machte es möglich, die Stärke von Symmetrien für beliebige Geraden bzw. Ebenen relativ leicht und schnell zu berechnen.



(a) Teller (b) Schale (c) Schale

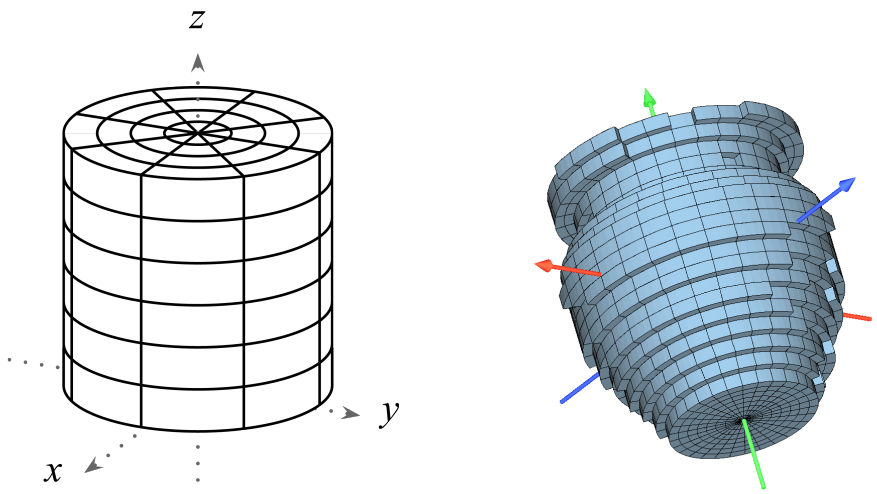
(d) Terrine (e) Schale

Abbildung 4: Beispiele für die Ausrichtung von Gefäßen an ihrer Rotationsachse.

Um nun die Rotationsachse eines beliebigen Objektes zu bestimmen, müssen wir nur diejenige Gerade finden, für welche ein gegebenes Objekt die größte Rotationssymmetrie aufweist. Da aber prinzipiell unendlich viele Geraden in Frage kommen, haben wir es hier mit einem Optimierungsproblem zu tun.

Als Optimierungsverfahren haben wir uns für die Partikelschwarmoptimierung (PSO) entschieden, die von Eberhart und Kennedy [KE95] entwickelt wurde. Sie wurde vom Verhalten von Fisch- und Vögelschwärmen inspiriert. In der Adaption als Algorithmus bewegen sich die Partikel durch einen mehrdimensionalen Suchraum, wobei die Position jedes Partikels eine mögliche Lösung darstellt. Die schwarmähnliche Bewegung der Partikel, basierend auf starker Interdependenz der einzelnen Partikel, sorgt dafür, dass mindestens eines der Partikel auf einer Position im Suchraum landet, die für eine optimale Lösung steht. Wir verweisen an dieser Stelle auf die entsprechende Literatur für weitere Details zu diesem Verfahren.

Das vorgestellte Verfahren hat sich als erstaunlich schnell und robust erwiesen. Die Bestimmung der Rotationsachse dauert im Schnitt etwa eine Sekunde und neben normal geformten Gefäßen werden auch sehr flache sowie annähernd kugelförmige Gefäße korrekt ausgerichtet. Einige Beispiel sind in Abbildung 4 dargestellt.



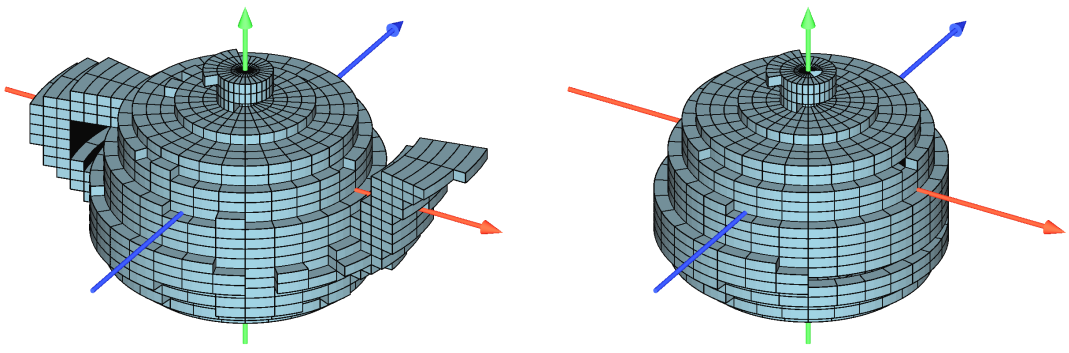
(a) Gitter (b) Rasterisierter Topf

Abbildung 5: Das Prinzip eines zylindrischen Voxelgitters (a) demonstriert an einem konkreten Beispiel (b).

Profillinie Nach der Ausrichtung des Gefäßes an seiner Rotationsachse müssen dessen in Abschnitt 3 beschriebene Eigenschaften aus der Profillinie abgeleitet werden. Zu diesem Zweck rasterisieren wir das Gefäß in ein zylindrisches Voxelgitter entlang der Rotationsachse. Dabei wird der Raum in paarweise disjunkte Zellen unterteilt, die sich in Anlehnung an das Konzept der Polarkoordinaten kreisförmig in der senkrecht zur Rotationsachse stehenden Ebene ausbreiten, wie in Abbildung 5 dargestellt. Es ist klar, dass einzelne Profillinien nun in Form von „Tortenstücken“ im Gitter vorliegen.

Zwei wichtige Aspekte müssen aber noch berücksichtigt werden, bevor die Profillinie extrahiert werden kann. Zum einen betrifft das mögliche Anhängsel wie Füße oder Henkel, welche vorher entfernt werden müssen. Wir haben dafür einen sehr einfache Ansatz entwickelt, der für jeden Kreis in der Gitterdarstellung die Anzahl der Oberflächenvoxel zählt und den Kreis löscht, falls er nicht zu mindestens 80 Prozent zur Oberfläche des Gefäßes zählt. Zum anderen sind Keramikgefäße selten wirklich perfekt rotationssymmetrisch, weshalb der Radius für jede Schicht entlang der Rotationsachse in geeigneter Weise geschätzt werden muss. Wir bestimmen zu diesem Zweck die Abstände aller Oberflächenvoxel einer Schicht von der Rotationsachse, sortieren diese Abstände und verwenden den Median als Repräsentanten.

Beide Ansätze haben sich als geeignet gezeigt, eine robuste Darstellung der Profillinie eines Gefäßes zu erzeugen, wobei die Rasterisierung der rechenaufwändigste Teil ist, der letztlich aber von der Qualität des vom Nutzer übergebenen Referenzmodells abhängt und daher nicht zu kontrollieren ist.



(a) Teekanne mit Ausguss und Henkel (b) Teekanne ohne Anhängsel

Abbildung 6: Beispiel für das Entfernen von Anhängseln einer Teekanne.

5 Ausblick

Wie in der Einleitung bereits erwähnt, ist die archäologische Forschung in den Bereichen der Generierung und der Arbeit mit Dokumentationen zu Fundstücken auf eine weitgehende Automatisierung der jeweiligen Prozesse angewiesen. In Zusammenarbeit mit dem Sächsischen Landesamt für Archäologie sind dafür an der Technischen Universität Chemnitz nicht nur die genannte Dokumentationssoftware *TroveSketch* entstanden, sondern auch andere Aspekte Gegenstand intensiver Forschungen. So ist auf Grundlage der Funddaten für ein Gräberfeld der Lausitzer Kultur unter Zuhilfenahme von aus der Künstlichen Intelligenz entlehnten Lernalgorithmen ein Typensystem für die Gefäße entwickelt worden.

Die zukünftigen Forschungen sollen der Frage nachgehen, ob dieses Typensystem auch auf ortsfremdes Material der Lausitzer Kultur anwendbar ist. Damit wird eine weiträumige Betrachtung dieser Kultur angestrebt, die wiederum genauere Aussagen über funktionale Aspekte ergeben kann.

Aus technischer Sicht sollen die Möglichkeiten einer Suche nach Gefäßen hinsichtlich der geplanten Auswertung erweitert werden. Dies zum einen, weil es eine gute Möglichkeit ist, sich effizient durch umfangreiche Datenmengen zu bewegen, zum anderen, weil sie eine sehr individuelle Sicht auf die Fundstücke erlaubt. Dabei sollen auch die Ergebnisse der o.g. automatischen Klassifizierung bei der Suche nach Gefäßen herangezogen werden. Für die Aufbereitung von dann möglicherweise mehrere Hundert Gefäße umfassenden Ergebnislisten sollen statistische Auswertungen implementiert werden, die zum Beispiel die räumliche Verteilung darstellen und den Archäologen somit Hinweise auf mögliche Migrationsbewegungen geben können.

Literatur

[BBO06] Brunner, David, Brunnett, Guido, and Oexle, Judith. Concept for an Application-Oriented Automated Classification System for Bronze Age Vessels. *EVA 2006 London Conference Proceedings, EVA 2006 London Conference*, pages 16.1–16.12, 2006.

[BKS+05] Bustos, Benjamin, Keim, Daniel A., Saupe, Dietmar, Schreck, Tobias, and Vranić, Dejan V. Feature-based similarity search in 3d object databases. *ACM Comput. Surv.*, 37(4):345–387, 2005.

[CM02] Cao, Yan and Mumford, David. Geometric structure estimation of axially symmetric pots from small fragments. In *Proceedings of the International Conference on Signal Processing, Pattern Recognition, and Applications*, 2002.

[Hal99] Halir, Radim. An automatic estimation of the axis of rotation of fragments of archaeological pottery: A multi-step model-based approach. In *Proc. of the 7th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media (WSCG'99)*, 1999.

[HB08] Hörr, Christian and Brunnett, Guido. Similarity Estimation on Ancient Vessels. In *to appear in GraphiCon Proceedings 2008*, 06 2008.

[HBB07] Hörr, Christian, Brunner, David, and Brunnett, Guido. Feature Extraction on Axially Symmetric Pottery for Hierarchical Classification. *Computer-Aided Design and Applications, Vol. 4, Nos. 1-4*, pages 375–384, 2007.

[Hör05] Hörr, Christian. Topologische klassifikation archäologischer gefäße. Studienarbeit, Fakultät für Informatik, Technische Universität Chemnitz, 2005.

[Kam03] Kampel, Martin. *3D Mosaicing of Fractured Surfaces*. PhD thesis, TU Wien, Institut für Automation, 2003.

[KE95] Kennedy, James and Eberhart, Russell C. Particle swarm optimization. In *Proceedings of IEEE International Conference on Neural Networks*, pages 1942–1948, NJ, 1995. Piscataway.

[KFR04] Kazhdan, Michael, Funkhouser, Thomas, and Rusinkiewicz, Szymon. Symmetry descriptors and 3d shape matching. In *SGP '04: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, pages 115–123, New York, NY, USA, 2004. ACM Press.

[MCF02] Min, Patrick, Chen, Joyce, and Funkhouser, Thomas. A 2d sketch interface for a 3d model search engine. In *SIGGRAPH '02: ACM SIGGRAPH 2002 Conference Abstracts and Applications*, page 138, New York, NY, USA, 2002. ACM Press.

[Min04] Min, Patrick. *A 3D Model Search Engine*. PhD thesis, Princeton University, Department of Computer Science, 2004.

[MKF04] Min, Patrick, Kazhdan, Michael, and Funkhouser, Thomas. A comparison of text and shape matching for retrieval of online 3d models. In *European Conference on Digital Libraries, Bath, UK*, 2004.

[PSG+06] Podolak, Joshua, Shilane, Philip, Golovinskiy, Aleksey, Rusinkiewicz, Szymon, and Funkhouser, Thomas. A planar-reflective symmetry transform for 3d shapes. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 549–559, New York, NY, USA, 2006. ACM Press.

[RRS03] Rowe, Jeremy, Razdan, Anshuman, and Simon, Arleyn. Acquisition, representation, query and analysis of spatial data: a demonstration 3d digital library. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 147–158, Washington, DC, USA, 2003. IEEE Computer Society.

[SMKF04] Shilane, Philip, Min, Patrick, Kazhdan, Michael, and Funkhouser, Thomas. The princeton shape benchmark. In *SMI '04: Proceedings of the Shape Modeling International 2004 (SMI'04)*, pages 167–178, Washington, DC, USA, 2004. IEEE Computer Society.

[ZPA95] Zabrodsky, Hagit, Peleg, Shmuel, and Avnir, David. Symmetry as a continuous feature. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(12):1154–1166, 1995.

Verschmelzendes Clustering in Artmap

Frederik Beuth und Marc Ritter

Technische Universität Chemnitz
Fakultät für Informatik
Professur Künstliche Intelligenz
{beuth,ritm}@informatik.tu-chemnitz.de

Zusammenfassung: Default-Artmap ist ein Datenanalyseverfahren, welches zur Objekterkennung einsetztbar ist. Dieses überwachte Clusterverfahren erreicht ähnliche Kategorisierungsgüten wie ein Künstliches Neuronales Netz, besitzt aber eine wesentlich geringere Laufzeit, da es die Eingabe nur einmal verarbeitet. Die Güte des Algorithmus ist aber von der Reihenfolge der Eingabedaten abhängig. Der hier vorgestellte Ansatz behebt das Problem, indem Default-Artmap um die Möglichkeit einer Clusterverschmelzung erweitert wird. Eine auf dem generischen 2d3cII-Datensatz durchgeführte Evaluation ergab verbesserte Kategorisierungsgüten und Clusterformen.

Schlagwörter: Objekterkennung, Clustering, Default-Artmap, Joint-Artmap

1 Einleitung

Neue Medien und Multimediatechniken erzeugen weltweit immer mehr Dokumente. Der technologische Fortschritt stellt die Suche, Dokumentation und Archivierung vor völlig neue Herausforderungen. Der Einsatz automatischer Annotationsverfahren bildet einen Ausweg den damit einhergehenden Mehraufwand vertretbar zu reduzieren. Dazu muss das Auftreten zuvor katalogisierter Objekte automatisiert innerhalb multimedialer Dokumente wie Bilder und Videos ermittelt und festgehalten (indiziert) werden.

Zu diesem Zweck kann das Gebiet der Künstlichen Intelligenz eine breite Palette von Ansätzen beisteuern, mit deren Hilfe speziell im Feld der Bild- und Videoverarbeitung relevante Objekte erkannt und beschrieben werden können. So versucht ein Großteil der Verfahren auf vorhandenen Videomaterialien Objekte in einer Einzelszene, d.h. einer Bildfolge gleichen oder ähnlichen Inhalts, zu erkennen. Die Objekterkennung funktioniert allerdings wesentlich robuster, wenn dem Verfahren kontextabhängiges Wissen über die ausgewählte Szene zur Verfügung steht.

Als ein vereinfachtes Beispiel kann die Aufgabe angesehen werden, einen roten Ball von einer roten Tomate zu unterscheiden, was anhand von Form und Farbe nur schwer möglich ist. Ein besseres Unterscheidungskriterium liefert der indiviuelle Szenenkontext mit Hintergrundinformationen, der Aufschluss darüber gibt, ob sich das zu suchende Objekt eher in einer Küche oder in einer Sporthalle auffinden lässt. Allerdings muss der Kontext für diese intellektuelle Verschlagwortung erst dem System vermittelt werden, was im Allgemeinen aufwendig ist. Dieses Problem soll mit Lernregeln aus der Künstlichen Intelligenz gelöst werden [Alp04], wobei das verwendete Verfahren robust gegenüber fehlerhaften Schlagwörtern sein und eine geringe Laufzeit aufweisen muss.

Eine Möglichkeit zur schnellen Berechnung bilden Single-Pass-Verfahren, weil sie die Eingabefolge nur einmalig verarbeiten [Kür06]. Diese besitzen aber gleichzeitig den Nachteil, dass die daraus resultierende Objekterkennung oftmals von der Eingabereihenfolge abhängig ist [Beu08](Kap 2.2). Der vorliegende Beitrag zeigt einen Ansatz auf, diese Abhängigkeit zu vermindern und damit Laufzeit und Qualität gleichermaßen zu verbessern.

2 Default-Artmap

Default-Artmap ist ein zur Objekterkennung einsetzbares überwachtes Datenclusteranalyseverfahren [Jai99]. Innerhalb der Clusterverfahren zählt es zur Unterkategorie der relativ schnellen Single-Pass-Verfahren [Ras92], [Alp04](Kap. 7.6). Die in dieser Arbeit vorgenommenen Verbesserungen können an jedem Single-Pass-Clusterverfahren eingesetzt werden, sollen hier aber an Default-Artmap demonstriert werden. Dieser Abschnitt fasst das Verfahren aus den englischen Orginalartikeln zusammen [Car03,CGM$^+$92,Car97]. Eine deutsche Beschreibung ist in [Beu08] zu finden.

Der Algorithmus erkennt Objekte, indem er sie in einzelne Gruppen (bezeichnet als Cluster) einteilt. Zu diesem Zweck muss jede Szene auf ein Bündel von M Merkmalen abstrahiert werden. Jede Szene stellt damit einen Punkt in einem M-dimensionalen Merkmalsraum dar. Zur besseren Übersicht beschränkt sich das vorliegende Minimalbeispiel auf nur zwei Merkmale, dessen Merkmalsraum sich vorteilhaft visualisieren lässt (siehe Abbildung 1 für $M = 2$).

Als Minimalbeispiel kann die Erkennung einer Szene anhand von Farben gewählt werden. Eine Eignung des reduzierten HS-Raums aus dem Modell des HSV-Farbraums (vgl. [GW08]) zur Erkennung bestimmter vordefinierter Szenen wird in [Krö09] beschrieben.

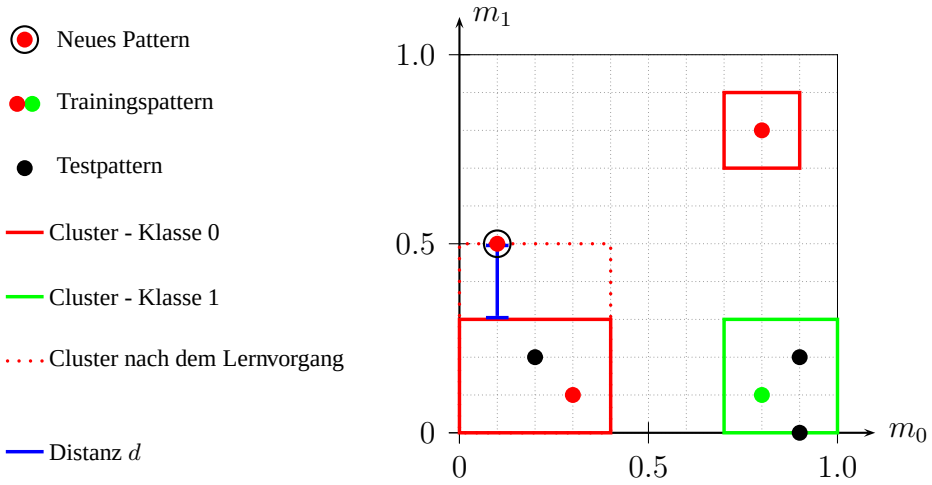


Abbildung 1: Beispiel zum Merkmalsraum: Auf den Achsen sind die beiden Merkmale aufgetragen. Die Punkte repräsentieren die Daten, die Vierecke die Cluster. (vgl. [Beu08](Fig. 2.1))

Das Merkmal m_0 entspricht dabei dem Kanal des Farbtons (engl. *hue*), m_1 dem Kanal der Farbsättigung (engl. *saturation*). Damit wird eine helligkeitsunabhängige Objekterkennung gewährleistet. Diverse natürliche Szenen wie Wüste, Arktis, Ozean und Wälder lassen sich anhand ihrer HS-Merkmale unterscheiden.

Ein Cluster repräsentiert eine Gruppe von Punkten im Merkmalsraum und eine Klasse die Art der Szene. Meist sind diese beiden Begriffe synonym zu gebrauchen. Eine Besonderheit stellt der Fall dar, wenn mehrere Cluster eine Klasse repräsentieren. Exemplarisch lassen sich so Tag- und Nachtaufnahmen einer Wüste trotz unterschiedlicher Merkmale einer gemeinsamen Klasse zuzuordnen. In Abbildung 1 enthält die Klasse 0 zwei Cluster mit differenten Merkmalen.

Um eine Semantik in das System einzubringen, verknüpft Default-Artmap in einem Lernvorgang innerhalb der Trainingsphase Merkmale und Klassenlabels (dargestellt durch die farbigen Punkte in Abbildung 1). In der anschließenden Testphase werden zu ungelernten Merkmalen Klassenlabels bestimmt und auf ihre Richtigkeit kontrolliert (schwarze Punkte in Abbildung 1).

Default-Artmap verarbeitet die Daten prinzipiell sequentiell. Die Trainingsphase ordnet neue Eingaben (Merkmale) dem ähnlichsten Cluster zu und vergrößert diesen bei Bedarf.

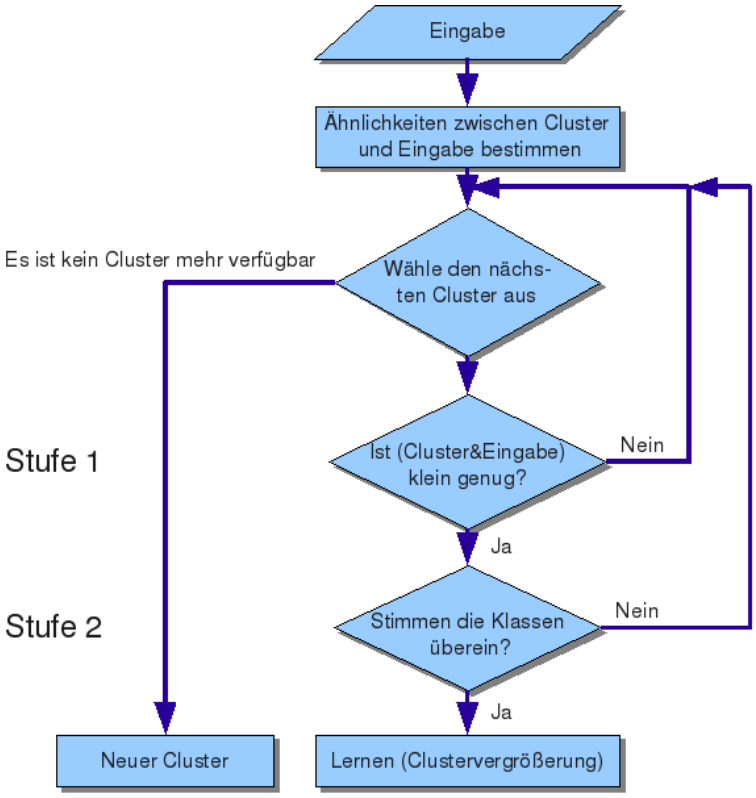


Abbildung 2: Programmablaufplan von Default-Artmap. (vgl. [Car03](Fig. 1))

In Abbildung 1 markiert die blaue Linie die größtmöglich ermittelte Ähnlichkeit zu allen Clustern. Ebenso ist eine Vergrößerung des Clusters nach dem Lernen als gepunktete Linie eingetragen. Findet sich hingegen kein ähnlicher Cluster, wird ein neuer Cluster aus der Eingabe erzeugt.

Default-Artmap (siehe Abbildung 2) vergrößert einen relevanten Cluster in einem zweistufigen Prozess. Zuerst wird der ähnlichste Cluster zu der Eingabe bestimmt, was der Manhattan-Distanz [Beu08](Kap. 2.3.2) zwischen der nächsten Clustergrenze und dem Eingabemerkmalsvektor entspricht. Anschließend prüft das Verfahren die Größe des entstehenden Clusters. Sie darf einen bestimmten Wert nicht überschreiten. Im zweiten Schritt werden die Klassenlabels von dem Cluster und der Eingabe auf Übereinstimmung verglichen. Falls ein Kriterium unerfüllt bleibt, wird der aktuelle Cluster nicht mehr betrachtet und ein anderer Cluster getestet. Dies wird solange fortgesetzt, bis ein passender

Cluster gefunden wurde. Alternativ erzeugt der Algorithmus aus der Eingabe einen neuen Cluster.

3 Joint-Artmap

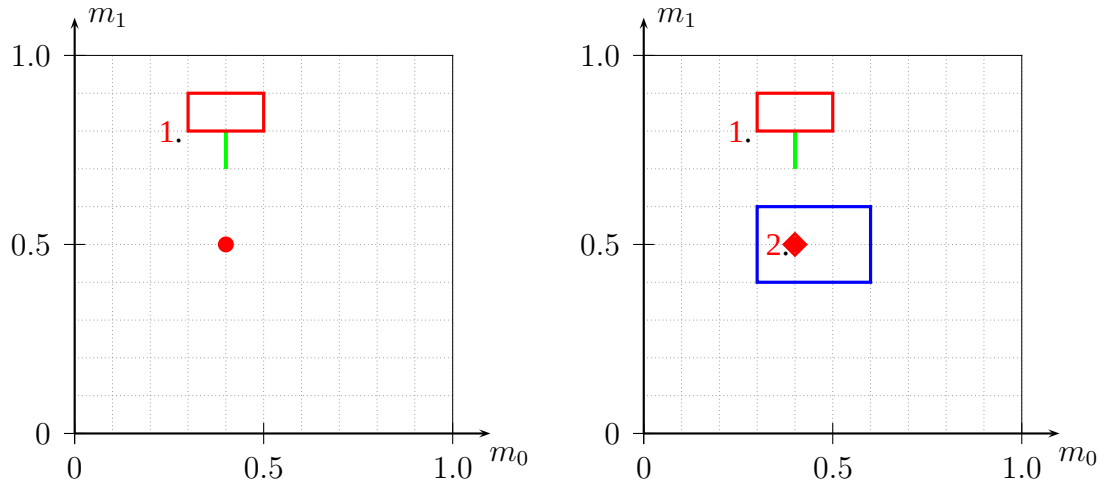
Bedingt durch seine Sequentialität verarbeitet Default-Artmap neue Eingaben nur mit den bis zu diesem Zeitpunkt gelernten Clustern. In bestimmten Fällen treten Fehler bei der Entscheidung auf, entweder einen neuen Cluster zu erzeugen oder einen bestehender Cluster zu vergrößern. Dieser Problemfall wird im nächsten Abschnitt beschrieben. Um diese Fehler zu verhindern, wurde eine Heuristik entwickelt, welche den Fall gesondert erkennt. Der Ansatz erzeugt im Zweifelsfall immer einen neuen Cluster, kann aber die Cluster nachträglich verschmelzen. Der neue Algorithmus wird als Joint-Artmap bezeichnet und stellt eine lose Weiterentwicklung des Default-Artmaps dar.

Das Problem tritt auf, wenn eine Eingabe verarbeitet werden soll, die zu keinem Cluster sehr ähnlich ist (Abbildung 3(a)). In diesem Fall ist zum momentanen Zeitpunkt unbekannt, ob die neue Eingabe besser dem Cluster 1 zugeordnet werden oder ein neuer Cluster entstehen sollte.

Werden zukünftige Eingaben in die Überlegungen mit einbezogen, lassen sich drei mögliche Fällen unterscheiden:

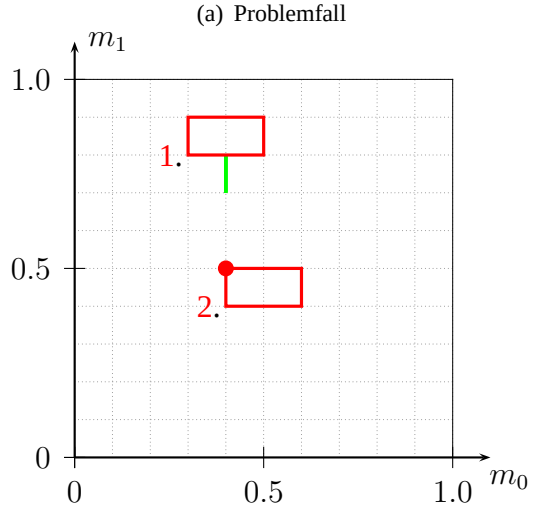
1. Die aktuelle Eingabe ist fehlerhaft oder ein Sonderfall. Es existieren an diesem Punkt keine weiteren Daten (Abbildung 3(b)). Da das Verfahren nicht explizit zwischen Sonderfällen und Datenfehlern unterscheiden kann, sollte ein neuer punktförmiger Cluster erzeugt werden. Diese Behandlung ist möglich, da ein Punktcluster die Generalisierung außer an dieser Stelle nicht ändert. Der Fall wird als Datenausnahme bezeichnet.
2. Aus der Eingabe wird später ein eigener Cluster entstehen (Abbildung 3(c)). Zum aktuellen Zeitpunkt sollte ein neuer Cluster erstellt werden.
3. Die Eingabe kann durch den aktuellen Cluster repräsentiert werden (Abbildung 3(d)). Eine Vergrößerung des entsprechenden Clusters ist sinnvoll.

Folglich ist es nicht möglich, immer nur eine Vergrößerung durchzuführen oder einen neuen Cluster zu erstellen. Der aktuelle Zeitpunkt erlaubt nicht die obige Fallunterscheidung, weshalb das gesamte Problem erst später entscheidbar ist.

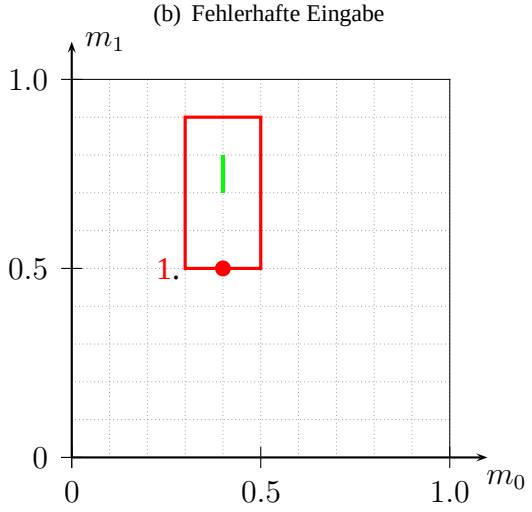


(a) Problemfall

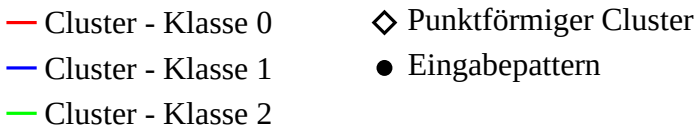
(b) Fehlerhafte Eingabe



(c) Eigener Eingabecluster



(d) Clustervergrößerung



(e) Legende

Abbildung 3: Problemfall in Default-Artmap zum aktuellen Zeitpunkt (a) und zukünftige Clusterverteilungen (b,c,d).

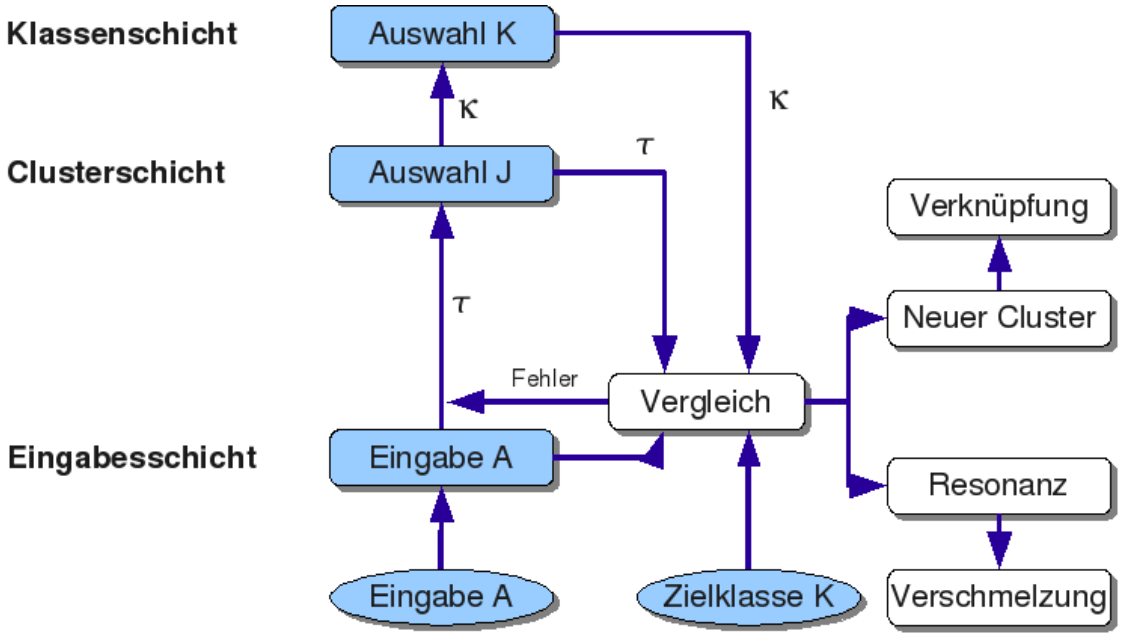


Abbildung 4: Joint-Artmap Informationsfluss

Die hier entwickelte Lösung geht folgendermaßen vor:

1. Es wird immer ein neuer Cluster j erzeugt.
2. Der bereits existierende Cluster t (Nr. 1 in Abbildung 3(a)) wird mit dem neuen Cluster j verlinkt. Diese Verknüpfung wird in einer Liste abgespeichert.
3. In regelmäßigen Abständen wird (die Liste) überprüft, ob beide Cluster genügend vergrößert wurden, so dass sie sich nun ähnlich sind. In diesem Fall werden sie zu einem Cluster verschmolzen. Mit dem Mechanismus kann der Algorithmus falsche Entscheidung korrigieren.

Der nächste Abschnitt stellt den neuen Joint-Artmap-Algorithmus vor. Abbildung 4 gibt einen Überblick in Form des Informationsflusses. Jeder einzelne Schritt wird als Pseudocode (Algorithmus 1, 2) beschrieben. Die Notation (Tabelle 5) lehnt sich an das orginale Default-Artmap-Verfahren an. Sie ist in [Car03](IV.B) und [Beu08](Anhang A) beschrieben.

```
1  Initialisiere: τ = 0, C = 0;
2  while Eingaben vorhanden do
3      Wähle die nächste Eingabe (Merkmalsvektor A mit Zielklasse K_prime) aus;
4      ∀j Berechne als Maß für die Ähnlichkeit: T_j = d(j, A) + (s_j / M) * α;
5      Sortiere die Cluster aufsteigend nach T in die Liste Λ ein;
6      while Λ nicht leer do
7          Wähle den nächsten Cluster J = front{Λ} aus;
           // Vergleiche Distanz und Clustergröße:
8          if d(J, A) > ξ * M then
               // Die Distanz ist zu groß.
9              Merke für den Verschmelzungslink: t = J;
10             Deaktiviere J: Λ = Λ \ J und überspringe den Schleifenkörper;
11         X = τ_J ⊕ A;    K = κ_J;
12         if 1 − s(X)/M < ρ then
               // Der Cluster ist zu groß.
13             Deaktiviere J: Λ = Λ \ J und überspringe den Schleifenkörper;
           // Vergleiche die Klasse:
14         if K ≠ K_prime then
               // Der falsche Cluster wurde ausgwählt.
15             Erhöhe die Vigilance auf ρ = 1 − s(X)/M + ε ;
16             Deaktiviere J: Λ = Λ \ J und überspringe den Schleifenkörper;
17         else
               // Resonanz bzw. Lernen
18             Vergrößere Cluster auf: τ_J = τ_J ⊕ A;
               // Verschmelzung: Überprüfe die Verschmelzungsliste
19             forall {t,j}=jointsList do
20                 if d(j, t) ≤ ξ * M und 1 − s(τ_j ⊕ τ_t)/M > ρ then
21                     verschmelze die Cluster j und t zu j: τ_t = τ_t ⊕ τ_j;
22                     lösche Cluster j;
23                     forall jointsList do
24                         suche nach j, ersetze durch t;
25                 if Es gab eine Verschmelzung then
26                     Überprüfe: führe den Schritt Verschmelzung erneut aus ;
27 ...
```

Algorithm 1: Joint-Algorithmus (Teil 1)

```
1 ...
2 if Λ leer then
      // Erzeuge neuen Cluster
3     Setze die initialen Gewichte für einen neuen Cluster:;
4     j = C; τ_C = A; κ_C = K_prime;
5     if t ≠ leer then
6         joinsList.add: (C, t);
7     C = C + 1;
8     Setze die Vigilance zurück ρ = ρ̄;
```

Algorithm 2: Joint-Algorithmus (Teil 2)

Variable	Beschreibung
M	Anzahl Merkmale
C	Anzahl Cluster
i	Merkmalsindex
j	Clustersindex
t	Nummer des bereits existierenden Clusters.
$joinsList$	Liste mit allen Verschmelzungs-Links.
ξ	Der Parameter regelt die nötigte Distanz/Ähnlichkeit zwischen Eingabe und Cluster für den Problemfall.
$\overline{\rho}$	Der Parameter begrenzt die maximale Clustergröße.
ϵ, α	Parameter für Datenausnahmen.
$\tau_{i,j}$	Die Variable speichert die zwei Eckpunkte des Clusters j ab. $0 \leq i < M \mathrel{\widehat{=}}$ unterer Punkt, $M < i < 2M \mathrel{\widehat{=}}$ oberer Punkt.
κj	Abbildung des Clusters j auf eine Klasse [Beu08](Kap. 4.6.2).
$d(j,t)$	Manhattendistanz $d(j,t) = \sum_i \begin{cases} \tau_{i,j} - \tau_{i+M,t} & \text{für } \tau_{i+M,t} < \tau_{i,j} \\ \tau_{i+M,t} - \tau_{i,j} & \text{für } \tau_{i+M,j} < \tau_{i,t} \\ 0 & sonst \end{cases}$
s	Summe der Clusterseitenlängen $s_j = \sum_i \tau_{i+M,j} - \tau_{i,j}$

Abbildung 5: Notation Joint-Artmap

4 Evaluation

In diesem Abschnitt werden die beiden Systeme auf ihre Kategorisierungsgüte und auf die Qualität der Clusterausbildung hin untersucht. Primäres Ziel der Entwicklung war eine Verbesserung der Clusterformen. Für den Praxiseinsatz ist es jedoch essentiell, dass sich die Kategorisierungsgüte nicht verschlechtert. Zusammenfassend konnte bei beiden Kenngrößen eine Verbesserung gegenüber dem herkömmlichen Default-Artmap Verfahren erzielt werden.

Als Datensatz kam bei beiden der 2d3cII-Datensatz zum Einsatz. Er enthält die ursprünglichen 30 Trainingsdaten des 2d3c-Datensatzes (Abbildung 6(a)) aus [Beu08] und eine Erweiterung auf 60 Testdaten (Abbildung 6(b)).

Die Aufgabe für ein Verfahren ist es, die drei großen Cluster zu erzeugen und ebenso mit den fehlerhaften Daten oder Ausnahmen umgehen zu können (vgl. TrainingsNr. 29). Die Trainingsreihenfolge bildete eine nach Klassen geordnete Reihenfolge, da diese den Worst-Case für Default-Artmap darstellt ([Beu08](Kap. 2.4.3 und Kap. 6.3.1).

Die Verfahrensparamter wurden für die Evaluation optimiert mit folgendem Ergebnis:

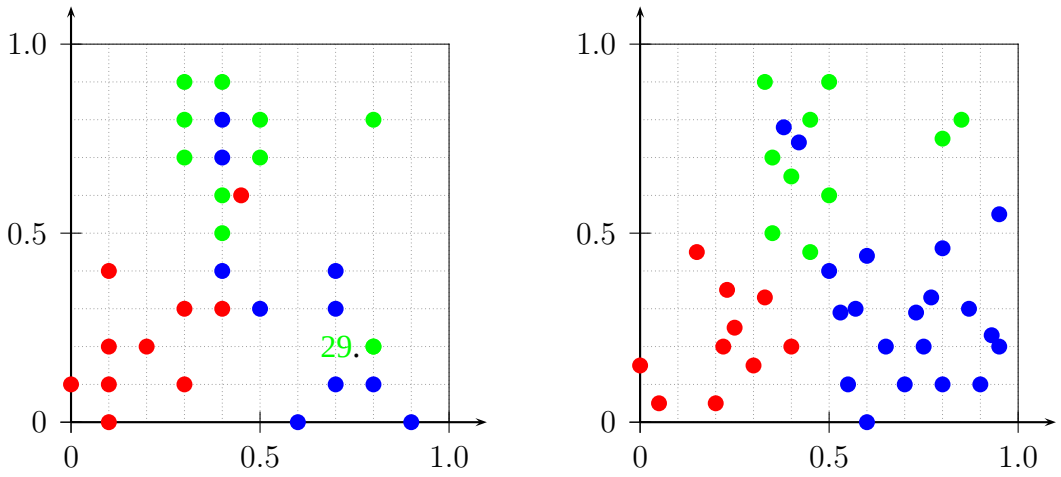
$$\begin{array}{rl} \text{Joint-Artmap:} & \overline{\rho} = 0.5, \quad \alpha = 0.1, \quad \epsilon = 0.1, \quad \xi = 0.12 \\ \text{Default-Artmap:} & \overline{\rho} = 0.6, \quad \alpha = 0.01, \quad \epsilon = -0.001 \end{array}$$

4.1 Clusterformen

Das Ziel der Clusteranalyse ist es, die Eingabedaten mit möglichst wenigen Clustern zu repräsentieren. Für einige Eingabedaten ist dies nicht möglich, zum Beispiel Sonderfälle oder falsche Eingaben. Das Verfahren behandelt solche Datenausnahmen als punktförmige Cluster. Die Clusterformen wurden durch Joint-Artmap sehr verbessert und es verringert sich die insgesamt benötigten Clusteranzahl.

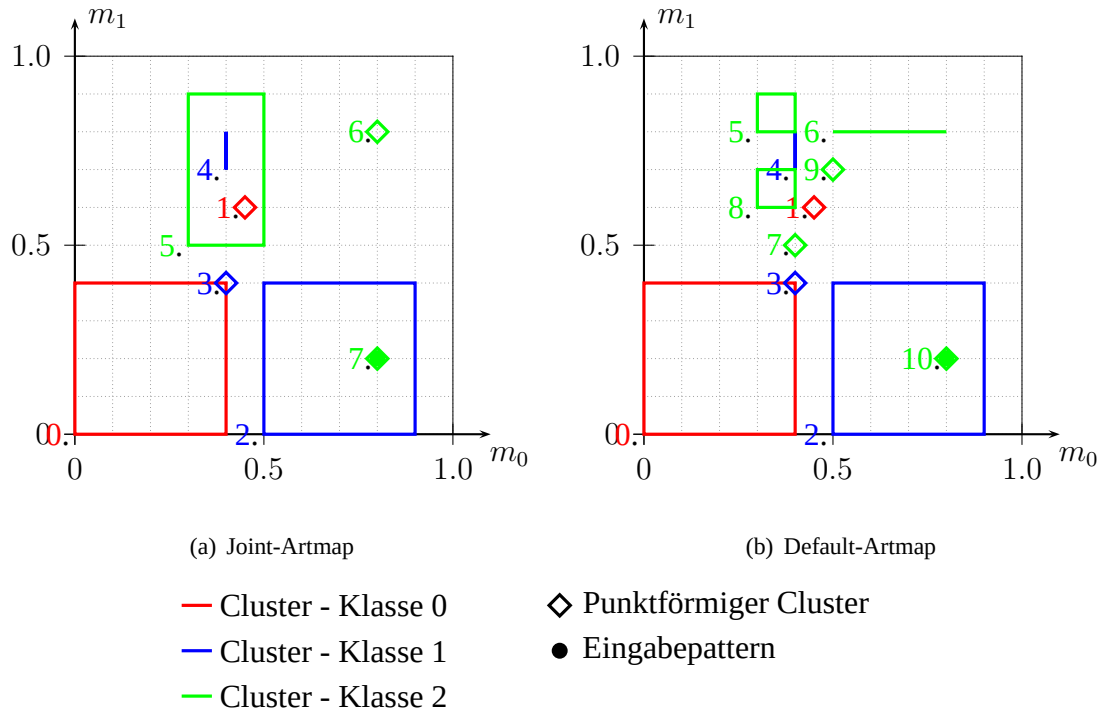
Abbildung 7(a) veranschaulicht die entstandenen Clusterformen anhand des 2d3cII-Datensatzes. Es handelt sich um eine ideales Clusterergebnis, da die meisten Daten durch drei große Cluster repräsentiert (Nr. $0, 2, 5$) werden, während die fünf kleineren Datencluster Ausnahmen darstellen.

Demgegenüber stellt die Abbildung 7(b) die Clusterung mit Default-Artmap dar. Auffällig ist, dass der obige große Cluster (Nr. 5, grün in Abbildung 7(a)) nicht ausgebildet wurde. Stattdessen wurden mehrere kleinere Cluster erzeugt. Dies erlaubt die Folgerung,



(a) Trainingsdaten (b) Testdaten

Abbildung 6: Datenmengen 2d3cII-Datensatz



(a) Joint-Artmap (b) Default-Artmap

(c) Legende

Abbildung 7: Clusterformen

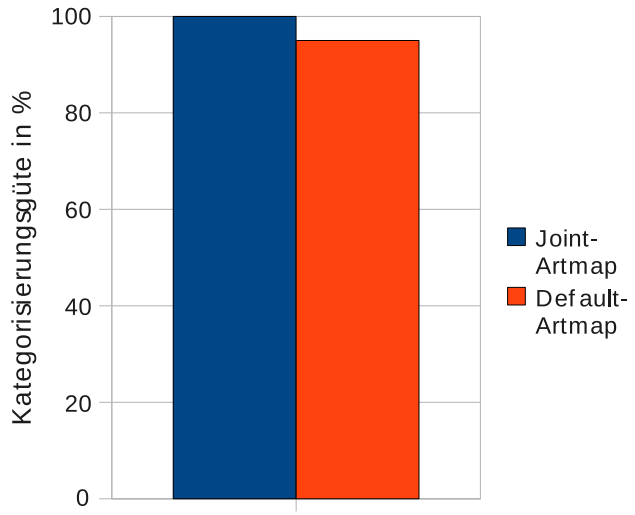


Abbildung 8: Kategorisierungsgüte

dass die Datengeneralisierung für diese Datenmenge fehlt. Die Ursache ist der Cluster Nr. 4 (blau). Er stört den Generalisierungsprozess. Er wurde erzeugt, bevor die Cluster der Klasse 2 (in grün dargestellt) erzeugt werden. Dies ist eine direkte Folge der vorhandenen Ordnung der Eingabedaten nach Klassen und ist ein Beispiel für die Abhängigkeit der Kategorisierungsgüte von der Eingabereihenfolge [Beu08](8.2.2) bei Default-Artmap. Somit lässt sich durch Joint-Artmap noch ein drittes Ziel indirekt erreichen: die Verminderung der Eingabedatenreihenfolgeabhängigkeit.

4.2 Kategorisierungsgüte

Eine neues Verfahren mit verbesserter Generalisierung muss eine gleich gute oder bessere Kategorisierungsgüte vorweisen können. Wie eingangs beschrieben, wurden die Parameter beider Systeme optimiert und die jeweiligen Güte auf der Testmenge gemessen. Abbildung 8 stellt das Ergebnis dar. Wie ersichtlich wird, erhöht die verbesserte Generalisierung in Joint-Artmap die Güte auf 100% (linke Säule). Default-Artmap ist demgegenüber etwas schlechter mit 95%.

5 Fazit

Das hier vorgestellte Joint-Artmap ist eine Weiterentwicklung von Default-Artmap mit dem Ziel die Clusterformen zu verbessern. Dieses wurde auf dem 2d3cII-Datensatz vollständig erreicht und der Algorithmus erzielt dabei eine optimale Clusterform und eine Kategorisierungsgüte von 100%. Darüber hinaus ist die Kategorisierungsgüte wesentlich robuster gegenüber der Eingabedatenreihenfolge als im orginalen Default-Artmap.

Nach dem erfolgreichen Abschluss der Pilotstudie auf dem 2d3cII-Datensatz kann die Anwendung und Evaluation derartiger Clusterfähigkeiten auf komplexen Realweltanwendungen, wie natürlichen Szenen, Gegenstand weiterer Forschungsarbeiten sein.

Literatur

[Alp04] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2004.

[Beu08] Frederik Beuth. Vergleich der Klassifikationsfähigkeit von Default Artmap mit Backpropagation-Netzen. Master's thesis, Technische Universität Chemnitz, 2008.

[Car97] Gail Carpenter. Distributed learning, recognition, and prediction by art and artmap neural networks. *Neural Netw*, 10(8):1473–1494, November 1997.

[Car03] G.A. Carpenter. Default ARTMAP. In *IJCNN*, pages 1396–1401, 2003.

[CGM+92] G.A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans Neural Netw*, 3(5):698–713, 1992.

[GW08] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*, chapter 6.2.3. Pearson Prentice Hall, third edition, 2008.

[Jai99] A.K. Jain. Data Clustering: A Review. In *ACM Computing Surveys*, volume 31. 1999.

[Krö09] Ulrike Krönert. Umgebungsklassifikation mittels Kamera und Lasersensoren zur Navigation eines mobilen Roboters. Master's thesis, Technische Universität Chemnitz, 2009.

[Kür06] Jens Kürsten. Systematisierung und Evaluierung von Clustering-Verfahren im Information Retrieval. Master's thesis, Technische Universität Chemnitz, 2006.

[Ras92] E. Rasmussen. *Information Retrieval: Data Structures & Algorithms*, page 427. Prentice-Hall, 1992.

Von der Bildrepräsentation zur Objekterkennung – Bewegungsanalyse als mächtiges Werkzeug der automatischen Bildinterpretation

Tobias John, Basel Fardi und Gerd Wanielik

Technische Universität Chemnitz
Fakultät für ET/IT
Professur Nachrichtentechnik

{tobias.john,basel.fardi}@etit.tu-chemnitz.de

Zusammenfassung: In diesem Beitrag wird ein Überblick über die Bewegungsanalyse und der Anwendung hinsichtlich der Szeneninterpretation gegeben. Die Berechnung des optischen Flusses als ein grundlegendes Werkzeug in der Videoverarbeitung wird eingeführt und anhand dessen die Detektion und Analyse von Bewegung in Bildern. Dabei wird schrittweise von der Bewegungsschätzung zur Detektion von Objekten bis hin zu deren Klassifikation mittels der Bewegungsinformation vorgegangen.

Schlagwörter: Optischer Fluss, Bewegungsdetektion, Bewegungsanalyse, Objektdetektion

1 Einleitung

Die Bewegungserkennung ist eine für die menschliche Wahrnehmung sehr entscheidende Information. Erst mit der Fähigkeit Bewegung wahrzunehmen sind wir in der Lage Geschwindigkeiten abzuschätzen oder zu erkennen, wo ein Objekt im nächsten Moment sein wird. Für uns ist es sogar möglich Emotionen anhand von Bewegungen zu erkennen, wie Wut oder Trauer. Aufgrund der Selbstverständlichkeit mit der wir diese Information aufnehmen und auswerten, unterschätzen wir leicht ihre Bedeutung.

Videos, als eine Folge von zeitlich nacheinander aufgenommenen Bildern, ermöglichen es Bewegungen festzuhalten. Bewegungen zwischen aufeinander folgenden Bildern sind nichts weiter als eine Verschiebung des Bildinhaltes von einer Aufnahme zu ihrer zeitlich nachfolgenden. Diese gerichtete Verschiebung wird als optischer Fluss bezeichnet (optical flow).

Im Folgenden wird erläutert, wie aus Videos die Bewegungsinformation gewonnen werden kann. Es wird dargestellt wie es allein anhand der Bewegung möglich wird Objekte zu detektieren und weiterhin, diese sogar zu klassifizieren.

2 Optischer Fluss – Grundwerkzeug der Bewegungsdetektion

Wie im Abschnitt 1 kurz dargestellt, ist die Bewegungsinformation für uns Lebewesen sehr bedeutend, weshalb die Bestimmung des optischen Flusses zu den wohl grundlegendsten Verfahren der Videoverarbeitung gehört.

Eine Möglichkeit die unterschiedlichen existierenden Methodiken einzuordnen, ist die Unterscheidung der „Ebene“ auf der sie arbeiten. Die Pixel als Rohdaten stellen die unterste Ebene dar.

Auf dieser wird kein weiteres Verfahren als Vorverarbeitung vorausgesetzt. Häufig lässt sich jedoch die Bewegung erst eindeutig erkennen, wenn das Objekt welches sie verursacht hat bekannt ist. Daher existieren Ansätze, die im Gegensatz zu den pixelbasierten Verfahren mindestens einen Verarbeitungsschritt, zum Beispiel eine Segmentierung des Bildes zur Generierung von Objekten, voraussetzen.

2.1 Pixelbasierte Verfahren

Ziel der pixelasierten Verfahren ist es Punktekorrespondenzen zu finden. Zu diesem Zweck existieren bereits unterschiedliche Methodiken [BB95]. U.a. gehören dazu Ansätze welche auf Korrelation, zeitlichen und räumlichen Ableitungen und Frequenz- und / oder Phaseninformation basieren.

Beim Korrelationsverfahren wird das Bild in Blöcke zerlegt und eine zeitliche Zuordnung dieser wird gesucht. Bei der Nutzung von Frequenz– oder Phaseninformation werden im Fourier-Bereich spezielle Filter (sensitiv auf Geschwindigkeit oder Orientierung) angewendet. Die differentiellen Ansätze nutzen die Ableitungen (örtlich, als auch zeitlich) um die Verschiebung zwischen aufeinander folgenden Bildern zu bestimmen.

Als eine der weit verbreitetsten, wird hier auf die differentielle Methode etwas genauer eingangen:

Die Grundlage hierfür bildet der Ansatz, dass die Intensität einer Oberfläche sich über die Zeit nicht ändert, zumindest für einen kleinen Ausschnitt. Demnach ist die Intensität

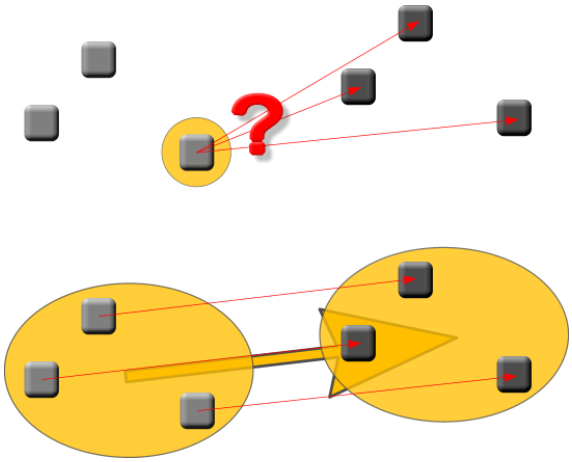


Abbildung 1: Blendenproblem: erst eine entsprechende große Umgebung ermöglicht eine Zuordnung

$I(x, y, t)$ zum Zeitpunkt t an der Stelle (x, y) gleich der Intensität $I(x + \triangle x, y + \triangle y, t + \triangle t)$ an der um $(\triangle x, \triangle y)$ verschobenen Stelle im nachfolgenden Bild.

Weil für jeden Bildpunkt (x,y) zwei Unbekannte $(\triangle x, \triangle y)$ zu berechnen sind, ist das Gleichungssystem unterbestimmt[1] und eine weitere Einschränkung muss vorgenommen werden. Wie z.B. die Betrachtung einer Umgebung des jeweiligen Bildpunktes um weitere Gleichungen zur Lösung des Problems zu gewinnen.

Die lokalen Verfahren setzen einen i.d.R. konstanten Fluss innerhalb der lokalen Umgebung $\mathcal{R}$ an [LK81], während die globalen Verfahren eine allgemeine „Glattheitsbeschränkung“ einführen, welche die Annahme beschreibt, dass angrenzende Bewegungen auch ähnlich geartet sein müssen [HS81]. Neben der Annahme der konstanten Intensität ist es auch denkbar abgeleitete Größen wie den Gradienten [ON95] oder sogar Strukturelemente als konstant anzunehmen. Um auch größere Verschiebungen sicher zu detektieren, werden die genannten Methodiken häufig als Multiskalen-Implementationen erweitert [Ana89,Bou02]. Dabei werden erste grobe Schätzungen sukzessive verfeinert.

2.2 Objektbasierte Verfahren

Objektbasierte Verfahren setzen nicht auf Pixelebene an, sondern fassen Pixel bereits zu Objekten zusammen. Diese werden anschließend von Bild zu Bild verfolgt (Tracking), so dass die komplette Objektbewegung erfasst werden kann.

---

[1] Dieser Sachverhalt wird in der Literatur als Blendenproblem bezeichnet (siehe Abb. 1)

Die Zuordnung von Bild zu Bild erfolgt anhand eines Abstandsmaßes der Merkmalsvektoren der Objekte. Die Merkmale können u.a. Farbe, Kompaktheit und weitere Momente sein. Das simple Beispiel in Abb. 2 zeigt eine Zuordnung der Objekte anhand Fläche (A) und Farbe (C).

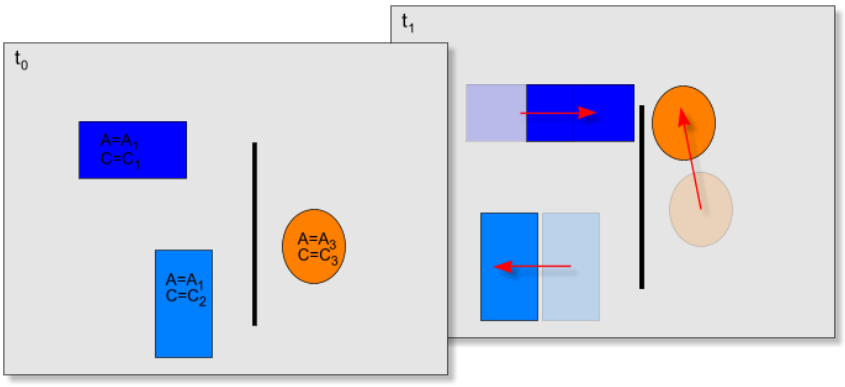


Abbildung 2: Verschiebung anhand objektbasierter Zuordnung

3 Objektdetektion

Wie im vorangegangenen Kapitel beschrieben, ist der optische Fluss eine Repräsentation der Verschiebung bzw. Bewegung zwischen zwei aufeinander folgenden Bildern. Wird die Richtung der Bewegung als Farbe kodiert und der Betrag als Helligkeit, so lässt sich das Bewegungsvektorfeld auch als Bild darstellen.

3.1 Objektdetektion ohne Modell

Die Objektdetektion ist der auf die Berechnung der Bewegung folgende Schritt. Liegt der Fall einer statischen, sich nicht bewegenden Kamera vor, so ist jegliche detektiere Bewegung eine reine Fremdbewegung und kann direkt ausgewertet werden.

Das Bewegungsbild kann somit hinsichtlich Richtung und/oder Betrag segmentiert werden um Objekte zu bilden, deren Pixel sich in Bezug auf ein definiertes Abstandsmaß ähnlich verhalten.

Anders, sobald auch die Kamera bewegt ist. Dann ist jegliche detektierte Bewegung die Summe aus Eigenbewegung und möglicher Fremdbewegung.

Dennoch kann auch hier der Ansatz einer Segmentierung zu brauchbaren Ergebnissen führen wie in [Ric04] gezeigt wird und auch Abb. 3 illustriert. Der Motorradfahrer zeichnet sich beim Überholen deutlich aus dem Bewegungsfeld ab und kann als bewegliches Objekt gekennzeichnet werden (blau).

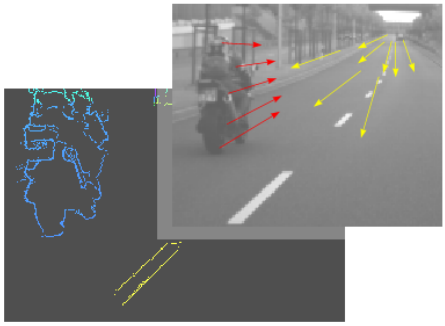


Abbildung 3: eindeutige Detektion des überholenden Motorrads

3.2 Objektdetektion anhand eines Modells

Modell der eigenen Bewegung zur Detektion bewegter Objekte

Um die überlagerte Fremdbewegung zu selektieren, ist es notwendig die eigene Bewegung der Kamera zu bestimmen. Dies kann über zusätzliche Sensoren erfolgen, oder durch Schätzung der Eigenbewegung aus dem optischen Fluss selbst.

Dazu ist es notwendig ein Modell für die Bewegung der Kamera zu erstellen. Die im Bild vorkommende Bewegung wird nun auf Anwendbarkeit des Modells überprüft. Liegt diese vor, so handelt es sich um einen sich in der realen Welt nicht bewegenden Punkt, dessen Bewegung im Bild allein durch die Kamerabewegung hervorgerufen wurde. Er kann also für die Bestimmung der Eigenbewegung genutzt werden. Lässt sich das Modell auf einen Punkt nicht anwenden, so ist es möglich, dass sich dieser selbst bewegt hat oder seine Verschiebung einfach falsch ermittelt wurde (ein Ausreißer). Er kann jedoch nicht zur Eigenbewegungsschätzung herangezogen werden [FJW09,HZ04].

Aus den zum Modell passenden Punkten kann die Eigenbewegung berechnet werden und anschließend von allen Bewegungen im Bild abgezogen werden. Dieser Vorgang wird als Eigenbewegungskompensation bezeichnet.

Jegliche in dem so kompensierten Bild verbleibende Bewegung rührt nun von sich selbst bewegenden Objekten her und kann dementsprechend weiter behandelt werden. Es ist demnach möglich sowohl die Eigenbewegung der Kamera zu ermitteln, als auch sich selbst bewegende Objekte zu detektieren.

[FJW09] setzen ein bilineares Bewegungsmodell ein um die Verschiebung zwischen zwei Bildern zu beschreiben:

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_0 x + a_1 y + a_2 + a_3 xy \\ a_4 x + a_5 y + a_6 + a_7 xy \end{bmatrix}$$

Abbildung 4 zeigt wie trotz bewegter Kamera die Fremdbewegung eines Fußgängers als solche erkannt (rot markiert) wird. Punkte die sich durch die Eigenbewegung verschoben haben, sind grün gekennzeichnet.

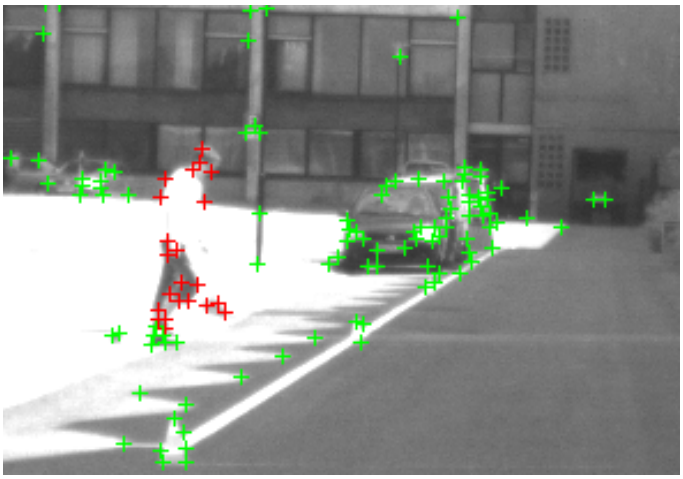


Abbildung 4: Fremdbewegung eines die Straße überquerenden Fußgängers

Modell der flachen Welt zur Detektion von Hindernissen

Zeigte der vorherige Abschnitt, wie sich bewegte Objekte detektieren lassen, so beschreibt dieser die Möglichkeit Hindernisse mit Hilfe des optischen Flusses zu erkennen.

In diesem Zusammenhang werden als Hindernisse Objekte mit einer Höhe verschieden von Null betrachtet ($h \neq 0$).

Ein Bild ist die projizierte 2D-Abbildung der realen Welt, d.h. die Tiefeninformation fehlt. Jeder Bildpunkt steht somit für einen Strahl in der realen Welt aber nicht für einen konkreten Punkt. Aus diesem Grund ist es nicht direkt möglich zu sagen, in welchem Abstand sich ein Punkt vor der Kamera befindet bzw. auf welcher Höhe h.

Ist die Position und Lage der Kamera sowie deren interne Parameter (Auflösung, Brennweite, ...) bekannt, kann dieser Strahl für jedes Pixel berechnet werden. Wird nun eine Höhe $h = h_0$ vorgegeben, so ist zu einem Bildpunkt der entsprechende Weltpunkt in dieser Ebene bekannt.

Damit kann seine aufgrund der Eigenbewegung sichtbar werdende Verschiebung zwischen zwei aufeinander folgenden Bildern berechnet werden. Diese Verschiebung ist diejenige, die der Weltpunkt hätte, wenn er auf der Höhe h_0 liegen würde.

Ein Vergleich der auf diese Weise berechneten Verschiebung und der über den optischen Fluss geschätzten Verschiebung ermöglicht es damit, zu schlussfolgern, ob sich der Punkt tatsächlich auf der Höhe h_0 befindet oder nicht. Dieser Sachverhalt wird in Abb. 5 verdeutlicht. Der Blau-Grüne „Strahl" der von der Kamera aus auf die Straße fällt, setzt sich aus zwei Teilen zusammen. Der blaue verläuft bis zum Hindernis auf der Höhe $h = h_{\text{Obstacle}}$. Der grüne Anteil stellt den Unterschied dar, wenn kein Hindernis vorhanden wäre. Grafik 6 zeigt den entsprechenden Unterschied beim optischen Fluss, ebenfalls in grün und blau kodiert.

Für die Hinderniserkennung würde $h_0 = 0$ angenommen werden und somit alle Punkte, die eine Verschiebung ungleich der theoretisch für h_0 berechneten aufweisen, als Hindernisse markiert.

Selbstverständlich würden damit auch Punkte markiert, die sich in der Welt zwischen den Aufnahmen der beiden Bilder bewegt haben. Das ist jedoch erwünscht, denn wenn sich ein Punkt eines Objektes bewegt, ist dieses auch ein Hindernis [FJW+08].

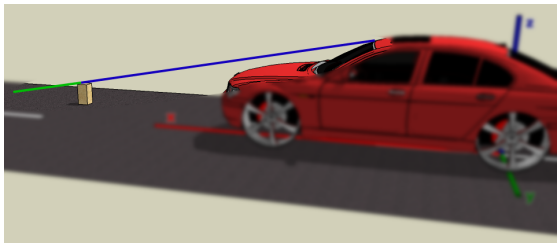


Abbildung 5: Strahlenverlauf mit und ohne Hindernis

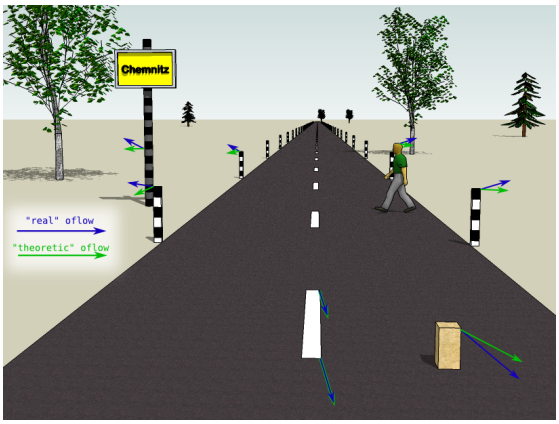


Abbildung 6: optischer Fluss mit und ohne Hindernis

4 Bewegungsklassifikation

Mit den bisher beschriebenen Verfahren zur Bewegungs- und Objektdetektion sind die Möglichkeiten den optischen Fluss auszuwerten noch nicht erschöpft. Mit der Bewegungsanalyse lassen sich die Objekte sogar klassifizieren.

Einerseits kann die Bewegung zwischen 2 Bildern ausgewertet und auf deren statistische Merkmale hin untersucht werden, andererseits kann auch eine Zeitreihe aufgestellt und das Bewegungsverhalten über mehrere Aufnahmen hinweg klassifiziert werden.

Ersteres erlaubt, wie in [FJW09] detailliert beschrieben, die Unterscheidung in starre oder nicht-starre Bewegung, was hauptsächlich durch Untersuchungen der Bewegungsrichtung erfolgt. Bei starren Objekten wie einem Fahrzeug, wird die Richtung wesentlich weniger variieren, als bei nicht-starren Objekten wie einem Menschen.

Die Zeitreihenanalyse erlaubt eine Beobachtung des Bewegungsverhaltens über die Zeit und ermöglicht somit das Erkennen der Periodizität des menschlichen Gangs [FSWG06]. Eine entsprechende Zeitreihe und ihr Leistungsspektrum ist in Abb. 10 dargestellt und zeigt deutlich die vorhandene Periodizität.

5 Schlussfolgerung

Der optische Fluss als ein wichtiges Werkzeug der Bildfolgeauswertung wurde im Zusammenhang mit einigen Möglichkeiten der Bewegungsanalyse vorgestellt. Unter jenen insbesondere die zur Detektion von Objekten, als auch zur Klassifikation derer.

Auch wenn damit die Einsatzmöglichkeiten des optischen Flusses oder der Bewegungsanalyse noch nicht erschöpft sind, ist deren weitreichende Bedeutung gezeigt geworden.

Weitere Verwendungen des optischen Flusses, die nicht angesprochen wurden, sind u.a. die Videokomprimierung, „Slow Motion“ und Rückgewinnung von Tiefeninformation.

Abbildung 7: Bewegungsvektoren sind gut ausgerichtet

Abbildung 8: Bew.vektoren streuen in ihrer Richtung

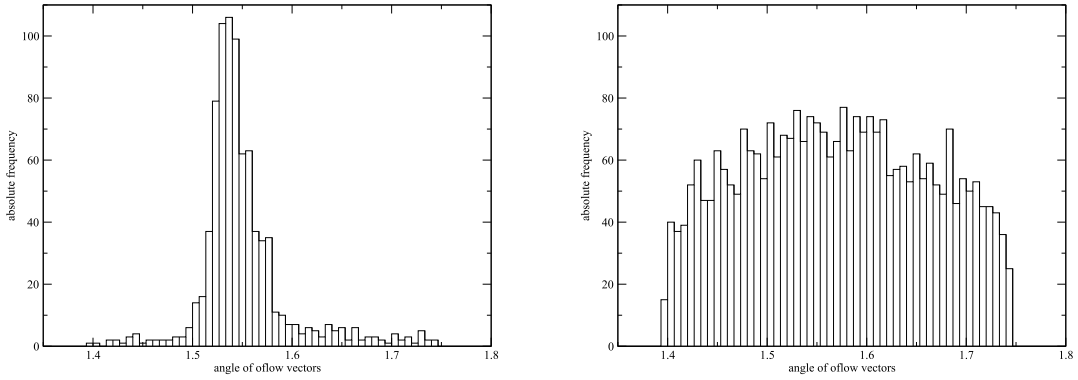


Abbildung 9: Richtungshistogramm zeigt Unterschied zwischen starrer (links) und nicht-starrer (rechts) Bewegung

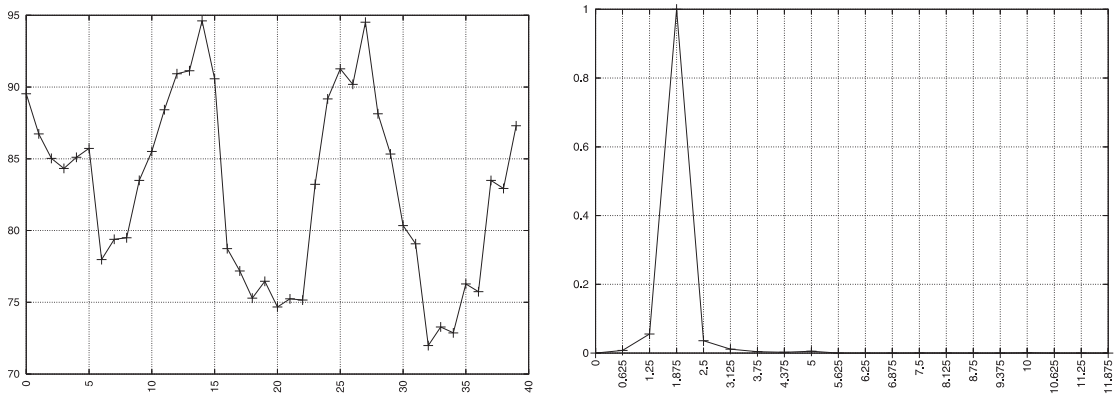


Abbildung 10: Zeitreihe und normiertes Leistungsspektrum der Bewegungsrichtung (aus [FSWG06])

Literatur

[Ana89] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989.

[BB95] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27:433–467, 1995.

[Bou02] Jean Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm, 2002.

[FJW+08] B. Fardi, T. John, H. Weigel, M. Walessa, and G. Wanielik. Unobstructed space recognition with a monochrome camera. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 577–583, June 2008.

[FJW09] Basel Fardi, Tobias John, and Gerd Wanielik. Non-rigid-motion recognition using a moving mono camera. In *Proc. Intelligent Vehicles Symposium 2009*, 2009.

[FSWG06] B. Fardi, I. Seifert, G. Wanielik, and J. Gayko. Motion-based pedestrian recognition from a moving vehicle. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 219–224, 2006.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1-3):185–203, 1981.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, March 2004.

[LK81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[ON95] Michael Otte and Hans-Hellmut Nagel. Estimation of optical flow based on higher-order spatiotemporal derivatives in interlaced and non-interlaced image sequences. *Artif. Intell.*, 78(1-2):5–43, 1995.

[Ric04] Dirk Richter. Bewegunsvektorschätzung zur Objekterkennung. Master's thesis, Chemnitz University of Technology, 2004.

Aspekte zur Archivierung audiovisueller Unterlagen im Sächsischen Staatsarchiv

Stefan Gööck

Sächsisches Staatsarchiv
Archivzentrum Hubertusburg
Sachgebiet Audiovisuelle Medien

stefan.goeoeck@sta.smi.sachsen.de

Zusammenfassung: Die Gefährdung audiovisueller Unterlagen resultiert nur zum Teil aus der Fragilität der Medien selbst. Soziale Phänomene, Marktinteressen und Vorurteile führen zu vorzeitiger Obsoleszenz. – Der gesetzliche Auftrag des Sächsischen Staatsarchivs schließt Filme, Ton- und maschinenlesbare Datenträger in die Gesamtüberlieferung ein. Jedoch führt die Spezifik der regionalen und lokalen Medienproduzenten zu einem anderen Profil der Medienarchivierung als in den großen deutschen Medienarchiven, deren Arbeitsweisen und Standards somit nur als Orientierung dienen können. – Im Archivzentrum Hubertusburg / Wermsdorf hat der Freistaat Sachsen weiter verbesserte Kapazitäten für die Übernahme, fachgerechte Bewertung, benutzerfreundliche Erschließung und materialgerechte Sicherung audiovisueller Unterlagen geschaffen.

Schlagwörter: Audiovisuelle Unterlagen, Landesarchiv, Bewertung, Erschließung, Benutzung, Sachsen

1 Tatsächliche Ursachen besonderer Gefährdung audiovisueller Unterlagen

Eingangs einige Bemerkungen zur generellen konservatorischen Herangehensweise eines Archivs und zu den alten wie neuen Vorurteilen, mit denen wir uns in der Öffentlichkeit und auf Seiten der öffentlichen Hand auseinandersetzen müssen.

Denn viele meinen es zu wissen: Fotografische Glasplatten sind bruchgefährdet, Schellacks nicht minder. Nitrofilm hat ganze Säle abgebrannt, auch Vinyl-Schallplatten legt man besser nicht auf die Heizung, das Spulentonband klingt nur auf dem Ursprungs-Gerät gut. Die Musikkassette verfitzt sich im Rekorder, für Großvaters 8-mm-Filme werden keine Projektoren mehr gebaut, für Vaters Hi8-Kassetten keine Video-Laufwerke. Und leider auch jüngst keine Entwarnung: Gebrannte Opto-Discs verweigern sich fremden Laufwerken, Schutzverletzungen drohen allenthalben. Man könnte meinen, nur was man schwarz auf weiß besitze, könne man getrost nach Hause tragen: Das Audiovisuelle ist eben fragil, nicht von Dauer, ungeeignet für die Ewigkeit, insofern flüchtig, für den Tag bestimmt, vernutzt sich, verschwindet zuletzt im Digitalen Desaster – so der Titel einer NDR-Fernsehproduktion aus dem Jahr 2005, die u. a. den Verfall von Audio-CD in der Deutschen Nationalbibliothek thematisiert.

Die Probleme mit der Langzeit-Erhaltung überlieferter audiovisueller Unterlagen sind in der Tat erheblich; dennoch ist zu bezweifeln, dass sie wirklich grundsätzlich größer und schlechter zu bewältigen sind als bei konventionellem Archivgut, das traditionell in analoger Form auf dem Träger Papier überliefert ist. Die seit der industriellen Revolution im 19. Jahrhundert erzeugten Papiere vergilben, zerbröseln, müssen aufwändig stabilisiert werden, um Verluste zu vermeiden. Und die moderneren, hektographierten wie thermokopierten Texte verschwinden in kürzester Zeit spurlos, sind nur in Kopie zu erhalten. Demgegenüber existieren relativ stabilere Magnettonbänder, die selbst nach einem halben Jahrhundert abspielbar geblieben sind, und hundert Jahre alte Kinofilme. Insofern besteht keine grundsätzlich andere Gefährdung audiovisueller Medien im Vergleich zu den konventionellen.

Als generelle Lösung für alle Probleme der Bestandserhaltung im Archiv ist heute oft zu hören, die Digitalisierung werde es richten, obwohl diese Erwartung offensichtlich der Erfahrung widerspricht, dass mit der Etablierung neuer Konzepte und Technologien auch neue Gefährdungen generiert werden. Das Beispiel Audio-CD wies ja schon in diese Richtung. Deshalb ist aus Sicht der Archive zu verlangen, die medientechnische ebenso wie die IT-Entwicklung kritisch zu hinterfragen und rechtzeitig tragfähige Entscheidungen hinsichtlich dauerhafter Stabilität und Authentizität zu verlangen – was derzeit zwingend darauf hinausläuft, die Originale weiter zu sichern. Jedoch wird diese Forderung zur Erhaltung von Kulturgut zunehmend konterkariert vom Internet- und Digitalisierungs-Hype.

1.1 Forcierter Innovationsdruck als soziales und Marktphänomen

Jeder Hype hat seine Zeit. Das galt für den photochemischen Prozess, den Phonographen, das Kino und das Radio, vorgestern für den Schmalfilm, gestern für die Videografie, und gilt fort mit dem, was aktuell angesagt ist. Subjektiv sind es an vorderster Front die Fortschrittsgläubigen und Technophilen, die sich am Machbaren begeistern und dafür das Althergebrachte aufzugeben bereit sind. Bereits Anfang der 1970er Jahre hörte ich Vorlesungen zur Datenverarbeitung, worin u. a. das papierlose Büro angekündigt wurde. Trotz mittlerweile erfolgten massenhaften Durchbruchs der IT in den Alltag weltweit harrt das papierlose Büro ironischerweise noch immer seiner Durchsetzung. Die audiovisuellen Medien hingegen, ohnehin eng verknüpft mit der Technologie, sind zügig der Digitalisierung verfallen: Zunächst, wegen des geringen Speicherbedarfs, Audio-Produktion und -vertrieb einschließlich Hörfunk und Schall-Archiv; derzeit ist die Umstellung der fernsehbasierten Produktion, Distribution und Archivierung im Gange. Entscheidender Anstoß ist hier die Umstellung auf bandlose Akquisition. Seitdem einige Camcorder-Fabrikate direkt auf IT-Medien aufzeichnen und sofortiges Andocken an digitale Video-Schnitt-Systeme zulassen, erscheint es antiquiert, zur Archivierung auf Videokassette auszuspielen, wie bisher üblich.

Beim Kinofilm verbleibt nach der Einführung digitaler Hilfs-, Manipulations- und low-budget-Aufnahmeverfahren als letzte traditionelle Bastion der analoge Filmvertrieb, den das digitale Kino in seinem Konstituierungsprozess zu überwinden trachtet. Hierfür zeichnen sich Konstrukte ab, die ganz ähnlich beim E-Book zur Anwendung kommen: Der „Content“ wird eingebunden in eine proprietäre, möglichst verplombte Black Box, um die wirtschaftliche Auswertung zu globalisieren, die Vertriebskosten zu senken und den Kunden zu binden, vor allem aber die Wahrung der Urheber- und Leistungsschutzrechte bei den jeweiligen Providern zu bündeln – also bei den Betreibern der E-Book-Portale bzw. den Kino-Major-Companies.

Trotz solcher Bestrebungen zur Nutzungsbeschränkung und Zentralisierung wird der digitale Umschwung allgemein akzeptiert, weil er modernisierte Gebrauchswerteigenschaften erwarten lässt. Daneben folgt er normativen Vorgaben (wie z. B. den HD-TV-Formaten), ist somit ggf. förderwürdig, entspricht dem digitalen Main-Stream und soll dementsprechend werblich genutzt werden. Dies, obwohl der digitale Umschwung zumindest in den ersten Schritten keineswegs wirtschaftlich oder zum Nulltarif realisierbar ist. So könnte aktuell die Neubeschaffung einer zertifizierten digitalen Kino-Projektionsanlage als wirtschaftlich irrational gelten, da die Amortisierung während der Standzeit des Systems in vielen Fällen fraglich sein dürfte. Während sich nämlich konventionelle Kinofilm-Projektoren nahezu unbegrenzt aktualisieren lassen, dürfte digitale Kino-Technik ähnlich kurzatmig sein wie die sonstige IT.

Insofern ist es nicht verwunderlich, dass der Fortschritts- und Technologie-Glaube doch der Stimulation durch die System-Anbieter bedarf. Einschlägige Produkt- und Imagewerbung, Promotiontouren, Rücknahmeaktionen für Altgeräte und Rabatte vermitteln dem Anwender: Die Zukunft ist digital, vom Mediaplayer bis zur Waschmaschine, vom Telefon bis zum Kühlschrank, und wer noch zögern wollte, den Innovationsschub umgehend mit zu finanzieren, verfiele als Ewig-Gestriger der Lächerlichkeit. Im Hintergrund wirkt eine scharfe Klinge, mit der die Hersteller überzeugend hantieren: 10 Jahre nach Produktionsende einer Studiotechnik-Komponente werden die Ersatzteile und Service-Handbücher monopolartig zurückgezogen. Je komplizierter ein System, umso wirksamer ist diese Maßnahme. So ist es heute kaum noch möglich, Videolaufwerke jener Studio-Systeme funktionstüchtig zu halten, die bis Anfang der 1990er Jahre weltweit gebräuchlich waren (U-Matic). Demgegenüber sind die relativ simpleren, meist verteilt gelabelten Consumer-Systeme weniger gefährdet, wie an Musik-Kassette oder VHS-Video zu sehen, die noch immer „lebende Systeme“ sind, obwohl sie inzwischen im Ramsch-Regal der Discounter lagern.

1.2 Das populäre Missverständnis aktueller Medien-Technologien

Die Welt der Medientechnologie besteht doppelt: Einerseits gibt es die Welt der Endverbraucher-Systeme, die für private Zwecke angeboten werden, darin die Video-DVD bzw. BluRay-Disc und der MP3-Player. Vom Consumer meist unbemerkt, gibt es

daneben die aufwändige Welt der Studiotechnik, die zur Medienproduktion unerlässlich ist. Beide Welten sind aufeinander bezogen, indem die professionell erzeugten Inhalte für die Wiedergabe mit Consumertechnik bestimmt sind. Das bedeutet aber nicht, das die verwendeten Konzepte einander entsprächen, und selbst innerhalb der großen Medienkonzerne erscheinen beide Welten streng geschieden. Der Unterschied ist folgender: Weil Endverbraucher-Systeme allgemein nicht zur Medienproduktion bestimmt sind, können Miniaturisierung und Datenreduktion weiter vorangetrieben werden. Hingegen arbeiten professionelle Medien- und Speichertechniken mit geringerer Datenreduktion, also höherer Redundanz, sind auf Produktionsbedürfnisse, darunter auf Sicherheit optimiert, weshalb auch sehr hohe Preise am Profi-Markt durchsetzbar sind.

Da aber diese „Profi-Welt“ quasi unsichtbar ist, kann es vorkommen, dass Entscheider z. B. aus dem Bereich von Haushalt und Verwaltung dem Irrtum aufsitzen, man könne günstige Allgebrauchs-Technologien für professionelle Zwecke adaptieren und so beträchtliche Rationalisierungs-Effekte, etwa im Archiv, generieren.

Im Gegenzug werden gelegentlich in der Archiv-Welt Retrodigitalisierungs-Projekte als Vorhaben zur digitalen Langzeit-Sicherung dargestellt, ohne tatsächlich dazu taugliche Komponenten zu enthalten. In diesen Fällen liegt die Vermutung nahe, dass entweder der Digitalisierungs-Hype zur Erlangung einer Finanzierung benutzt wird, um z. B. die digitale Benutzung voranzubringen – oder ein irrationales „Prinzip Hoffnung“ wirkt.

1.3 Die archivarische Bevorzugung konventioneller Quellengattungen

Der klassische Archivar ist traditionell vor allem dafür ausgebildet, mit schriftlicher Überlieferung umzugehen. Seine Spezialität besteht darin, Verwaltungsstrukturen zu durchdringen, Registraturen, Aktenvorgänge und Texte schnell zu erfassen, ihren Wert als mögliches Archivgut einzuschätzen, nachvollziehbare und benutzbare Bestände zu formieren, schließlich deren Inhalt wertfrei und bündig zu verzeichnen, um den weiteren Zugriff der Behörden auf ihre Akten zu ermöglichen sowie Historiker, Juristen und Familienforscher kompetent zu bedienen. Sind jedoch in den Beständen nontextuale Medien enthalten, z. B. Rede-Mitschnitte, oder gar nonverbale, man denke neben Bildern an Musik, entstehen dem Archiv manuelle Aufwendungen zur Interpretation, versagen selbst neuere IT-gestützte Rechercheverfahren. Im Übrigen wird der typische Benutzer im „Papier-Archiv“ nur ausnahmsweise andere Quellengattungen erwarten. Denkt man die medientechnische Hürde hinzu, die jeder maschinenlesbare Träger mit sich bringt, wird schnell klar, warum die Archivierung audiovisueller Unterlagen zumindest in der Vergangenheit überwiegend Spezial-Archiven überlassen wurde.

1.4 Inflation und Trivialisierung der Medienproduktion

Diese Enthaltsamkeit wurde mit Heraufziehen des Medien-Zeitalters zunehmend problematisch, sollte auch künftig ein zutreffendes Bild der gesellschaftlichen Verhältnisse und des staatlichen Handelns im Archiv nachvollziehbar sein. Jeder Besitzer eines Foto-Handys sei ein potentieller Film-Regisseur, suggerierte die Werbung. Und alles, was medientechnisch möglich erscheint, wird früher oder später auch von den Behörden praktiziert, also von den Bestandsbildnern vieler Archive. Seit den 1950er Jahren liegen uns erste Tonband-Mitschnitte vor, seit den 1980ern finden sich in größerer Zahl Tonband-Kassetten. Ende der 1980er setzt auch in der örtlichen DDR-Überlieferung das Medium Video ein. Ab 1989/90 ist eine regelrechte Inflation der Verwendung audiovisueller Möglichkeiten festzustellen, weil die DDR-staatliche Deckelung ebenso entfiel wie die Abschottung vom Weltmarkt. Schlagartig war die professionelle Medienproduktion nicht mehr den elitären DDR-Staatsfirmen - typisch mit Sitz in Berlin - vorbehalten. Behörden, mehr noch zahllose kleine Produzenten, viele darunter Newcomer und wirtschaftlich schwach, „machen“ seitdem mit jeder Art von Technik in bunter Vielfalt Rundfunk, Fernsehen, Film und Multimedia, zuweilen mit geringem inhaltlichen und handwerklichen Anspruch, meist ohne Ressourcen und Ambition zur Archivierung ihrer Werke.

1.5 Der Zwang zur zyklischen Erneuerung der Medientechnologie

Die großen Medienproduzenten, wie die öffentlich-rechtlichen Fernsehanstalten, erwarteten konventionell 20-Jahres-Zyklen für ihre Medientechnik. Wer demnach in der Nachwendezeit mit Medienproduktion oder auch nur mit audiovisuellen Mitschnitten begonnen hat, dürfte den End-of-Support seines Systems spätestens jetzt erreichen, wenn er nicht überhaupt mit billiger Uralt-Technik am Start war und so schon viel früher seinen ersten Medien-Bruch vollziehen musste. Idealerweise wäre zuvor der Produktions-Bestand in das neue Format umzukopieren gewesen. Wo die Kraft hierzu nicht ausreichte, ist der Zugang zu älteren Produktionen bereits verloren gegangen. Um dem entgegenzuwirken, muss Medienarchivierung vergangene Technik-Generationen funktionstüchtig vorzuhalten trachten.

1.6 Physikalisch-chemische Zerfallsprozesse der Trägermedien

Absichtsvoll seien die Zerfallsprozesse der Träger audiovisuellen Contents erst an letzter Stelle genannt: Hier handelt es sich um objektiv unvermeidliche, jedoch durch geeignete Klimatisierung verzögerbare Faktoren. Der Zahn der Zeit nagt meist an mehreren Schwachstellen. Feuchte und Wärme sind ebenso wie energiereiche Strahlung generell von Übel, magnetische Aufzeichnungsschichten können angelöscht werden, insbesondere Gelatine-Schichten bilden biologische Nährböden. Zellulose-Ester als Unterlage der Aufzeichnungsschicht, wie beim photochemischen Nitro- oder

Acetat-Film, ebenso bei gewissen Magnetton-Typen, entfalten autokatalytische Dynamik bei ihrem hydrolytischen Zerfall.

Relativierend sei angefügt: Da das Virtuelle gleichfalls eines Trägers bedarf, schaden etliche der oben angefügten Störfaktoren ähnlich auch der IT-Archivierung.

2 Die Spezifik der AV-Archivierung in einem Landesarchiv am Beispiel Sachsen

Die großen deutschen Medienarchive, wie die Abteilung Filmarchiv des Bundesarchivs, die Stiftung Deutsches Rundfunkarchiv und die Archive der Landesrundfunkanstalten, verwahren in großem Umfang eine jeweils relativ homogene, von zuverlässigen Metadaten begleitete Überlieferung in professionellen Medienformaten, die sie typisch auf direktem Wege erreicht. Jedoch fällt weder Kinofilm-Archivierung noch Fernsehproduktion jeglicher Herkunft zwangsläufig in die Zuständigkeit eines deutschen Landesarchivs.

2.1 Behördliche, sozial-kulturelle, kleingewerbliche und private Quellen

Entsprechend dem sächsischen Archivgesetz (1993) steht die Überlieferung der Landesbehörden an erster Stelle.

Dementsprechend fallen im Bereich der Politik Rede- und Interview-Mitschnitte an, die zum Teil aus presserechtlichen Gründen oder zunächst als Grundlage einer späteren Verschriftlichung mit nicht-professionellen Mitteln hergestellt werden. In der Bau-Verwaltung werden dokumentarische Aufnahmen und Image-Filme gefertigt. Die Landespolizei verfügt über audiovisuelle Mittel bis hin zu professionellen Standards, um Einsätze zu dokumentieren und Lehrfilme herzustellen, die mit anderen Bundesländern ausgetauscht werden. Das sächsische Landtags-Archiv verwahrt Videomitschnitte aller Plenarsitzungen seit Neubeginn in einem Consumer-Format, ist allerdings dem Staatsarchiv nicht anbietungspflichtig. Hingegen fällt die privatrechtliche Fernseh- und Hörfunkproduktion in die Zuständigkeit einer staatlichen Genehmigungs- und Aufsichtsbehörde, deren Mitschnitte und Belegkopien dem Staatsarchiv zustehen.

Bestandsergänzend bemüht sich das Sächsische Staatsarchiv um solche audiovisuellen Quellen, deren Medienproduktion aus staatlichen Mitteln gefördert wurde. Dies ist einerseits im Bereich der Sozial-, Kinder- und Jugendarbeit der Fall, andererseits bei konkreten Projekten professioneller Medienproduzenten außerhalb der öffentlich-rechtlichen Hörfunk- und Fernsehanstalten.

Auch aus Privathand oder von gewerblichen Medienproduzenten kommen Medien-Übernahmen in Frage, wenn es der Sicherung kultureller Werte oder sonstigen Interessen des Freistaates Sachsen dient. Insbesondere in Folge des Umbruchs 1989/90 wurden in z. T. beträchtlichem Umfang audiovisuelle Unterlagen durch Privatinitiative gesichert und später dem Staatsarchiv übergeben. Gerade solche Bestände machen wegen ihres Umfangs und ihrer Bedeutung heute den Kern der audiovisuellen Überlieferung in Sachsen aus (Filmstudio der DDR-Landwirtschaftsausstellung AGRA, Zentrales Amateurfilmarchiv der DDR, Bleichert- / VTA-Filme, Bezirksfilmstudio Leipzig).

2.2 Bewertungsansatz

Während seltene, ältere Unterlagen kaum selektiert werden, ist für die jüngere Überlieferung eine bewertende Auswahl unumgänglich, um den Aufwand zu begrenzen.

Ein Bewertungsansatz für AV-Unterlagen, der im Sächsischen Staatsarchiv entwickelt wurde, geht von verschiedenen Überlieferungstypen aus. Dabei haben Produktionsbestände, welche direkt mit der Entstehung der audiovisuellen Aufzeichnungen verbunden sind oder waren, eine Sonderstellung. weil sie hohe Authentizität versprechen, eine Klärung verwertungsrechtlicher Fragen i. d. R. möglich sein wird und weil frühe Kopiergenerationen / hochwertige Medienformate (darunter ggf. Originalnegative bzw. Master-Tapes) erwartetet werden können. Entsprechend kommt Verleih-Beständen, die nur Vertriebsmedien enthalten, sowie Mediensammlungen und Einzelstücken eine abgestufte Bedeutung zu.

Allerdings haben sich pauschalisierte Bewertungsverfahren als eher problematisch erwiesen. Auch wegen des u. U. hohen materiellen Wertes der audiovisuellen Überlieferung ist eine Einzelfall-Entscheidung vorzuziehen.

2.3 Ausschluss von Doppelüberlieferung zum öffentlich- rechtlichen Hörfunk und Fernsehen

Die allgemein bestehende Erwartung, man könne im Landesarchiv die Sendungen der Landesrundfunkanstalt recherchieren, trifft in Sachsen wie in den meisten deutschen Bundesländern nicht zu. Stattdessen betreiben die öffentlich-rechtlichen Anstalten eigene Archive, die zunehmend eng mit dem Sendebetrieb, der weiteren Programmproduktion und den Verwertungsrechten verbunden sind.

2.4 Low-Tech unterhalb professioneller Standards

Jenseits der medientechnischen Standards der DDR-Staatsmedien bzw. des Mitteldeutschen Rundfunks (MDR) verbleibt dem Sächsischen Staatsarchiv eine inhomogene, überwiegend in semiprofessionellen bzw. Consumer-Medienformaten

erzeugte AV-Überlieferung, die häufig mangelhaft dokumentiert und mit Problemen jeglicher Art behaftet ist.

So verfügen wir im kinematografischen Bereich über zahlreiche 16-mm-Filme als vertonte Umkehr-Originale, als Umkehr-Kopien und als Vertriebsstücke, darunter oft mit Magnetton-Randspur. Typische Schwierigkeiten bereiten offene Klebstellen, Materialschrumpfung und Ablösung bzw. Verformung der Magnetpiste.

Im Audio-Bereich ist bei uns jegliches Format seit dem Magnettonband mit 76 cm/s bis hin zur Audio-CD überliefert, jedoch dominieren die Consumer-Spulentonband-Formate, etwa mit 9,5 cm/s und Viertelspur. Minderwertiges Aufzeichnungsmaterial, abweichende Spurlage und allgemein Mängel der verwendeten Mikrofon- und Übertragungstechnik bereiten Probleme. Eine mit dem Nutzsignal aufgezeichnete Störung ist auch bei bester heutiger Wiedergabetechnik ein untrennbarer Bestandteil der Aufnahme, dessen nachträgliche Tilgung unter das Thema Archivethik fiele.

Video für den Consumer existierte in der DDR bis zur Herbstmesse 1989 offiziell nicht, somit auch kaum in örtlichen Strukturen, dies im Unterschied zu den professionellen DDR-Staatsmedien (Berlin). Dennoch gab es eine Grauzone beispielsweise im Bereich von Sport, Bildung und Kultur, aus der uns z. T. heute historische Videoformate vorliegen: VCR und VHS als Consumer-Formate, U-Matic (Lowband) im semiprofessionellen Bereich.

2.5 Die AV-Erschließungsrichtlinie des Sächsischen Staatsarchivs

Aus der Handhabung audiovisueller Unterlagen im Sächsischen Staatsarchiv seit 1997 wurde, orientiert am Umgang mit der schriftlichen Überlieferung, eine AV-Erschließungsrichtlinie entwickelt. Sie geht davon aus, die überlieferten physischen Objekte (-„Stücke“) den übergeordneten Inhalten (-„Titeln“) zuzuordnen. Unter Wahrung der Provenienz sind Konvolute zu bilden, die alle vorhandenen Träger von Teil-Inhalten des jeweiligen Titels bündeln.

2.6 Sicherungskonzept

Damit wird es möglich, die Bedeutung jedes einzelnen Stücks für die Überlieferung des Titels aufzuklären, bis hin zu der Frage, welche Stücke gesichert werden müssen und welche nicht. Grundsätzlich orientieren wir uns am Sicherungsstandard des Bundesarchiv / Filmarchiv: Der Sicherung dienende Stücke sind für die Benutzung gesperrt, im Sicherungspaket muss jeder Teilinhalt (Bild / Ton) doppelt überliefert sein. Bei notwendigen Sicherungs-Kopierungen wird möglichst das Ursprungs-Format gewahrt, anderenfalls werden anerkannte professionelle Nachfolge-Formate, möglichst ohne Datenreduktion, gewählt. Für Audio-Originale erscheint es schon heute

unvermeidlich, zu digitalen Datei-Formaten überzugehen. Zur Benutzung kinematografischer Inhalte haben wir aus Kostengründen von Anfang an keine Filmkopien, sondern Videofassungen erzeugt.

2.7 Benutzungspraxis und –perspektive

Tatsächlich finden sich unter unseren Benutzern kaum cineastische Puristen, die Wert darauf legen würden, am Verschleiß des Originals im Projektor teilhaben zu können – wie es leider manche Kunstfreunde sehen. Wir legen Video- und Audio-Benutzerkopien derzeit als VHS-Video oder als Musikkassette vor, wobei im Video der Timecode des Masterbands als Referenz einkopiert wurde. So wird es möglich, mit wenig Aufwand in allen Abteilungen des Sächsischen Staatsarchivs provenienzgerecht in AV-Unterlagen zu recherchieren, ohne mehrere Studiotechnik-Einheiten unterhalten zu müssen.

Mittelfristig streben wir an, die analogen Benutzerstücke durch Benutzungs-Digitalisate, konkret Mediendateien, zu ersetzen. Dies dürfte der Mehrzahl unserer AV-Benutzer, die aus dem Bereich TV-Produktion kommen, noch besser entsprechen.

Wir fertigen für professionelle wie private Interessenten Kopien vom Master im gewünschten Zielformat – meist geht es um DigiBeta bzw. Video-DVD, vorausgesetzt, die verwertungsrechtlichen Fragen wurden geklärt und eine dementsprechende Nutzungsvereinbarung abgeschlossen. Die Sächsische Archivgebührenordnung differenziert zwischen Nutzungsarten, die ganz oder überwiegend im öffentlichen Interesse liegen, sowie rein geschäftlichen Interessen. Ferner wird unterschieden, ob das Archivmaterial nur örtlich, regional oder national / international verwertet werden soll.

2.8 Beispiele für kinematografische, videografische, Audio- und multimediale Archivalien

Drei viel beachtete AV-Archivalien des Sächsischen Staatsarchivs seien exemplarisch vorgestellt.

Aus dem Jahre 1912 stammt ein 35-mm-Nitromaterial, das privat aufgenommen und überliefert wurde. Es zeigt den deutschen Kaiser und den sächsischen König auf der Durchreise in Coswig bei Meißen, auf dem Wege zum Kaisermanöver. Im Jahr 2002 erhielten wir das Originalmaterial auf dem Wege der Schenkung, übernahmen die Herstellung eines Sicherungspakets, und stellten es 2007 an den Anfang unserer Video-DVD-Veröffentlichung „Land, Leute und Maschinen – Film in Sachsen 1912 – 1940“.

Im Jahre 1953 wurden, nach den Juni-Ereignissen, auch in Leipzig Schauprozesse gegen die mutmaßlichen Rädelsführer veranstaltet. Der Leipziger Stadtfunk strahlte zwei propagandistische Audio-Reportagen hierzu aus, die auf Magnettonband im Format 76 cm/s in Vollspuraufzeichnung erhalten sind. Diese Tapes gelangten als

Bestandteil des AV-Depositums des Leipziger Stadtarchivs zu uns, wurden digitalisiert, die enthaltenen Pegelsprünge vorsichtig angepasst und schließlich eine Audio-CD gefertigt, die als Bestandteil der Leipziger Gemeinschafts-Ausstellung zum 50. Jahrestag des 17. Juni 1953 im Dauerbetrieb lief.

Seit 2009 recherchieren TV-Redaktionen erstmals Material aus der Zeit nach 1989. Wir konnten verweisen auf „Pennhouse TV“, ein medienpädagogisch angelegtes Videomagazin von Leipziger SchülerInnen, das über einen Zeitraum von 10 Jahren etwa vierteljährlich auf Kassette verbreitet wurde und somit eine Materialbasis zur Entwicklung des Zeitgeist', von Themen und Haltungen bietet. Unsere hochwertigen Kopien wurden im Format Betacam-SP von den Original-Masterbändern gezogen, geliehen von der „Multimedienwerkstatt Die Fabrik e.V.“, deren Projekt vom Freistaat Sachsen und der Stadt Leipzig gefördert worden war.

Frühes Beispiel für Multimediales sind die zahlreichen Dia-Ton-Serien aus der DDR-Zeit, die typisch aus der Triade Dia-Automatenkassette, Magnettonband (mit Steuer-Impulsen für die Lichtbilder) und Beiheft bestehen. Multimedial im modernen Sinne sind diese Dia-Ton-Serien, weil sie aus lose verkoppelten Wahrnehmungsebenen bestehen, die auch losgelöst rezipiert werden können, und weil gestaltende Mitwirkung des Referenten möglich war. Der archivarische Wert dieser Überlieferung und die Frage, wo sie langfristig zuzuordnen ist, sind freilich weiter zu diskutieren.

2.9 Perioden: Vorkrieg, DDR, Nach-„Wende“

Mit diesen Beispielen ist eine mögliche Periodisierung unserer AV-Überlieferung angerissen. Aus der Zeit bis 1940 sind nur wenige kleine Bestände bei uns überliefert, so zu den Firmen Hille AG (Dresden) und Bleichert (Leipzig), ferner Einzelstücke zu Firmen wie der Heine AG (Leipzig / Riesa) und Rudolph Sack (Leipzig).

Von 1940 bis zur frühen Nachkriegszeit liegt uns keine AV-Überlieferung vor.

Aus der DDR-Zeit datieren seltene frühe AV-Archivalien um 1950; eine höhere Überlieferungsdichte setzt erst Mitte der 1960er Jahre ein. Maßgeblich dafür mag die allmähliche Überwindung der Kriegsfolgen sowie die Etablierung von öffentlich geförderten Filmgruppen an Betrieben, Kulturhäusern und Massenorganisationen gewesen sein. Neben deren sogenannten Amateurfilmen finden sich in Firmen-Überlieferung Belegkopien professioneller Werbe- und Image-Filme, allerdings seltener. Verwaltungs- und Partei-Instanzen der DDR bedienten sich über Jahrzehnte hinweg der Magnettonband-Aufzeichnung.

Die Zeit ab 1990 ist begleitet von einem sprunghaften Anstieg der lokalen und regionalen Medienproduktion.

2.10 Digitalisierung als Hilfsmittel am Beispiel Audio-Massenüberlieferung

So enthält die Überlieferung der Sächsischen Landesanstalt für neue Medien und privaten Rundfunk (SLM) einen beträchtlichen Umfang an Audio- und Video-Mitschnitten, der zumindest für das Nachwende-Jahrzehnt erhalten und erschlossen werden soll. Der Erschließungs-Aufwand wird für die Audios bewältigt, indem das Material zunächst digitalisiert und danach auf der Timeline analysiert wird. So gelingt es, Sequenzen anhand grafischer Merkmale schnell zu erkennen und die enthaltenen, relevanten Wortbeiträge anzuspringen, um sie nach Erschließungsrichtlinie zu verschlagworten. So lässt sich Audio-Material gewissermaßen „diagonal lesen“. Es wird aufgrund der sehr unterschiedlichen Programmformate und bestehender Qualitätsmängel nicht erwartet, dass absehbar eine automatische Spracherkennung anwendbar wäre.

3 Perspektiven im Archivzentrum Hubertusburg

Im Schlosskomplex Hubertusburg in Wermsdorf (Nähe Grimma) nimmt der Freistaat Sachsen seit März 2009 ein neues Archivzentrum für die Papier-Restaurierung, die Reprografie und für das Sachgebiet AV-Medien in Betrieb. Damit soll die bestehende Bündelung der AV-Archivierung im Sächsischen Staatsarchiv, bisher am Standort Leipzig-Paunsdorf, im Archivzentrum Hubertusburg auf höherem Niveau weitergeführt werden. Wie bisher stehen professionelle Film-Bearbeitungstische sowie ein Video- / Audio-Studio zur Verfügung. Daneben werden Magazine für Sonderklimate errichtet, die den Erfordernissen für Colorfilm, Nitrofilm, schwarz-weiß-Acetatfilm sowie für magnetische / optoelektronische Medien besser entsprechen sollen. Um die Möglichkeiten im Archivzentrum Hubertusburg effektiv zu nutzen, wird eine enge Vernetzung mit Medienproduzenten und -vereinen sowie Konsultation mit jeglichen Interessenten angestrebt.

FusionSystems GmbH Systeme zur Sensor-Daten-Fusion und Szeneninterpretation

Ullrich Scheunert und Basel Fardi

FusionSystems GmbH
www.fusionsystems.de

{scheunert,fardi}@fusionsystems.de

1 Transferprojekt – Ziele und Struktur

Die Gründung der FusionSystems GmbH erfolgte im Februar 2005 als Spin-Off Unternehmen der Technischen Universität Chemnitz. Der Gründung ging die gemeinsame wissenschaftliche Arbeit der Firmengründer an der Professur für Nachrichtentechnik der TU Chemnitz voraus.

Wichtigstes Ziel der Firmengründung war es, neueste wissenschaftliche Erkenntnisse der Sensor-Daten-Fusion so effektiv und schnell wie möglich markttauglich zu machen. Aussicht auf wirtschaftlichen Erfolg der Gründung haben die Firmengründer aus dem Umstand abgeleitet, dass sich in sehr verschiedenen Anwendungsgebieten die messtechnische Erfassung von Objekten, Systemzuständen oder Situationen mit gleichzeitig mehreren Sensoren zu einer üblichen Praxis zu entwickeln begann. Gleichzeitig war den Gründern bewusst, dass man nur bei gezielter wissenschaftlicher Weiterentwicklung der theoretischen Basis, dieses Gebietes die Firma erfolgreich am Markt etablieren kann. Für die Umsetzung der Projektziele wurden die folgenden drei Aufgaben formuliert:

- Entwicklung marktfähiger Produkte und Dienstleistungen auf der Grundlage der bisherigen wissenschaftlichen Forschungsprojekte
- Erschließung weiterer Anwendungsbereiche und potentieller Märkte
- Kritische Analyse der Anwendungen in Hinblick auf weitere Schritte ingenieurwissenschaftlicher Forschung und Entwicklung

Die enge Bindung an die Professur für Nachrichtentechnik wird permanent gezielt weiterentwickelt. Grundlage dafür bilden gemeinsame Kooperationsprojekte, bei denen Entwicklungsschwerpunkte bei FusionSystems liegen, wohingegen die Wissenschaftler der Professur für Nachrichtentechnik vordergründig die Grundlagenforschung betreiben. Außerdem werden gemeinsam Diplomarbeiten und Promotionen betreut. Im Jahr 2007 formierte sich die Forschungsgruppe „Multisensorgestützte Bildverarbeitung“, in der Ingenieure der FusionSystems GmbH und der Professur für Nachrichtentechnik zusammen an der Lösung praxisbezogener Forschungsaufgaben arbeiten. Ein Gründungsmitglied von FusionSystems ist seit 2008 mit einem Lehrauftrag an der TU Chemnitz betraut.

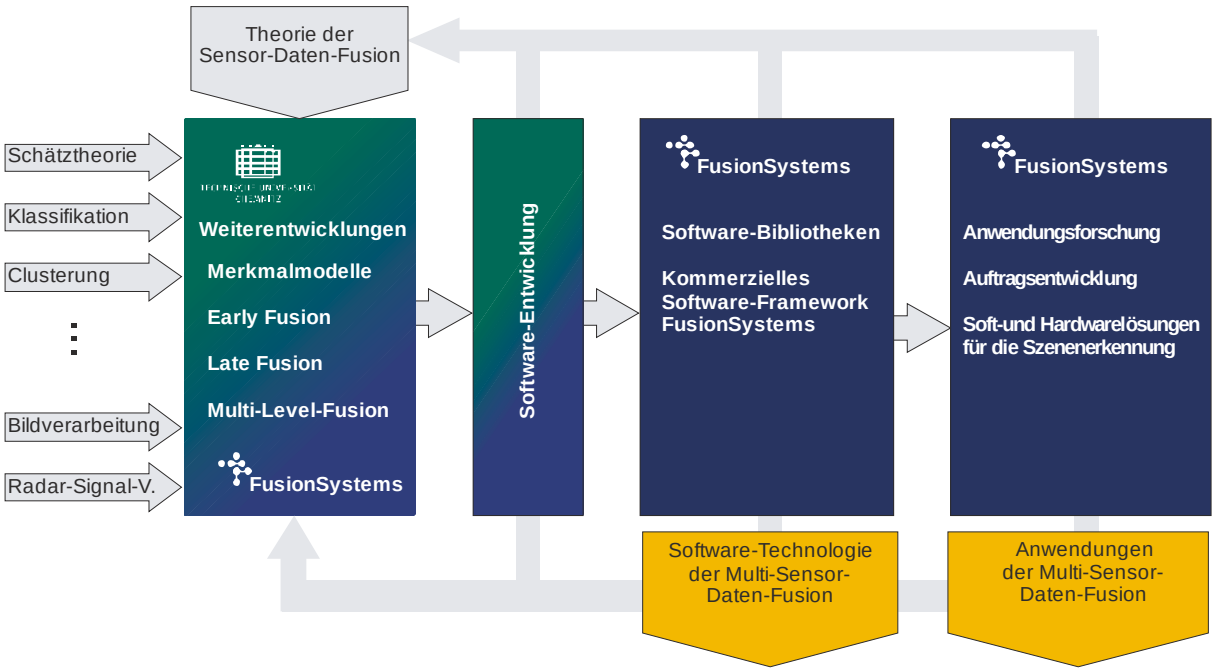


Abbildung 1: Kompetenzstruktur des Transferprojekts

2 Wirtschaftlicher Erfolg

Die FusionSystems GmbH ist Spezialist auf dem Gebiet der Multi-Sensor-Daten-Fusion sowie verschiedener sensorgeprägter Spezialgebiete wie der automatischen Bildauswertung sowie der Laser- und Radar-Signal-Verarbeitung. Unser Unternehmen FusionSystems besitzt heute umfassendes wissenschaftliches Know-how und ist in der Lage, vielfältige ingenieurtechnische Lösungen anzubieten. Produktschwerpunkte von FusionSystems sind

- Systemlösungen für die Szenenerkennung in Fahrerassistenzsystemen
- Optische Navigation von Fahrerlosen Transportsystemen
- Überwachung von Gebäuden, Geländen und Produktionsszenarien
- Industrielle Bildverarbeitung

Die wirtschaftlichen Ergebnisse und der Kundenkreis von FusionSystems sind beredtes Zeugnis für die erfolgreiche Umsetzung der Unternehmensziele. So gehören mehrere deutsche Automobilhersteller, internationale Vertreter dieser Branche, sowie eine Reihe von Automobilzulieferern zu den Kunden. Als Beispiele dafür seien die Honda R&D Europe (Deutschland) GmbH und die auch in Chemnitz ansässige IAV GmbH – Ingenieurgesellschaft Auto und Verkehr als einer der führenden Engineering-Partner der Automobilindustrie angeführt. Über den Rahmen der Fahrzeugumfelderkennung hinaus gehen Projektlösungen für andere Partner, wie die Carl Zeiss Optronics GmbH, und belegen die breite Anwendbarkeit der Ansätze von FusionSystems.

Mit Systemen wie „*NightVisionFS* – Sichere Fußgängererkennung im Fahrzeug“, „*FTSNavBox* – zuverlässige und flexible Lösung für die Navigation von Fahrerlosen Transportsystemen“ und „*ArgoGard* – Sichere Objektüberwachung“ setzte FusionSystems den Wissenstransfer auf dem Gebiet der Multi-Sensor-Systeme zur Szenenanalyse von der Technischen Universität Chemnitz in das als Spin-Off gegründete Unternehmen erfolgreich um.

2.1 NIGHTVISIONFS – Sichere Fußgängererkennung im fahrzeug

In einer Zeit mit immer höherem Verkehrsaufkommen auf unseren Straßen gewinnen Systeme, die den Fahrer beim Fahren unterstützen, immer mehr an Bedeutung. Vor allem elektronische Systeme, die den Fahrer vor Fußgängern warnen, sind für alle Verkehrsteilnehmer von großer Bedeutung.

Eine Möglichkeit, Fußgänger auch bei Dunkelheit und schlechter Sicht für den Fahrer sichtbar zu machen, bietet die FIR-Kamera. Sie erzeugt ein Wärmebild, in dem Menschen sich vom kalten Hintergrund abheben.

Damit der Fahrer aber nicht durch das Betrachten dieses Bildes vom Verkehrsgeschehen abgelenkt wird, hat FusionSystems die Software NightVisionFS entwickelt, die gefährdete Personen automatisch in Wärmebildern erkennt.

Dadurch kann der Fahrer effektiv gewarnt werden und die Aufmerksamkeit des Fahrers muss nur dann auf das FIR-Bild gelenkt werden, wenn sich wirklich eine gefährdete Person vor dem Fahrzeug befindet. Die Person wird im Bild farbig markiert, und der Fahrer kann rechtzeitig und sicher auf die Situation reagieren.

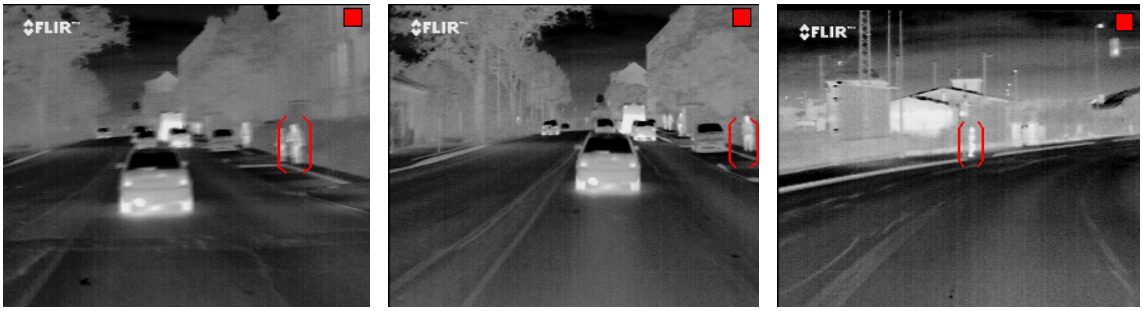


Abbildung 2: *NightVisionFS* – Beispiele für die Fußgängererkennung

2.2 FTS*NAVBOX* – Zuverlässige und flexible Lösung für Fahrerlose Transportsysteme

Im Gegensatz zu etablierten optischen, induktiven oder mechanischen Bahnführungen ist das neue System mit einem Bildsensor ausgestattet und kann auf dem Boden aufgebrachte Markierungen wesentlich sicherer auswerten als bisherige Systeme. Durch speziell entwickelte Auswertealgorithmen für die Bildmessung lässt sich die Spur auch bei Verschmutzung oder Fehlstellen sicher detektieren. Anstelle aufwendiger

Bodenarbeiten für andere Spurführungssysteme reichen nun sogar auf dem Boden aufgeklebte Führungslinien, die sich in Minutenschnelle umlegen lassen. Dadurch ist die Spurführung leichter als bisher an veränderte Bahnen in der Produktion anzupassen und hält mit ihrer Robustheit auch dem rauen Alltag stand. FusionSystems bietet hier den Kunden ein modulares System welches auch auf die Verwendung mehrerer Sensormodule erweitert werden kann.

Abbildung 3: FTS*Nav*Box und seine Komponenten

2.3 *Argo2.5D* und *ArgoGard* – Sichere Objektüberwachung

Muss eine Maschine oder Anlage während der Produktion aus Sicherheitsgründen abgeschaltet werden, kostet das oft viel Zeit und damit auch Geld. Abhilfe schafft hier das neue kamerabasierte Sicherheitssystem Argo2.5D. Es erkennt und unterscheidet zwischen erlaubten und unerlaubten Bewegungsrichtungen von Objekten. Durch die flexibel definierbaren Warn- und Schutzzonen erleichtert es die ständige Überwachung von Gefahrenbereichen an Förder- und Montagebändern, Verpackungsmaschinen sowie in den Umgebungen von Industrierobotern.

Die erlaubten Bewegungsrichtungen - wie etwa die gewünschten Bewegungen von Maschinenteilen - werden dem Kamerasystem angelernt, indem ihm diese im Beobachtungsfeld vorgeführt werden. Kommt es in der Anlage zu einer verbotenen Bewegung - zum Beispiel einer Hand in der Warnzone - kann das überwachte Montageband verlangsamt werden. Wird eine Hand in der Schutzzone geortet, wird das Band angehalten. Das sorgt für mehr Sicherheit im Arbeitsumfeld und hilft, unnötige Stillstandzeiten zu vermeiden. Zudem ist Argo2.5D praktischer und leichter einzurichten als bisherige Überwachungssysteme.

Dieses System basiert auf Forschungsergebnissen, wie sie auch zur Objekterkennung im Fahrzeugumfeld eingesetzt werden. Mit ähnlichen Verfahren wie bei der Erkennung und Positionsbestimmung von Fußgängern kann Argo2.5D die Bewegungsrichtungen im gesamten Raum in Echtzeit überwachen. Ähnliche Anwendungen aus der Argo-Produktreihe sind der Personenzähler und die Fahrkorbüberwachung in Aufzügen.

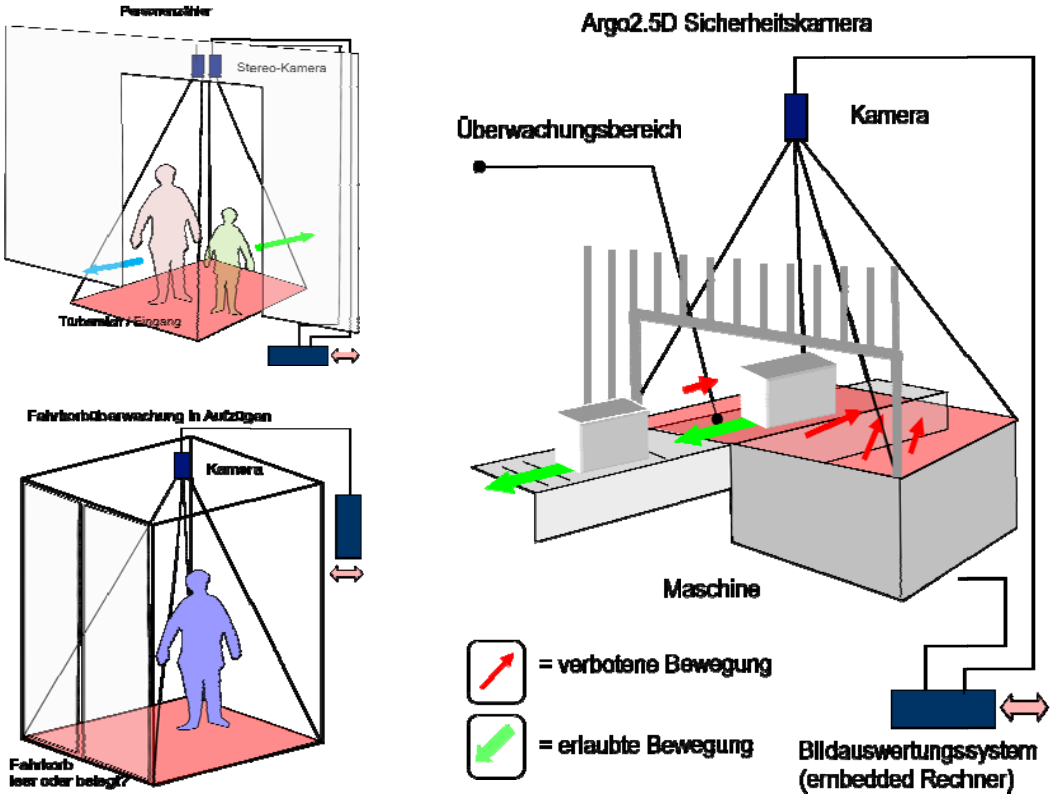


Abbildung 4: *Argo2.5D* und *ArgoGard*

3 Fazit

Die bei der Firmengründung definierten Ziele konnten bis zum heutigen Tag erfolgreich verfolgt werden. Insbesondere konnte sich die FusionSystems GmbH am Markt für multisensorielle Lösungen etablieren. Das gilt sowohl für die Fahrzeugumfelderkennung im Automobilsektor als auch für die Multi-Sensor-Daten-Fusion und Szenenerkennung generell. Aus vielen Forschungspartnern wurden Kunden für anwendungsreife Produkte und Dienstleistungen der FusionSystems GmbH; zahlreiche neue Kunden konnten hinzugewonnen werden.

Visualisierung von Prozessketten zur Shot Detection

Marc Ritter

Technische Universität Chemnitz
Fakultät für Informatik

ritm@informatik.tu-chemnitz.de

Zusammenfassung: Dieser Artikel befasst sich zuerst mit einem Framework zur Videoanalyse, der an der Professur Medieninformatik der TU Chemnitz entwickelt worden ist und betrachtet darauf aufbauend aktuelle Erweiterungen wie ein Visualisierungstool für den Workflow von Bild- und Videoverarbeitungsprozessen und die Entwicklung eines erweiterbaren Annotationstoolkits zur Erzeugung von Referenz-Annotationen für Videos. Die Anwendung des Frameworks wird letztlich exemplarisch auf dem Gebiet der modernen Szenenerkennung demonstriert.

Schlagwörter: Bildverarbeitung, Prozessketten, Shot Detection, Visualisierung, Annotation

1 Hintergründe

Die Dokumentation und Archivierung zunehmend größerer Datenmengen, besonders im Bereich der multimedialen Dokumente, stellt die moderne Forschung vor neue Herausforderungen. Einen Aspekt des Aufgabenbereichs der *Retrieval-Gruppe* des Projekts *sachsMedia*[1] *— Cooperative Producing, Storage, Retrieval and Distribution of Audiovisual Media* bildet die Extraktion und Indexierung wichtiger Informationen in Form vordefinierter Objekte aus dem Videomaterial lokaler Fernsehsender und für nachfolgende nutzergesteuerte Suchprozesse. [sac09]

Die enge Verwandtschaft der Arbeitsbereiche Sprach- und Videoanalyse und des Metadaten-Handlings legte die Entwicklung eines gemeinsam nutzbaren Frameworks nahe. Abschnitt 2 führt in die Entwicklungsumgebung *AMOPA* ein und stellt darauf aufbauend zwei weitere Software-Werkzeuge vor, die zum die Nutzung des Frameworks erleichtern und zum anderen bei der zukünftigen Annotation von Videodaten hilfreich sein sollen.

[1] Gefördert durch Unternehmen Region, der BMBF Innovationsinitiative Neue Länder

Damit sich zuvor katalogisierte Objekte überhaupt auf dem Videomaterial auffinden lassen, ist der Einsatz klassischer Verfahren zur Objektdetektion angedacht. Unglücklicherweise versagt ein großer Teil dieser Algorithmenklasse oftmals bereits bei einem abrupten Wechsel des Inhalts zwischen aufeinanderfolgenden Bildern—häufig auch als Kontextänderung oder Szenenwechsel bezeichnet. Um dem präventiv entgegen zu wirken, hat sich die vorverarbeitende Unterteilung des Videostroms in einzelne Schnitte bewährt. Abschnitt 3 betrachtet ein modernes Verfahren zur automatischen Erkennung von Schnittgrenzen (engl. *shot detection*) aus dem Repertoire des wissenschaftlichen Wettbewerbs *TRECVID* genauer und erläutert dessen Integration in unseren Framework.

2 Ein Framework zur Videoanalyse

Der als Lehr- und Forschungsinstrument konzipierte Framework *AMOPA - Automated MOving Picture Annotator* soll Entwurf und Implementierung von nahezu beliebigen prozessgesteuerten Workflow-Konzepten erlauben, wie diese im Bereich der Bildverarbeitung als Bildverarbeitungsketten (engl. *image processing chain* oder *IPC*) traditionell anzutreffen sind.

Dieser Abschnitt stellt neben dem Framework, zwei eigene darauf aufbauende Toolkits vor. Der *IPC-Editor* vermag die Entwicklungs- und Konfigurationszeit mittels einer graphischen Oberfläche zu verkürzen, die für die Erstellung einer Prozesskette notwendig ist. Dieser kann sowohl zur einfachen Annotation von Videodaten als auch zur Erzeugung von Referenz-Datensätzen genutzt werden, die die Entwicklung maschineller Algorithmen in derselben Weise als Lern- und Verifikationsbasis begünstigen kann.

2.1 Struktur

Die offene C-Bibliothek *FFMPEG* (siehe Abbildung 1) öffnet, speichert und gibt multimediale Dateien wieder. Sie wird von einer großen Community stetig weiter entwickelt und ist integraler Bestandteil erfolgreicher Software-Tools wie *Video LAN*[2]–Multimediaplayer oder *VirtualDub*[3] von Avery Lee, die beim Schnitt und Editieren von Videos Anwendung finden. Die Ansteuerung der C-Funktionen erfolgt über eine Weiterentwicklung des *Java Native Interface* (Abk. *JNI*) von Java aus, was die Open Source-

[2] http://www.videolan.org

[3] http://www.virtualdub.org

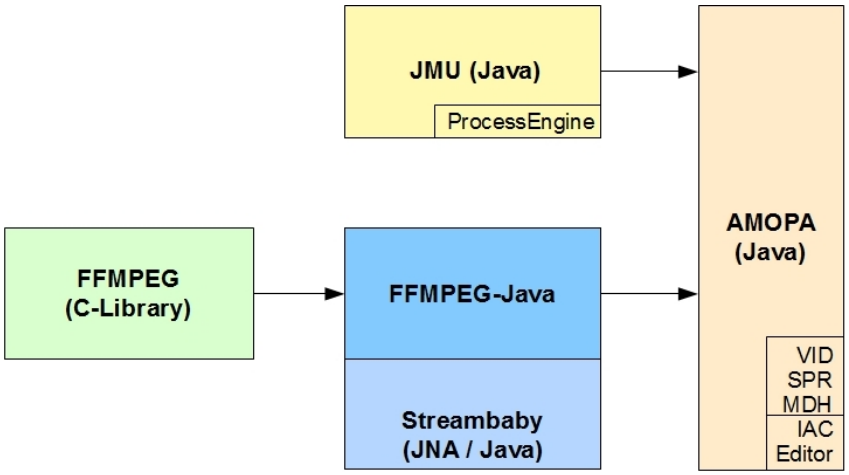


Abbildung 1: Die Aufbau des Frameworks *AMOPA* und die Zusammenschaltung seiner Subkomponenten.

Bibliothek *FFMPEG-Java* ermöglicht, die seit Ende 2007 Bestandteil des Projektes *Streambaby* ist [str09].

Die Grundlagen des Prozesskonzeptes zur Umsetzung der Workflow-Ketten liefert der Engine-Framework der preAlpha-Version des Open Source Projektes *Java Media Utility* (Abk. *JMU*), das bis 2006 von Paolo Mosna entwickelt und betreut wurde [Mos07].

Die Klasse *Engine* stellt einen Verwaltungsprozess dar, der eine Prozesskette aus einer XML-Datei erstellt und startet. Abbildung 2 illustriert den Aufbau einer verallgemeinerten linearen Prozesskette. *InputProcess* liest zu verarbeitende Daten (beispielsweise Bilder einer Videosequenz) ein und stellt diese dem nachfolgenden Verarbeitungsprozess im Rahmen des Prozesskontextes über einen internen Puffer sukzessive zur Verfügung. Eine Menge nachfolgender Nutzerprozesse (engl. *CustomProcess*) entnimmt und verarbeitet schrittweise die jeweils hinterlegten Daten. Die Verarbeitungskette endet mit dem *OutputProcess*, der keine Konsumenten besitzt.

Alle Prozessglieder werden als Instanz von *EngineProcess* als einzelne Threads gestartet und bilden somit eine ideale Basis um Multi-Core-Prozessorsysteme optimal auszulasten. Das Abstraktionsniveau der Bildverarbeitungskette und der Einzelprozesse bleibt komplett dem Nutzer überlassen.

Jede Instanz von *EngineProcess* verfügt über Grundfunktionalitäten, die im Laufe der Lebensdauer des Objekts nacheinander aufgerufen werden. Der Prozess wird über **init()**

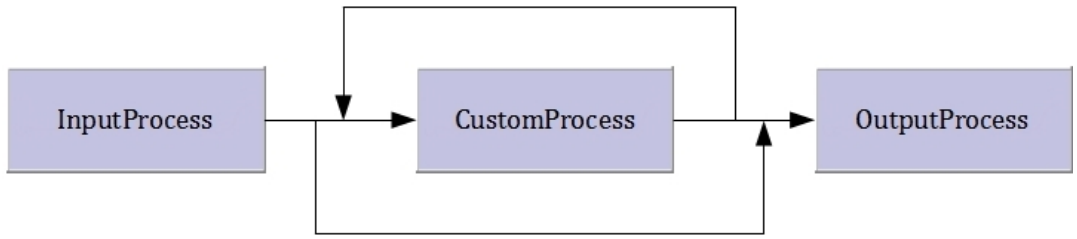


Abbildung 2: Generalisiertes Schema zum Aufbau von Prozessketten mittels *JMU* in *AMOPA*.

initialisiert, indem dessen Parametrierung aus der zugehörigen XML-Datei ausgelesen und verarbeitet wird. Nach der Initialisierung aller Prozesse, ermöglicht **prepare()** Objekte im Prozesskontext gemeinsam zu nutzen (engl. *object sharing*). Die eigentliche Verarbeitungsroutine **execute()** entnimmt das Produkt des Produzenten aus dem Puffer und wird solange wiederholt ausgeführt bis der vorhergehende Prozess terminiert. Zuletzt zeichnet sich **terminate()** für notwendige Aufräumarbeiten beim Beenden des Prozesses verantwortlich.

Als Top-Level Framework aggregiert und vereint das Java-basierte *AMOPA* die Funktionalität der vorgestellten Systeme. Die zuvor aufgeführten Arbeitspakete spiegeln sich in den gleichnamigen Pakten *SPR*, *VID* und *MDH* wider.

Das Paket *IP* ist eines der beiden zentralen Pakete der Bildverarbeitung (engl. *image processing*). Es enthält eine Reihe von Funktionen zur Ein- / Ausgabe, Datenbankanbindung, Datenvorverarbeitung, Datenkonversion, Visualisierung sowie verschiedenste Operatoren der Bildverarbeitung. Es nutzt nicht nur die graphischen Java-Erweiterungen *AWT* und *SWT*, sondern auch die Bildprozessoren des freien Bildverarbeitungs-Werkzeugs *ImageJ*[4], die es um algebraische Funktionen ergänzt.

Konzentriert sich die Funktionalität von *IP* hauptsächlich auf Einzelbilder, widmen sich die Implementationen der Prozessketten der *ProcessEngine* der Verarbeitung von ganzen Bildgruppen, wobei sie prinzipiell auf die *IP*-Funktionen zurück greifen.

Dieser Aufbau gestattet die strukturierte Weiterentwicklung des Frameworks nach den Richtlinien moderner Entwurfsmuster, wie sie beispielsweise durch Erich Gamma in [Gam99,Gam04] vorgeschlagen werden.

---

[4] http://rsbweb.nih.gov/ij

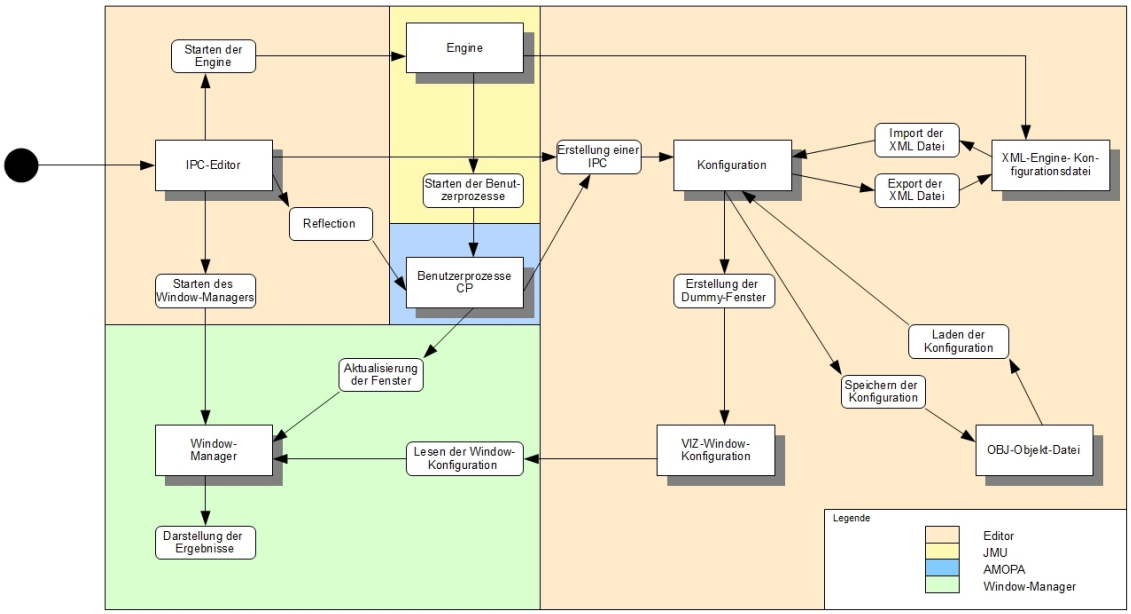


Abbildung 3: Aktivitätsdiagramm des *IPC-Editors*. (Quelle: [Ber09])

2.2 Erweiterungen

Bisher gestaltete sich die Erstellung und Konfiguration der Prozessketten per XML-Datei sehr zeitintensiv und umständlich. Das *IPC-Visualisierungstoolkit* von Christian Berger in [Ber09] besteht aus zwei Komponenten. Abbildung 3 zeigt das Programmablaufdiagramm des Editors. Dessen graphische Oberfläche orientiert sich an dem komfortablen Design von *GraphEditPlus*[5], reduziert deutlich die Entwicklungszeit und erlaubt die Parametrierung der Kettenmitglieder über die *Java Reflection API*[6].

Die Komponente des *Window-Managers* erweitert die Funktionalität von *AMOPA* um einen Visualisierungs-Layer, den alle Prozesse ansteuern können. Er verwaltet die Positionen aller *Windows* und gestattet die direkte Ausführung der Prozesskette über ein Kontextmenü. Zudem besteht die Möglichkeit die Positionen der Fenster der einzelnen Prozesse bereits im Editor festzulegen.

Die Annotation von Videos ist oftmals ebenso langwierig wie schwierig. Zwar wurden in der Vergangenheit verschiedene Tools zur Videoannotation geschaffen, jedoch sind nur

[5] http://www.thedeemon.com/GraphEditPlus

[6] http://java.sun.com/docs/books/tutorial/reflect/index.html

wenige davon frei verfügbar. Nahezu keines ist problemlos anpass- / erweiterbar oder besitzt gleichsam eine gute Usability. Der *Image Annotation Client* (Abk. *IAC*) baut den Framework um eine solche Komponente aus und ermöglicht die direkte Annotation von Zwischenresultaten einer Bildverarbeitungskette mit verschiedenen Zeichenelementen. Diese werden dann in einer separaten XML-Datei in einem erweiterten MPEG7-Format gespeichert. Desgleichen ist die Zuordnung der Elemente zu Objekten innerhalb ontologischer Strukturen möglich. Nähere Details sind ausführlich in [Mül09] beschrieben.

3 Der Anwendungsfall Shot Detection

Der internationale wissenschaftliche Wettbewerb *TRECVID (Text Retrieval Conference series on Video Retrieval Evaluation)* wird vom amerikanischen *National Institute of Standards and Technology* (Abk. *NIST*) seit 2001 jährlich veranstaltet. Mit einer großen Testkollektion von Videodaten bietet *TRECVID* Wissenschaftlern eine Plattform zur Weiterentwicklung bestehender *state-of-the-art*-Algorithmen und zur Schaffung kreativer Innovationen, was mit dem langjährigen Task zur automatischen Shot Boundary Detection (2001-2007) besonders gelungen ist. In diesem konnte die Firma AT&T Research Labs in den Jahren 2006 und 2007 mit besonders guten Resultaten überzeugen. [SOK06]

Nachstehende Absätze geben zuerst einen kurzen Überblick über allgemeine Grundlagen auf dem Gebiet der Shot Detection. Danach werden die Vorteile des Ansatzes von AT&T erläutert und die Einbindung des Referenz-Modells in modifizierter Form in das Prozesskonzept von *AMOPA* dargestellt.

3.1 Allgemeine Grundlagen

Videos, die nicht statischen Kameraaufnahmen entstammen, bestehen im Allgemeinen aus verschiedenen Szenen. Dabei setzt sich eine Szene aus einer Folge von Einzelbildern des gleichen Ortes mit räumlicher und zeitlicher Kontinuität zusammen. So besteht beispielsweise ein Dialog aus zwei oder mehr wiederkehrenden Kameraeinstellungen, wobei jede einzelne als *Shot* bezeichnet wird und somit eine Untermenge der Szene bildet.

Grundsätzlich sind nach [SGG+99] vier Arten von Szenenübergängen unterscheidbar:

1. **Cut:** Die häufigste Form des Shotwechsels bilden harte Schnitte ohne jeglichen Übergang zweier aufeinanderfolgender Frames.

2. **Fade:** Das langsame Einblenden der Szene aus einem monochromen Bild heraus wird als Fade-In definiert. Analog folgt die Definition des Fade-Out.
3. **Dissolve:** Weiche Übergänge respektive Blenden überblenden von einer Szene in eine andere.
4. **Wipe:** Oftmals als Wischblende bezeichnet, deklariert dies einen Szenenübergang mit einer sich durch das Bild bewegenden Linie, der die zu verbindenden Shots jenseits der Linie anzeigt. Dieser Typ ist mathematisch am anspruchsvollsten zu definieren, da der Wechsel in nahezu jedweder Form in beliebigem Winkel erfolgen kann.

Klassische Methoden zur Erkennung von Shotgrenzen berechnen Differenzen verschiedenster Merkmale über eine bestimmte Anzahl von Bildern einer Bildfolge und lösen bei Überschreitung eines zumeist konstanten Schwellwerts eine Detektion aus. Eine Klasse von Verfahren analysiert dabei die Merkmale Farbe und Helligkeit einzelner Bildpixel oder ganzer Pixelregionen in teils unterschiedlichen Farbräumen mittels Histogrammen.

Weit verbreitet ist auch der Ansatz die Verteilung der Kantenstrukturen in einem Bild zu nutzen. Rapide Kantenänderungen sind häufiger beim Szenenwechsel als innerhalb der Szene anzutreffen. Die Berechnung des sogenannten Kanten-Änderungs-Verhältnisses (engl. *edge change ratio* oder ECR) zweier aufeinanderfolgender Bilder wird in [ZMM95] erklärt; distinktive Charakteristiken des ECR zur Identifikation der unterschiedlichen Szenenübergänge beschreiben [dBvDdC$^+$08,Lie99].

Zu den Herausforderungen der korrekten Erkennung von Übergängen zählen Explosionen, eine Szene durchlaufende Menschen, akute Beleuchtungsänderungen, schnelle Bewegungen—gleichermaßen bei Kameraschwenks wie bei Objekten— und nicht zuletzt sehr langsame Übergänge. [MPC06]

Für eine ausführliche Beschreibung der Problemfälle, aufgezeigter und weiterer Methoden sei auf die Grundlagenarbeit zur Shot Detection von Thomas Linowsky [Lin08] verwiesen.

3.2 Das Verfahren von Liu und Gibbon

Das von AT&T in [LZG$^+$06] vorgeschlagene System zur automatischen Detektion von Szenenwechseln skizziert Abbildung 4(links). Zuerst wird das zu verarbeitende Video sequentiell Bild für Bild dekodiert. Ein Ringpuffer speichert die letzten 256 Bilder eines Videos, ein anderer die jeweils zugehörigen extrahierten visuellen Merkmale. Beide versorgen sechs verbundene Shotdetektoren nach deren Bedarf mit benötigen Informationen.

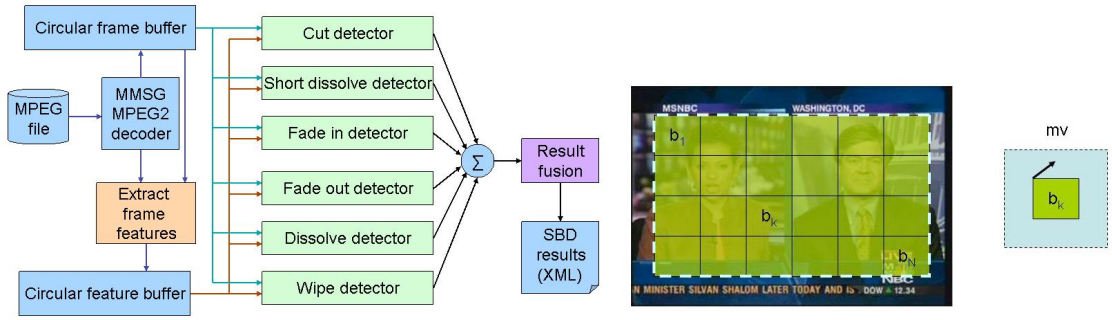


Abbildung 4: Überblick über das System zur Shotgrenzenerkennung von AT&T (links). Unterteilung des Bildes zur Extraktion visueller Merkmale (rechts). (Quelle: [LZG+06])

Eine Routine zur Ergebnisfusion sorgt mit einem einfachen regelbasierten Algorithmus dafür, dass gefundene Shots nicht überlappen.

Das mögliche Repertoire an Shots wird mit der Erkennung von harten Schnitten und weicheren Übergängen wie (schnellen) Über-, Ein- und Ausblenden sowie einer bestimmten Sorte von Wipes auf den verbundenen *TRECVID* Datensätzen weitestgehend abgedeckt (vgl. 3.1).

Da zu den häufigsten Übergängen in Filmen die harten Schnitte und die graduellen weichen Blenden gehören, beschränkt sich die Darstellung der nächsten Abschnitte insbesondere auf die Extraktion der notwendigen Merkmale für den Cut- und auszugsweise den $Dissolve$-Detektor. Für eine vollständige Betrachtung sei auf die Veröffentlichungen von Liu und Gibbon in [LZG+06,LZG+07] verwiesen.

Visuelle Merkmalsextraktion

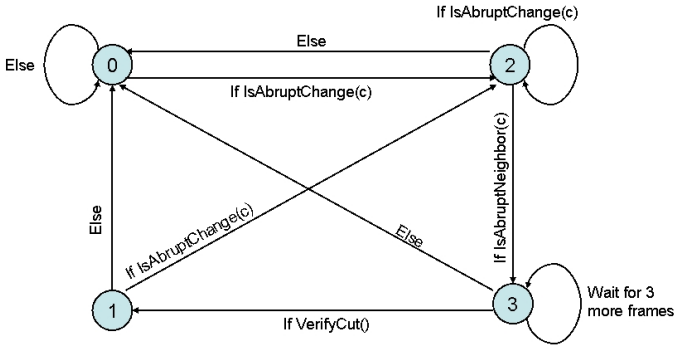
Aus jedem Bild wird eine Menge an Features extrahiert. Zu den Merkmalen eines Einzelbildes gehören bekannte Größen wie das horizontale und vertikale ECR und im Besonderen die Histogramme der einzelnen Kanäle des RGB-Farbraumes und des Grauwertkanals sowie deren statistische Größen Mittelwert, Varianz, Schiefe, Exzess und dynamische Breite. Ein großer Teil der Merkmale zwischen mehreren Bildern (engl. *interframe features*) errechnet sich über die Differenz der Einzelbildfeatures, gleichermaßen im Abstand von zwei oder sechs aufeinanderfolgenden Bildern.

Weitere inter-frame Merkmale können über *Blockmatching*[7] gewonnen werden. Dies bedingt die Aufteilung des Bildes in disjunkte Teilbereiche, was der linke Teil in Abbildung 4(rechts) illustriert. Zu jedem Block wird innerhalb einer vorgegebenen Suchweite ein Verschiebungsvektor (engl. *motion vector*) zum nächsten Bild berechnet (siehe Abbildung 4(rechts) kleines Bild). Um dieses Matchingproblem zu lösen, wird eine Maske in Größe des originären Blocks systematisch über den Suchraum bewegt, mit dem Ziel den punktweisen Vergleich der Bildelemente zwischen der aktuellen Position und dem Muster anhand einer vorgegebenen Abstandsfunktion zu minimieren. Dieser Abstand (engl. *matching error*) beschreibt die Ähnlichkeit zweier Blöcke, wohingegen das Verhältnis des Fehlers zwischen dem bestem Match und dem Durchschnittswert innerhalb der Suchweite als *matching ratio* bezeichnet wird. Für die drei lokalen Größen werden u.a. der Durchschnittswert und Median global über alle Blöcke ermittelt.

Bewährt haben sich quadratische Blöcke der Größe 48×48 mit einer Suchweite von 32×32, da sie zuverlässigere Bewegungsvektoren hervor bringen als das bei kleineren Blockgrößen wie z.B. 8×8 der Fall ist.

---

[7] Eine systematische Einführung in die zugehörigen Themenbereich Objektverfolgung enthält [Bis06].



State Variables
State_ID
Start_frame
End_frame
AverageME

State_ID	Description
0	Initial state
1	Cut detected state
2	Possible cut detected state
3	Cut verification state

Abbildung 5: Automatengraph des Cut-Detektors, wobei c die Nummer des aktuellen Frames angibt. (Quelle: [LZG+06])

Detektoren zur Shotgrenzenerkennung

Die Detektoren sind als endliche Automaten (engl. *finite state machines* oder *FSM*) implementiert. Sie starten im Zustand 0. Zustand 1 markiert einen detektierten und vom Automaten verifizierten Shot. Alle weiteren Zustände markieren Übergangszustände, deren Anzahl und Übergangsfunktionen je nach Detektor variieren kann. Darüber hinaus besitzen die Detektoren Zustandsvariablen, wobei $State\_ID$ den aktuellen Zustand, $Start\_Frame$ den letzten Frame der vorherigen Kameraeinstellung und End_Frame das erste Bild des neuen Shots repräsentieren.

Der Cut-Detektor besitzt vier verschiedene Zustände (siehe Abbildung 5). Die wichtigste Variable $AverageME$ wird mit dem Default-Wert 5.0 initialisiert und immer im Zustand 0 aktualisiert. Sie beschreibt den durchschnittlichen Fehler einer Bildfolge unter Zuhilfenahme des durchschnittlichen Fehlers eines Bildes ME_A in einem Filter mit unendlicher Impulsantwort (engl. *infinite impulse response* oder IIR):

$$AverageME \;=\; AverageME * 0.85 + ME_A * 0.15 \;.$$

Ist der aktuelle ME_A um ein Vielfaches größer als $AverageME$ und der ME_A der letzten fünf Frames, veranlasst die Funktion $IsAbruptChange$ einen Wechsel des Automaten in den Zustand 2 einer möglichen Detektion , wobei End_Frame den Wert $c-1$ und $Start\_Frame$ den Wert c erhält. Diese Form eines adaptiven Schwellwerts erlaubt die verlässliche Detektion auftretender *Hardcuts*. Falls der aktuelle ME_A größer als der vorherige ist, wechselt $IsAbruptNeighbor$ in den Verifikationszustand 3, anderseits kehrt die *FSM* zu Zustand 0 zurück.

Zustand 3 verweilt für drei nachfolgende Bilder und vergleicht danach die Ähnlichkeit der Frames vor und nach dem vermeintlichen Shot. Bei positivem Resultat meldet er eine erfolgreiche Detektion über den Zustand 1, ansonsten springt er in den Ursprungszustand zurück.

Eine weiche Überblendung ist oftmals gekennzeichnet durch einen bitonischen Verlauf der Varianz im Helligkeitskanal HV_l, die typischerweise zuerst monoton abfällt und anschließend wieder steigt (siehe Abbildung 6). Dies beruht auf der Idee, dass sich alle Zwischenbilder Z_i eines weichen Szenenwechsels zweier verschiedener Kameraeinstellungen X und Y über deren konvexe Linearkombination berechnen lassen.

$$Z \;=\; \alpha_i X + (1-\alpha_i)Y \quad mit \quad \alpha_i \in [0,1] \;.$$

Der $Dissolve$-Detektor (Abbildung 7) besteht aus fünf Zuständen, um einen derartigen Kurvenverlauf zu modellieren. Die zentrale Variable $AverageVariance$ mittelt die Va-

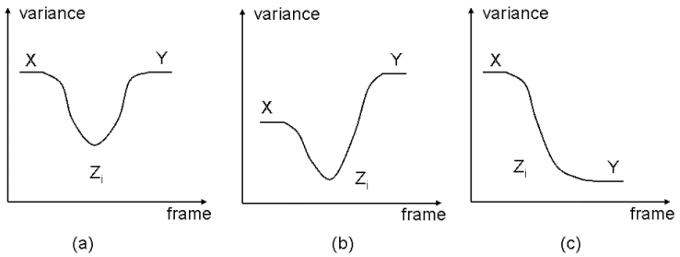


Abbildung 6: Charakteristischer Kurvenverlauf der Varianz bei weichen Szenenübergängen, der aus der Linearkombination verschiedener Szenenbilder X und Y resultiert. Links: Symmetrischer quadratischer Funktionsverlauf bei Unabhängigkeit. Mitte: Typisch asymmetrischer Verlauf bei einer gewissen Abhängigkeit. Rechts: Äußerst geringe Varianz in Y. (Quelle: [LZG+06])

rianz über eine Bildfolge, wird mit dem Wert 3.5 initialisiert und definiert die IIR im Zustand 0 über:

$$AverageVariance \;=\; AverageVariance * 0.85 + HV_l * 0.15 \; .$$

Zustand 2 modelliert den Teil der abfallenden Varianz, Nummer 3 den der steigenden. Das Kernstück dieser *FSM* bildet eine ausgefeilte Verifizierungsfunktion, die zum einen die Shotgrenzen bestimmt, zum anderen die unterschiedlichen Kurvenverläufe adäquat zu handhaben in der Lage ist, indem sie die Fähigkeiten einer *Support Vector Machine* mit insgesamt 66 verschiedenen Merkmalen nutzt (weitere Details siehe [LZG+06](S. 5ff)).

3.3 Migration nach AMOPA

Das Verfahren von Liu und Gibbon hat sich auf den *TRECVID*-Datensätzen von 2005 bis 2007 bewährt[8]. Jedoch unterscheiden diese sich in den Punkten Auflösung, Kodierung und Inhalt unterschiedlich stark von dem Videomaterial lokaler Fernsehsender, was im Regelfall eine Neuparametrierung oder Modifikation der vorgestellten Algorithmen mit sich bringt.

Ziel der Integration dieser Methoden in das Prozesskonzept von *AMOPA* (siehe Abbildung 8) war es, die vorgestellte Technik nicht nur zu verifizieren, sondern auch deren Parametrierung über die Visualisierung von einzelnen Zwischenschritten zu erleichtern.

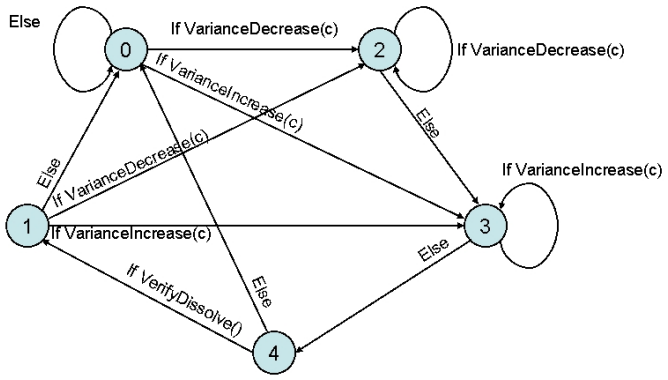
---

[8] http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6.sb.slides-final.pdf

Während $FrameReader$ die Bilder sequentiell ausliest, beginnt $HistogramExtraction$ mit der Berechnung aller Histogramme und weiterer intra-frame Merkmale. Zur besseren Systemauslastung bei der Extraktion der Bewegungsvektoren $MotionExtraction$ kommen auf einem Doppel-Quad-Core-System (Xeon 5430 mit 3.0 GHz) mehrfache Threads zur Anwendung, wobei jeder einen Block bearbeitet. Dabei konnte sich die N-Schritt-Suche gegenüber der vollständigen Suche mit einem Zehntel der Laufzeit durchsetzen, ohne dass die geringe Abnahme der damit verbundenen Genauigkeit $Recall$ und $Precision$ der gefundenen Shots mindert. Unter der Beachtung programmiertechnischer Constraints wie dem Recycling von Objekten statt deren Neukreation, können aktuell etwa 50 Bilder pro Sekunde verarbeitet werden.

Die Berechnung der inter-frame Merkmale erfolgt in $CalculateStatistics$. $ShotDetection$ stellt einen Verwaltungsprozess dar, der alle verbundenen Shot-Detektoren sukzessive aufruft. Die beiden Shot-Detektoren verwalten entgegen der vorherigen Beschreibung nun kleinere lokale Puffer, bevor $ResultFusion$ abermals Überlappungen verhindert.

Besonderer Wert wurde bei dieser Implementierung darauf gelegt, dass nahezu jede Komponente Möglichkeiten zur Visualisierung ihrer Variablen, Zustandsgrößen oder gar einzelner Zwischenschritte besitzt. Beispielsweise nutzt Abbildung 9 Methoden der frei-



State Variables
State_ID
Start_frame
End_frame
AverageVariance

State_ID	Description
0	Initial state
1	Dissolve detected state
2	Variance decreasing state
3	Variance increasing state
4	Fade in verification state

Abbildung 7: Automatengraph des Dissolve-Detektors. (Quelle: [LZG+06])

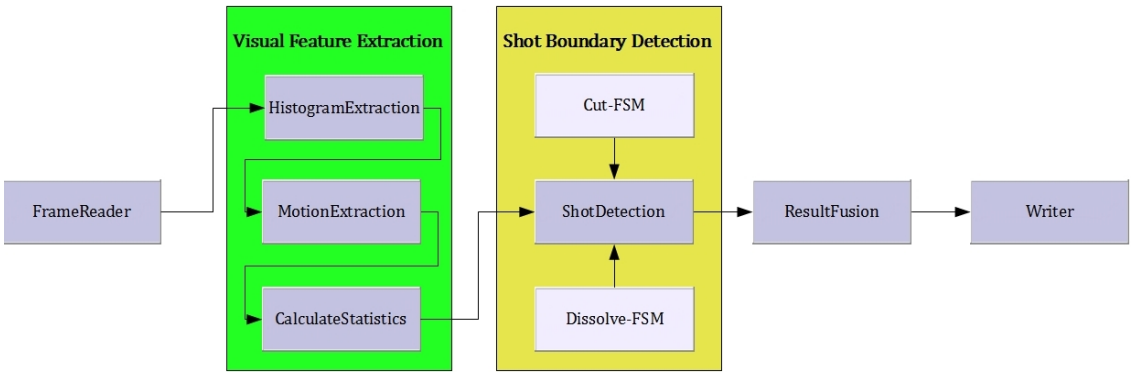


Abbildung 8: Die Prozesskette des AT&T-Verfahrens im Framework *AMOPA*. Die Subsysteme Merkmalsextraktion und Shot Boundary Detection aus 3.2 sind grün- bzw. gelblich markiert und werden auf mehrere Prozesse abgebildet. Alle Prozesse sind dunkelgrau gekennzeichnet; $ShotDetection$ aggregiert die beiden hell unterlegten Detektoren.

en *OpenSourcePhysics-Engine* [9], um den Funktionsverlauf der einzelnen Parameter des Cut-Detektors anzuzeigen. Das ebenfalls Java-basierte OSP bietet ein multifunktionales Arsenal zur weiteren Datenanalyse und -regression.

Die roten Spitzen werden durch die rapiden Änderungen des aktuellen $mean\ average\ errors$ hervorgerufen, was meist mit einer Zustandsänderung in der *FSM* und einem detektierten Hardcut einher geht (blaue Spitze). Die Durststrecke der ersten 1.000 Bilder ist einer Vielzahl weicher Überblendungen geschuldet, die durch den Cut-Detektor nicht bearbeitet werden. Animationsbedingte rote und blaue Spitzen treten im Bereich der Bilder 2.950 bis 3.100 auf. Eventuelle $FalseAlarms$ verhindert die Verifikationsfunktion im Zustand 3.

4 Zusammenfassung und Ausblick

Dieser Beitrag hat den an der TU Chemnitz entwickelten Framework *AMOPA* näher betrachtet. Aktuelle Arbeiten beschäftigen sich mit der Erweiterung des Prozesskonzeptes. Bisher konnten nur lineare Prozessketten erzeugt und verarbeitet werden. In Zukunft

[9] http://www.compadre.org/OSP/

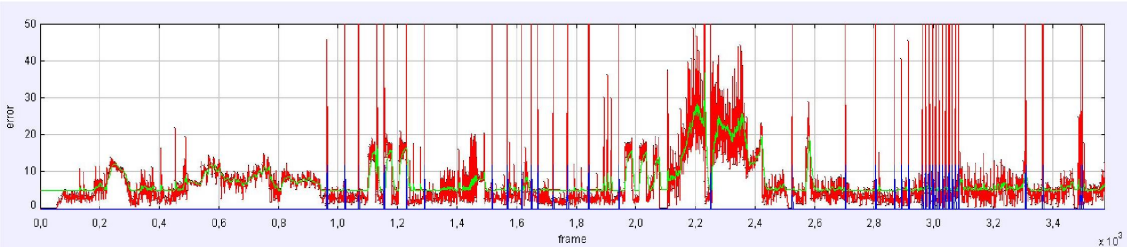


Abbildung 9: Visualisierung des Funktionsverlaufs der Variablen des Cut-Detektors $AverageME$ (grün), ME_A (rot) und eines Vielfachen der $State\_ID$ (blau) auf dem zufällig ausgewählten Video BG_26797 aus dem *TRECVID* Datensatz von 2008 mit einer Gesamtzahl von 3.605 Frames.

werden, neben der $1:1$ Relation, beliebige Relationen ($1:N$, $N:1$, $N:M$) zwischen miteinander verschalteten Produzenten und Konsumenten möglich sein. Diese Vorgehensweise ist zwar bereits im Konzept von Mosna angedacht, jedoch nicht umgesetzt worden. Die vorgestellten Erweiterungen *IPC* und *IAC* befinden sich aktuell in der Endphase der Entwicklung.

Skizziert Abschnitt 2 lediglich eine verallgemeinerte Form einer Prozesskette, führt Abschnitt 3 in die Grundlagen der Szenenwechselerkennung ein und demonstriert die Umsetzung eines Algorithmus mittels einer *IPC* in *AMOPA* am Beispiel eines Verfahrens von AT&T.

Eine akkurate Evaluation des Verfahrens von Liu und Gibbon bleibt zukünftiger Forschungsarbeit vorbehalten, wobei lediglich das Material der Datensätze der TRECVID-Wettbewerbe von 2005 bis 2007 geeignet erscheint, da die verfügbaren GroundTruth-Daten verschiedene Shottypen beherbergen. Aktuelle TRECVID-Datensätze enthalten zwar automatisch berechnete Master- und Shotgrenzen, geben aber nicht die Art des Szenenwechsels an. Ebenfalls sind diese maschinell gewonnenen Annotationen aufgrund einer Restfehlerquote nicht vollkommen fehlerfrei.

Der Standortbestimmung am TRECVID-Datensatz wird sich eine geordnete und systematische Evaluation auf ausgewählten Datensätzen der lokalen Fernsehsender anschließen, mit Hilfe derer die verbleibenden Parameter der endlichen Automaten für unterschiedliche Kategorien an Videodaten adaptiv und vollautomatisch bestimmt werden sollen.

Literatur

[Ber09] Christian Berger. Entwurf und Implementierung eines Visualisierungstoolkits für den Workflow von Ketten von Bild- und Videoverarbeitungsprozessen. Diplomarbeit, Technische Universität Chemnitz, 2009.

[Bis06] Alexander Bischof. Systematisierung und Evaluierung von Verfahren der Objektverfolgung in Videosequenzen. Diplomarbeit, Technische Universität Chemnitz, 2006.

[dBvDdC+08] Sarah de Bruyne, Davy van Deursen, Jan de Cock, Wesley de Neve, Peter Lambert, and Rik van de Walle. A compressed-domain approach for shot boundary detection on h.264/avc bit streams. *Image Commun.*, 23(7):473–489, 2008.

[Gam99] Erich Gamma. Design Patterns at Work. *International Conference on Technology of Object-Oriented Languages*, 0:4, 1999.

[Gam04] Erich Gamma. *Entwurfsmuster: Elemente wiederverwendbarer objektorientierter Software*. Addison-Wesley, München, 1 edition, 2004.

[Lie99] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases*, number SPIE 3656, pages 290–301, January 1999.

[Lin08] Thomas Linowsky. Grundlagen der Shot Detection. Studienarbeit, Technische Universität Chemnitz, 2008.

[LZG+06] Zhu Liu, Eric Zavesky, David Gibbon, Behzad Shahraray, and Patrick Haffner. AT&T RESEARCH AT TRECVID 2007. Workshop Beitrag, AT&T Labs-Research, 200 Laurel Avenue South, Middletown, NJ 07748, 2006. http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/att.pdf, 13.05.2009.

[LZG+07] Zhu Liu, Eric Zavesky, David Gibbon, Behzad Shahraray, and Patrick Haffner. AT&T RESEARCH AT TRECVID 2007. Workshop Beitrag, AT&T Labs-Research, 200 Laurel Avenue South, Middletown, NJ 07748, 2007. http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/att.pdf, 13.05.2009.

[Mül09] Sven Müller. Entwurf und Implementierung eines erweiterbaren Annotationstoolkits zur Erzeugung von Referenz-Annotationen für Videos.

Diplomarbeit, Technische Universität Chemnitz, 2009.

[Mos07] Paolo Mosna. JMU: Java Media Utility, 2007. http://sourceforge.net/projects/jmu, 13.05.2009.

[MPC06] N. Manickam, A. Parnami, and S. Chandran. Reducing false positives in video shot detection using learning techniques. pages 421–432, 2006.

[sac09] InnoProfile Projekt sachsMedia — Cooperative Producing, Storage, Retrieval and Distribution of Audiovisual Media, 2009. http://www.tu-chemnitz.de/informatik/Medieninformatik/sachsmedia, http://www.unternehmen-region.de/de/1849.php, 14.05.2009.

[SGG+99] Alan F. Smeaton, J. Gilvarry, G. Gormley, B. Tobin, S. Marlow, and N. Murphy. An evaluation of alternative techniques for automatic detection of shot boundaries. In *School of Electronic Engineering*, pages 8–9, 1999.

[SOK06] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. http://www-nlpir.nist.gov/projects/trecvid/, 14.05.2009.

[str09] Streambaby - Tivo HME Streaming Application, 2009. http://code.google.com/p/streambaby, 13.05.2009.

[ZMM95] Ramin Zabih, Justin Milller, and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings ACM Multimedia 95*, pages 189–200, 1995.

Textdetektion und -extraktion mit gewichteter DCT und mehrwertiger Bildzerlegung

Stephan Heinich

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

stephan.heinich@informatik.tu-chemnitz.de

Zusammenfassung: In diesem Artikel beschreibe ich die Arbeit an einer Textdetektion in Videos als Vorbereitung für die folgende Texterkennung. Ich benutze das Archiv der lokalen Fernsehsender als Grundlage. Der Inhalt kann meist als Nachrichtensendung kategorisiert werden. Als erstes werden mit einer einfachen aber schnellen Heuristik Textkandidaten selektiert. Die folgenden Schritte verarbeiten und bereiten den Kandidaten vor, damit er mit einer Standard-OCR-API (optical character recognition) verarbeitet werden kann. Der erste Teil der Textverarbeitung ist die Detektion. Dies geschieht zum einen mit einer gewichteten DCT und zum anderen mit einem Verfahren zur mehrwertigen Bildzerlegung. Anschließend werden die Bilder mit verschiedenen Schwellwertverfahren für eine Texterkennung vorbereitet. Im letzten Schritt verarbeitet eine freie OCR-API die aufbereiteten Frames.

Schlagwörter: Diskrete Cosinus-Transformation, gewichtete DCT Koeffizienten, Textdetektion, Texterkennung

1 Einleitung

In dem Projekt sachsMedia versuchen wir die Archive von lokalen Fernsehsendern zu analysieren und sie für die Firmen und das Fernsehpublikum zugänglich zu machen. Um ein multimediales Retrieval System zu entwickeln, verwenden wir alle Modalitäten die die Videodaten bereitstellen: Text-, Sprache- und Bildinformationen. Die Größe der Datenkollektion beträgt ungefähr einen Terrabyte. Der Inhalt des Videomaterials kann zum größten Teil als Nachrichtensendung kategorisiert werden. Sie beinhaltet trotzdessen eine große Menge an Werbematerial. In diesem Paper beschreibe ich verschiedene Methoden zur Erkennung und Extraktion von textueller Information. Der Aufbau des Papers ist im Folgenden erläutert. Im zweiten Kapitel wird die Frameverarbeitung in den einzelnen Prozessschritten erklärt. Die Frameverarbeitung teilt sich in zwei Bereiche. Im ersten Teil wir die gewichtete DCT behandelt. Im 2 Teil wird das Verfahren auf Basis der Bildsegmentierung erläutert. In den beiden letzten Kapiteln wird ein Fazit und ein Ausblick in zukünftige Arbeit gegeben.

Seit dem die resultierende Textinformation, speziell in Nachrichtensendungen, in multimedialen Retrieval Systemen nützlich ist, haben sich viele Wissenschaftler mit der automatischen Textextraktion in Videos beschäftigt. Ein genereller Überblick über Textinformationsextraktion (TIE) ist in [Ju04] gegeben. Sie elaborieren eine generelle Architektur von TIE-Systemen, welche meist aus fünf Schritten bestehen: Detektion, Lokalisierung, Extraktion und Weiterverarbeitung sowie die Erkennung. In meinen Paper stelle ich zwei Ansätze zur Lokalisierung und deren detailierte Extraktion vor. Ich habe eine detaillierte Erklärung der Erkennungsphase vermieden, da ich verschiedene state-of-the-art OCR-Systeme verwende.

In dem letzten Jahrzehnt haben einige Wissenschaftler die DCT-Koeffizienten von MPEG-komprimierten Videos zur Textdetektion und Lokalisierung verwendet. Es gibt zwei herausragende Gründe DCT-Koeffizienten zu nutzen. Der erste Grund ist, dass die DCT-Koeffizienten invariant zur Größe des Textes und dessen Schriftart sind. Die Berechnung verschiedener Energien (durch die Summierung der Amplituden der DCT-Koeffizienten) lässt sich durch parametergesteuerte Algorithmus an bestimmte Anwendungsfälle, wie zum Beispiel Overlay-Text und spezielle Arten von Szenentext, anpassen. Der zweite Grund ist kürzere Verarbeitungsdauer auf komprimierten Videomaterial. Zhong et. al. [Zho00] berechnete die horizontalen und vertikalen Energien in einem DCT-Block eines I-Frames, in komprimierter Form, um Text zu detektieren. Bei einem ähnlichen Ansatz benutzt Qian et. al. [XQi07] eine Anzahl an (horizontalen, vertikalen und diagonalen) DCT-Koeffizienten um chinesischen und englischen Text zu detektieren. Ein schweres Problem ist die Bestimmung des Schwellwertes für die Erkennung von Textblöcken. In Shiratori et. al. [Shi06] und Lu et. al. [Lu08] wird der Fishersche Diskriminatenmittelpunkt verwendet. Zusätzlich testete Lu et. al. [Lu08] verschiedene gewichtete DCT-Koeffizienten. In meinem Versuch verwende ich zwei Algorithmen um dem Problem der Schwellwertbestimmung entgegenzutreten.

Als Referenzverfahren wurde eine Methode ausgewählt die auf eine mehrwertige Bildzerlegung in Kombination mit einer Connected-Component-Analyse besteht. Dieses Verfahren wurde von Jain et. a. [AKJ98] entwickelt. Mit Hilfe dieses Verfahrens wurde getestet wie gut das Verfahren der DCT gegenüber Alternativverfahren abschneidet.

2 Frame-Verarbeitung

Die folgenden Schritte werden zur Videovorverarbeitung genutzt. Als erstes werden Textkandidaten mit einem einfachen, aber schnellen heuristischen Algorithmus selektiert. Die DCT liefert Informationen über das Frequenzspektrum der betrachteten Bildregionen. Bestimmte Frequenzbereiche werden verstärkt und andere gedämpft. Danach wird eine Normalisierung und eine Binarisierung angewendet um eine Maske

zu erstellen. Diese wird auf den aktuellen Frame angewendet, der anschließend wird der Frame mit Hilfe einer regionbasierenden Technik binarisiert. Im letzten Schritt wird das Binärbild einer OCR-API übergeben um den beinhalteten Text zu extrahieren.

2.1 Frame Selection

Ich benötigte eine schnelle Methode zur Vorselektierung von Textkandidaten und entschied mich das heuristische Verfahren von Lim et. al. [Lim00] zu verwenden. Der Algorithmus summiert Kantenpixel im Frame. Für jeden Pixel muss der Unterschied der Helligkeit zu seinen Nachbarn eine Schwelle überschreiten um zur Summe addiert zu werden. Das ist eine gute Abschätzung zur Detektion von Text, trotz dessen ruft diese Methode einige Falschalarme hervor. Beispielsweise werden Frames ausgewählt die eine feingranulare Textur besitzen, aber keinen Text enthalten. Desweiteren werden manche Frames nicht erkannt in dem sich Text befindet. Wenn die Anzahl der akzeptierten Kantenpixel einen Schwellwert überschreitet nehme ich an, dass dieser Frame Text enthält. Die Kantenpixel werden wie folgt berechnet:

$$S = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} K(x,y)$$

X und *Y* begrenzen die Dimension des Bildes. $K(x,y)$ ist folglich definiert als:

$$K(x,y) = \begin{cases} 0, & |I(x,y) - I(x+1,y)| < 170 \\ 1, & |I(x,y) - I(x+1,y)| > 170 \\ 1, & |I(x,y) - I(x,y+1)| > 170 \end{cases}$$

$I(x,y)$ ist der aktuelle Pixel im Bild. Ob der Frame ein Textkandidat ist oder nicht wird durch folgende Formel bestimmt:

$$isText = \begin{cases} false, & S_f < 500 \\ true, & S_f > 500 \end{cases}$$

S_f gibt die Summe der Kantenpixel im Frame f an. Der Textkandidat wird, wie im folgenden Abschnitt beschrieben, weiter verarbeitet. Zwei Beispiele für Textkandidaten sind in der Abbildung 1 beschrieben.

Abbildung 1: (a), (c) originale Frames und (b), (d) repräsentieren die Pixelmaps der beinhalteten Kantenpixel (von links nach rechts, von oben nach unten)

2.2 DCT

Die DCT wird in der Regel im Bereich der Bildkompression angewendet. Hierbei wird das Bild in gleich große Pixelblöcke (Makro-Blöcke) aufgeteilt. Es wird für jeden Makro-Block das Frequenzspektrum berechnet. Menschen sind nicht in der Lage sehr Hohe Frequenzen in kleinen Bildern zu erkennen. Diese Regionen können aus dem Bild entfernt werden, wobei die Frequenz abgesenkt und die Information reduziert wird um eine Datenreduktion zu erzielen. Umgekehrt schaffen wir eine Methode um signifikante Frequenzen zu verstärken. Normalerweise haben Buchstaben aller Art ein klare Silhouette gegenüber ihrem Hintergrund. Ein hohes Vorkommen von Kanten, wie zum Beispiel ein Zeile Buchstaben, produzieren ein verstärktes Frequenzprofil.

2.2.1 DCT-Koeffizienten

Nach der Idee von Lu et. al. [Lu08] verwende ich eine blockweise DCT, bestehend aus 16 × 16 Pixel großen Blöcken mit einer 8 Pixel breiten Überlappung in allen Richtungen. Das ergibt eine Gesamtüberlappung von 75%. Bilder beinhalten

normalerweise viele niederfrequente Bereiche, wobei Text sich vorwiegend im mittleren Frequenzbereich befindet. Die Summe der DCT-Koeffizienten bestimmt die Energie eines Blockes. Die Summe der gewichteten DCT-Koeffizienten stellt ein Maß über den Grad der Textpräsenz dar. Über ein zum Ergebnis manuell kalibrieten Schwellwert können Textbereiche in verschiedenen Bildern bestimmt werden.

2.2.2 Gewichtete DCT-Koeffizienten

Für eine weitere Verbesserung des Textlokalisierungsprozesses werden Gewichte für die DCT-Koeffizienten benutzt. Es kann nicht nur mit der Summe der Frequenzen maskiert werden. Es muss auch eine Verstärkung mit Hilfe eines geeigneten Gewichtungsschemas angewendet werden. Die Energie E eines Makro-Blockes mit einer Größe von $N \times N$ kann als gewichtete Summe der Koeffizienten $W(p,q)$ beschreiben werden.

$$E = \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} W(p,q)$$

Verschiedene Ansätze für die Gewichtung von Koeffizienten wurden in [Lu08] abgedeckt. Ich bevorzuge eine andere Methode um gewichtete Koeffizienten $W_e(p,q)$ zu berechnen:

$$W_e(p,q) = \begin{cases} 0, & p+q < 4 \\ C(p,q)^2, & 4 \leq p+q < 12 \end{cases}$$

Die gewichteten Koeffizienten $W_e(p,q)$ resultieren aus dem Quadrat des Koeffizienten $C(p,q)$ um irrelevante Frequenzen abzuschwächen. Wenn die Summe der Indizes p und q einen bestimmten Bereich der Frequenzen erreicht wird der korrespondierende Koeffizient null gesetzt.

2.2.3 Normalisierung

Im letzten Schritt, bevor die Textkandidaten erweitert und binarisiert werden, benötige ich eine Normalisierung der Energie, die zuvor berechnet wurde. Dies ist für die Vorbereitung der Schwellwert-basierenden Binarisierung und der Eliminierung von Ausreisern innerhalb der erkannten Textkandidaten-Blöcke wichtig. Ich benutze die Gesamtenergie der Blöcke zur Normalisierung. Der daraus resultierende Normalisierungsfaktor ist im folgenden beschreiben:

$$N = \frac{1}{X \cdot Y} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} E(x,y)$$

x und y sind die Indizes der einzelnen Blöcke im Bereich von X, Y und $E(x,y)$ repräsentiert die seine Energie. Die normalisierten, gewichteten Koeffizienten NW ergibt sich wie folgt:

$$NW(x,y) = \frac{E(x,y)}{N}$$

Die normalisierten Energien ergeben einen Maske des Original Bildes. Die Maske ist an eine Grauwertdarstellung im Bereich von 0 bis 256 angepasst:

$$M(x,y) = \begin{cases} 0, & NW(x,y) < 0 \\ 255, & NW(x,y) > 255 \end{cases}$$

Alle Werte zwischen 0 und 255 bleiben unverändert. Bei dem folgenden Schritt wird zur Binarisierung der Maske ein einfaches globales Schwellwertverfahren verwendet.

2.2.4 Globales Schwellwertverfahren

Ein globaler Schwellwertalgorithmus wird zur Binarisierung benutzt. Hierbei wird das Histogramm H zur Maske M errechnet. Anschließend werden zwei Maxima Max1 und Max2 wie folgt berechnet:

$$Max_1 = \max(H)$$

$$Max_2 = \max(Max_1 \notin H)$$

Im nächsten Schritt wird der Schwellwert wie folgt bestimmt:

$$T = \min\left(H_{[Max_1,\dots,Max_2]}\right)$$

Zusätzlich selektierte ich das Minimum der beiden Maxima. Letztlich kann die binäre Maske M_b mit der Hilfe des Schwellwertes erstellt werden:

$$M_b(x,y) = \begin{cases} 0, & M(x,y) < T \\ 1, & M(x,y) > T \end{cases}$$

2.3 Mehrwertige Bildzerlegung

Der Algorithmus der mehrwertigen Bildzerlegung wurde von Anil et. al. [AKJ98] entwickelt und von mir als Referenzalgorithmus verwendet. Bei diesem Verfahren werden potenzielle Textregionen durch eine Connected-Component-Analyse auf einzelnen Vordergrundbildern bestimmt. Der Ablauf des Algorithmus im folgenden erläutert.

2.3.1 Farbraumverkleinerung

Die Connected-Component-Analyse zielt darauf ab das Bild in verschiedene Fraben zu unterteilen. Aus dem Grund muss die Anzahl der Farben auf ein sinnvolles Maß reduziert werden. Diese Reduzierung wird dem Bit-Dropping-Verfahren erreicht. Dabei wird ein vorgegebene Anzahl höher wertiger Bits jedes Farbkanals eines Pixels erhalten, wobei alle niederwertigen Bits null gesetzt werden.

2.3.2 Mehrwertige Bildzerlegung

Ein Bild ist mehrwertig, wenn es mindestens zwei verschiedene Farben *l* enthält:

$$l \in L = \{0,1,\ldots,L-1\}, \qquad L > 1$$

Abbildung 2: Farbraumverkleinerung: (a) Bit Dropping auf 2 Bit pro Farbkanal, (b) Color Clustering (von links nach rechts)

Ein L-wertiges Bild kann in eine Menge von L Teilbildern $I = I_i$ zerlegt werden, wobei gelten muss:

$$\bigcup_{i=0}^{L-1} I_i = I$$

$$I_i \bigcap_{i \neq j} I_j = \emptyset$$

Anschließend wird jede aus jeder im Bild enthaltenen Farbe ein Vordergrundbild erzeugt, in dem ausschließlich die Pixel der entsprechenden Farbe gesetzt sind. Das Ergebnis der mehrwertigen Bildzerlegung ist eine Liste der verschiedenen Vordergrundbilder.

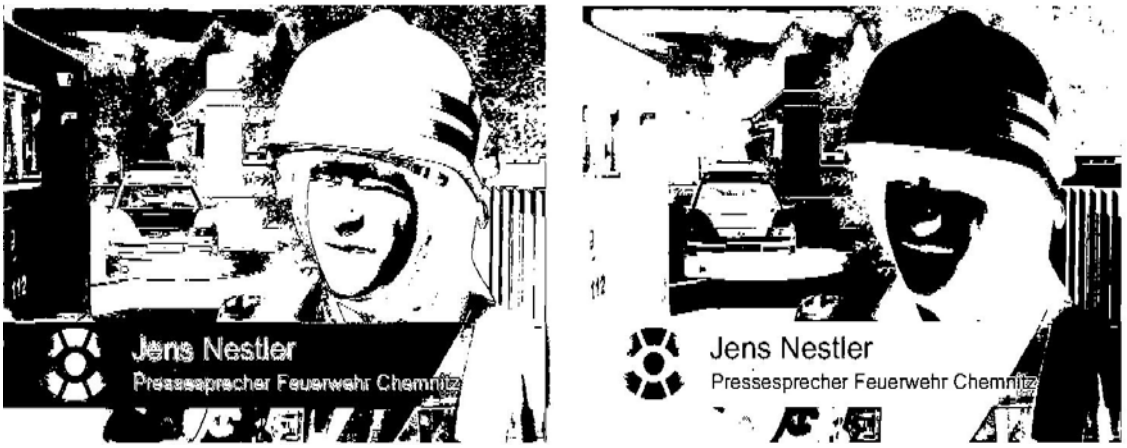


Abbildung 3: Ergebnis der Mehrwertigen Bildzerlegung: 2 Vordergrundbilder

2.3.3 Region-Growing

Durch die Anwendung eines Region-Growing-Algorithmus werden Connected Components der einzelnen Teilbilder extrahiert. Liegen die Werte für Höhe und Breite eine CC nicht in bestimmten Intervallen, so wird diese Komponente als Falschalarm betrachtet und verworfen.

Abbildung 4: (a) Region Growing, (b) resultierende Textobjekte (von links nach rechts)

2.3.4 Textzeilen erzeugen

Benachbarte CCs werden zu Textzeilen zusammengefasst, wenn folgende Bedingungen erfüllt sind:

i. Differenz der oberen CC-Grenzen ist kleiner als ein Schwellwert t_{oben}
ii. Differenz der unteren CC-Grenzen ist kleiner als ein Schwellwert t_{unten}
iii. der horizontale Abstand der CCs ist kleiner als ein Schwellwert t_{hor}

2.3.5 Analyse der Projektionsprofile

Die Entscheidung, ob eine potentielle Textzeile wirklich Text enthalten kann wird anhand einer Analyse der Projektionsprofile getroffen. Das horizontale Projektionsprofil einer Textzeile ist wegen der vielen Buchstaben und Zwischenräume durch mehrere Erhebungen charakterisiert, während im vertikalen Profil nur sehr wenige Hügel existieren. Es müssen daher folgende Bedingungen erfüllt sein [AKJ98] :

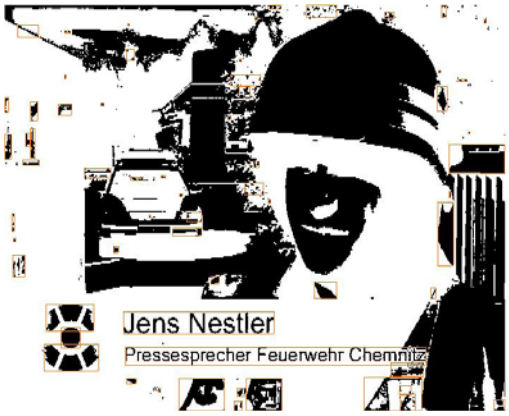


Abbildung 5: Zusammenfügen benachbarter Textobjekte zu Textzeilen

i. in der horizontalen Signatur sind mindestens 5 Hügel enthalten
ii. jeder Hügel der horizontalen Signatur ist Kürzer als $1{,}4 \cdot Höhe_{Textzeile}$
iii. die Standartabweichung der Breite der Hügel ist kürzer als $1{,}2 \cdot Mittelwert_{Hügel}$
iv. $1{,}2 \cdot Mittelwert_{Hügel} < 0{,}11 \cdot Höhe_{Textzeile}$
v. die vertikale Signatur enthält weniger als 3 Hügel

Abbildung 6: Textzeilen nach Analyse der Projektionsprofile

2.3.6 Zusammenfügen der Vordergrundbilder

Die Ergebnisse der einzelnen Vordergrundbilder werden nun in einem Ergebnisbild zusammen gefasst. Dabei werden zwei potentielle Textregionen zu einer verschmolzen, wenn sich ihre Flächen um mehr als 70% überlappen.

2.4 Maskierung und Regionen basiertes Schwellwertverfahren

Im letzten Schritt wird die binäre Maske M_b auf den Frame F angewendet, welcher aktuell verarbeitet wird:

$$F_m(x,y) = F(x,y) \cdot M_b(x,y)$$

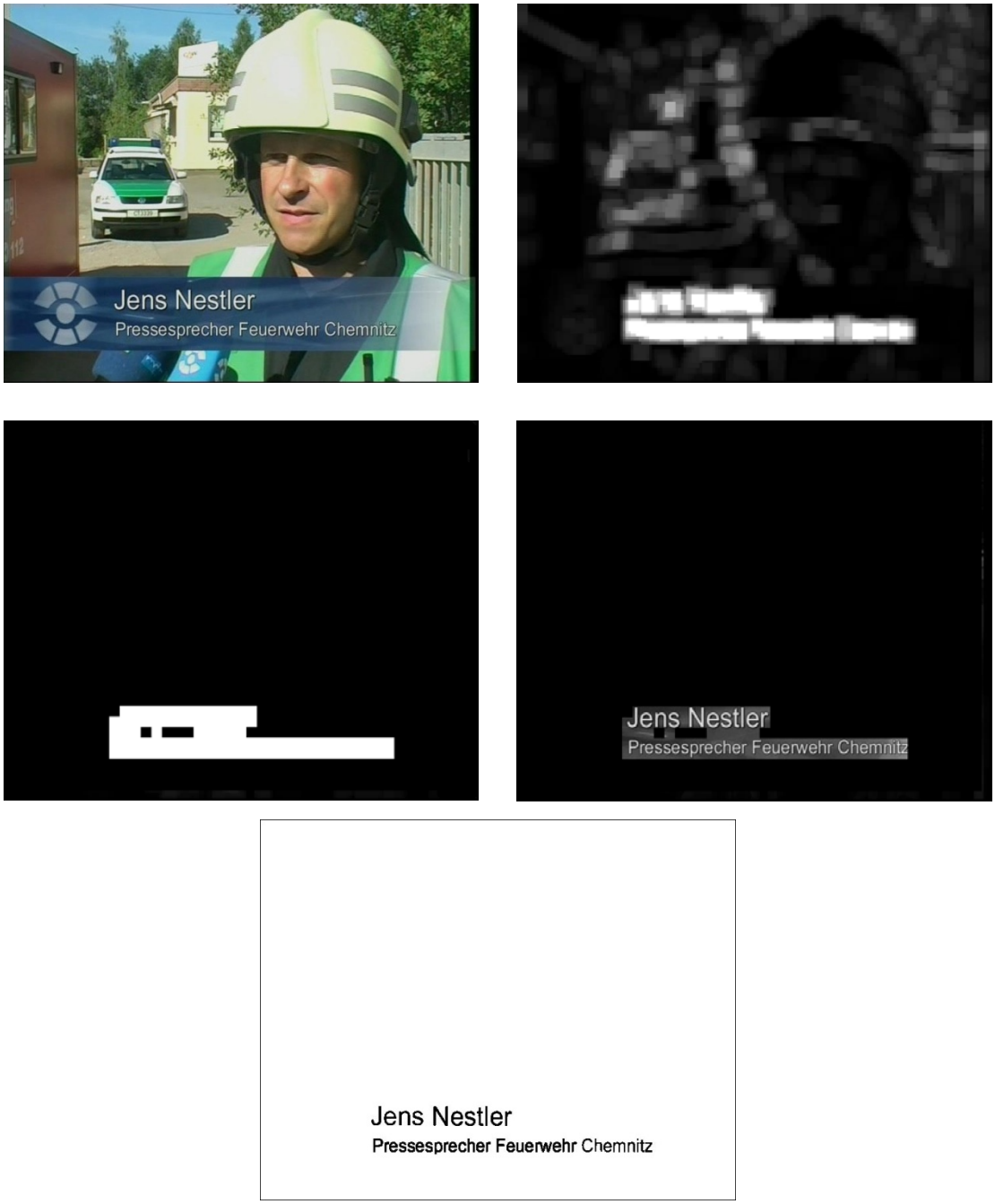


Abbildung 7: Sequentielle Frameverarbeitung (a)-(e) (von links nach rechts, von oben nach unten): (a) original Frame, (b) normalisierte Maske, (c) n. Maske nach globalem Threshold, (d) Maskierungsergebnis, (e) Resultat für OCR-API

F_m repräsentiert den verarbeiteten Frame. In diesem Frame wurden potentielle Textbereiche verstärkt und Bereiche ohne Text eliminiert. Für eine weitere Verbesserung des binären Frames eine ideale Repräsentation für die OCR zu erhalten muss ein weiteres Schwellverfahren angewendet werden. Diesmal werden nur kleine Teile des Bildes betrachtet. Hierbei wird ein Fenster mit festgesetzter Größe (50×50 Pixeln) über den ganzen Frame geschoben. Für jeden Bildausschnitt wird ein Schwellwert errechnet. Es werden nur Werte über 0 verarbeitet. Sollte ein Fenster keine relevanten Pixel aufweisen wird es verworfen. In der Abbildung 7 sind einige Beispiele dargestellt.

2.5 Texterkennung

Zur Texterkennung wurde eine OCR-API verwendet. Es handelt sich dabei um die tesseract-API. Diese API liefert sehr gute Resultate. Weiterhin wurden Omnipage und Finereader getestet. Diese kommerziellen Produkte können nicht verwendet werden, da sie auf einer GUI-Eingabe basieren und somit keine automatisierte Verarbeitung zulassen.

3 Fazit

Es wurde eine Verarbeitungskette erstellt für die Lokalisierung und Extraktion von Text in Videos verwendet wird. Die Kombination aus der gewichteten DCT und den verschiedenen Schwellwertverfahren zeigt gute Ergebnisse, wenn das System auf eine spezielle Art von Overlaytext justiert wurde. So müssen um den Text von jedem Fernsehsender korrekt erkennen zu können die Algorithmen angepasst werden. Die Mehrwertige Bildzerlegung verhält bei solchen Problemen ähnlich und bedarf ebenfalls einer Justierung. Der bisherige Evaluationskorpus beträgt 30 Stunden Video verschiedener Fernsehsender. Die bisherige Evaluation auf den Testdaten ergab keine zufriedenstellenden Ergebnisse. Aus dem Grund ist eine Anpassung der Algorithmen und Parameter notwendig.

4 Zukünftige Arbeit

Die nächsten Schritte auf dem Gebiet sind Anpassungen für das verschiedene Fernsehmaterial zu finden und diese zu testen. Der Evaluationskorpus wird feiner strukturiert um differenziertere Probleme bearbeiten zu können. Später sollen die Ergebnisse zur Unterstützung der Sprechererkennung genutzt werden. Die Texterkennung bildet somit einen Teil des in Entwicklung stehenden Retrieval-Systems. Dieses Multimedia-Retrieval-System wird zur internen (Be- und Verarbeitung sowie Annotation) und externen (Vermarktung) Wiederverwendung der Archive von unseren Projektpartnern benutzt.

5 Literaturverzeichnis

[AKJ98] Jain, A. K., & Bin, Y. (1998). Automatic text localization in images and video frames. *Pattern Recogniton*, (S. 2055-2076).

[Ju04] Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information eytraction in images and videos. *Pattern Recognition*, (S. 977-997).

[Lim00] Lim, Y.-K., S.-H.Choi, & Lee, S.-W. (2000). Text extraction in mpeg crompressed video for content-based indexing. *International Conferenceon Pattern Recognition*, (S. 409-412).

[Lu08] Lu, S., & Barner, K. E. (2008). Weighted dct based text detection. *Acoustics, Speech and Signal Processing*, (S. 1341-1344).

[Par97] Parker, J. R. (1997). Algorithms for image processing and computer vision. *Wiley Computer Publishing* .

[XQi07] Qian, X., Liu, G., Wang, H., & Su, R. (2007). Text detection, localization and tracking in compressed video. *Signal Processing: Image Communication*, (S. 752-768).

[Shi06] Shiratori, H., Goto, H., & Kobayashi, H. (2006). An efficient text capture method for moving robots unsing dct feature und text tracking. *Pattern Recognition*, (S. 1050-1053).

[Zho00] Zhong, Y., Zhang, H., & Jain, A. K. (2000). Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Mashine Intelligence*, (S. 385-392).

Sprechererkennungssystem auf Basis der Vektorquantisierung mit Störgeräuschfilterung

Stephan Heinich

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

stephan.heinich@informatik.tu-chemnitz.de

Zusammenfassung: In diesem Artikel beschreibe ich die bisherige Arbeit an einen Sprechererkennungssystem. Es wurde Wert auf eine Signalvorverarbeitung gelegt, die überflüssige Anteile aus dem Signal filtert. So werden Hintergrundrauschen und Musik sowie stimmlose Laute gefiltert. Das Sprechererkennungssystem benutzt eine auf dem Mel-Frequenzspektrum basierende Cepstralanalyse als Merkmalsextraktion.
Als Erkenner wird die Vektorquantisierung herangezogen. Der SYSDATA-Algorithmus, basierend auf dem k-Means, trainiert den Datensatz. Es ist durch die Filterung möglich stimmhafte Laute von allen anderen zu trennen. Diese Tatsache macht das entstandene Sprechererkennungssystem sehr robust.

Schlagwörter: Sprechererkennung, SYSDATA, Vektorquantisierung, ASR, Fouriertransformation

1 Einleitung

In dem Projekt sachsMedia versuchen wir die Archive von lokalen Fernsehsendern zu analysieren und sie für die Firmen und das Fernsehpublikum zugänglich zu machen. Um ein multimediales Retrieval System zu entwickeln, verwenden wir alle Modalitäten die Videodaten bereitstellen: Text-, Sprache- und Bildinformationen. Die Größe der Datenkollektion beträgt ungefähr einen Terrabyte. Der Inhalt des Videomaterials kann zum größten Teil als Nachrichtensendung kategorisiert werden. Sie beinhaltet trotzdessen eine große Menge an Werbematerial. In diesem Paper beschreibe ich verschiedene Methoden zur Erkennung von Sprechern. Der Aufbau des Papers ist im Folgenden erläutert. Im zweiten Kapitel wird Signalvorverarbeitung, low-level-feature-Extraktion und Signalfilterung behandelt. Desweiteren wird die Merkmalsextraktion mit Hilfe der Mel-Frequenz-Cepstralanalyse erläutert. Im Abschnitt 2.2 wird das Verfahren zum Training und Klassifikation der Merkmalsvektoren vorgestellt. Es handelt sich dabei um die Vektorquantisierung mit einen SYSDATA-Kernel. In den beiden letzten Kapiteln wird ein Zusammenfassung und ein Ausblick in zukünftige Arbeit gegeben.

Die Wissenschaft beschäftigt sich schon seit langer Zeit mit dem Thema der Sprechererkennung. Dabei gibt es zwei Arten dieser Systeme. Die erste ist eine wortabhängige Erkennung von Sprechern. Diese Art der Sprechererkennung findet Anwendung in Systemen bei denen eine biometrische Authentifizierung notwendig ist. Die Sprecher müssen vorgegebene Sätze oder zufällige Zahlenfolgen vorlesen. Somit können keine Aufzeichnungen verwendet werden. Bei einer Annotation beziehungsweise Indexierung von Videos in denen Sprecher identifiziert werden sollen ist eine solche Art der Erkennung eher ungeeignet. Im Bereich der Videoannotation werden wortunabhängige Sprechererkennungssysteme verwendet. Zu diesen Systemen zählt die Vektorquantisierung, die viel Anwendung in der Bildverarbeitung und im Textretrieval findet.

Die Vektorquantisierung ist eine klassische Methode zur Sprechererkennung. Sie basiert auf Cluster-Algorithmen. Der gebräuchlichste Algorithmus ist k-Means. Ich benutzte SYSDATA als Kern der Vektorquantisierung. Dies ist ein abgeänderten Algorithmus.

Um das Training zu verbessern und die Erkennungsrate zu steigern wird das Audiosignal vor der Merkmalsextraktion gefiltert. So werden Hintergrundgeräusche, Musik, Rauschen und stimmlose Laute zu verschiedenen Verarbeitungsschritten aus dem Signal gefiltert. Anhand vom MPEG7-Standart wurden low-level Features bestimmt. Unter diesen Merkmalen befindet sich die Zero-Crossing-Rate, die Spektral-Rollover-Frequency sowie die spektrale Energie. Diese Merkmale werden zum Ausschluss von störenden Signalteilen verwendet.

2 Audiosignalverarbeitung

Die folgenden Schritte werden zur Audiosignalverarbeitung genutzt. Im ersten Schritt werden über ein mehrstufiges Filtersystem bestimmt Anteile, wie Sprachpausen oder Rauschen, aus dem Audiosignal eliminiert. Im zweiten Schritt erfolgt eine Transformation in Frequenzspektrum die von einer erneuten Filterung gefolgt wird. Mit Hilfe der Cepstralanalyse werden Merkmale aus dem Audiosignal gewonnen. Diese müssen normalisiert und gewichtet werden. Es folgt ein Training der Daten unter der Verwendung einer Vektorquantisierung. Die Klassifizierung erfolgt auf den berechenten Codebüchern der Vektorquantisierung.

2.1 Signalvorverarbeitung

Um Merkmale aus dem Audiosignal extrahieren zu können, muss es zuvor verarbeitett werden. So wird das Signal neu gesampelt. Anschließend wird anhand Amplitudenschwellwerte ermittelt ob Sprachpausen enthalten sind. Es folgt eine Fensterung und die Ermittlung der Zero-Crossing-Rate. Zur Vorverarbeitung gehört auch eine Transformation ins Frequenzspektrum sowie die Bestimmung der spektralen

Rollover-Frequency. Danach wird das Signal in das Mel-Frequenzspektrum überführt und anschließend einer Cepstral-Analyser unterzogen, bei der die notwendigen Merkmale berechnet werden.

2.1.1 Signalresampling

Um die weitere Verarbeitung der Daten möglichst zu vereinfachen wurde das Resampling des Signals durchgeführt. Diese Idee stammt von Lu, et al. [LuL02] . Der Audio-Stream wird auf 10 kHz, 16bit pro Sample und Monokanal in ein uniformes Format umgewandelt. Dadurch sind keine variablen Anpassungen an weitere Verarbeitungsschritte nötig. Da meist die Signale herunter gerechnet werden, werden bei der Verarbeitung Rechenaufwand und Speicherplatz gespart. Es wird von der in [LuL02] benutzten 8 kHz Samplingrate abgesehen, da die gewählten 10 kHz größere Vorteile bei der Unterteilung des Sprachsignals haben. Das Signal kann einfacher in Fenster von 20 ms unterteilt werden. Das bedeutet, dass jedes Fenster aus 200 Samples besteht. Der einzige Nachteil der bei dem Umwandeln entsteht, ist der Datenverlust des Audio-Streams, der aber zu keiner gravierenden Verschlechterung der Erkennungsergebnisse führt. Spätere Anpassungen der Samplerate sind ohne weitere Probleme durchführbar.

2.1.2 Erkennen und Markieren von Sprachpausen

Ein gutes Hilfsmittel in der Sprechererkennung ist das Herausfinden von Sprachpausen [LEG03] . Diese Pausen werden markiert und müssen nicht in die Analyse einfließen. Anhand der Sprachpausen können Übergänge zwischen Sprechern leichter erkannt werden. Die einfachste Methode der Sprachpausenerkennung ist eine Schwellwertfunktion. Im Sprachpausenvektor b werden die Daten nach der Form

$$b(n) = \begin{cases} 1; s(n) > L \\ 0; s(n) < L \end{cases}$$

berechnet. Der Schwellwert L gibt eine relative Lautstärke an, ab der die Sprache von den Hintergrundgeräuschen unterschieden wird.

2.1.3 Fensterung

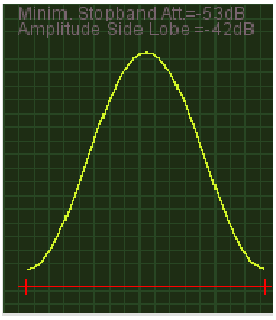
Da das Signal nicht im Ganzen analysiert werden kann, wird es in Segmente unterteilt. Diese Segmente werden Fenster genannt. Es hat sich herausgestellt, dass sich Fenster mit einer Fensterbreite *M* von 10-20 ms am besten verarbeiten lassen. Um weiterhin gute Ergebnisse zu erzielen, werden die Fenster, überschnitten.

Da bei ungewichteten Fenstern Probleme entstehen können, werden Fensterfunktionen w(n) angewendet. Im Folgenden die Funktionen dargestellt, die den Segmenten Gewichte gibt, welche auf das Signal durch die Vorschrift

$$s'(n) = w(n) \cdot s(n)$$

multipliziert wird.

Angewendet wurde das Hamming-Fenster. Bei dieser Fensterfunktion entstehen nahezu keine Nebeneffekte bei Überlagerung der Fenster.



$$w(n) = 0.54 + 0.46 \cdot \cos\left(\frac{2\pi n}{M}\right)$$
$$\text{mit } n = -\frac{M}{2}, \ldots, \frac{M}{2}$$

Abbildung 1: Hamming-Fenster

2.1.4 Zero-Crossing-Rate-Filter

Mit Hilfe der Zero-Crossing-Rate (ZCR) [Kim05] kann ein sehr nützlicher Filter erzeugt werden. Die ZCR zählt die Häufigkeit, in der ein Signal in einer bestimmten Zeit die 0-Achse kreuzt. Sie lässt sich mit der Formel

$$ZCR = \frac{1}{2}\left(\sum_{n=1}^{N-1} \left|\text{sign}\big(s(n)\big) - \text{sign}\big(s(n-1)\big)\right|\right)\frac{F_s}{N}$$

beschreiben. *N* gibt die Anzahl der Sampels in *s*(*n*) an. F_s ist die Samplingfrequenz. Die Signum Funktion sign(x) ist wie folgt definiert:

$$\text{sign}(x) = \begin{cases} 1, & wenn\ x > 0 \\ 0, & wenn\ x = 0 \\ -1, & wenn\ x < 0 \end{cases}$$

Da die ZCR anhand der Zeitdarstellung und nicht anhand der Spektraldarstellung gewonnen werden kann, kann sie im Vorfeld Rechenaufwand sparen, der für die Fast-Fourier-Transformation verwendet wird. Harmonische Signale überqueren die 0-Achse

mit einer recht stabilen Frequenz. Stimmhafte Laute weisen eine ZCR von 500-2000 Hz auf und passieren den Filter. Stimmlose Laute wie zum Beispiel F- und H-Laute weisen ein rauschähnliches Muster auf, bei dem die ZCR unter 450 oder meist weit über 2500 Hz liegt.

2.1.5 Zero Pattern Fast Fourier-Transformation

Die meisten Informationen können aus dem Spektrum oder auf dem Spektrum basierenden Transformationen gewonnen werden. Die schnelle Fourier-transformation (Fast Fourier Transformation-FFT) wird benutzt, um ein Audiosignal in die spektrale Darstellung zu transformieren. Um eine gute Auflösung des Spektrums zu erhalten, sollte die Länge mindestens 512 Samples betragen. Das bedeutet für ein Fenster von 20 ms muss das Signal mit einer Samplingrate von 25600 Hz codiert sein. Mit der Erhöhung der Samplingrate steigt auch der Rechenaufwand und Speicherbedarf. Aus dem Grund wird die Zero Pattern Fast Fourier Transformation (ZPFFT) angewendet. Bei dieser Methode ist es vollkommen ausreichend ein Signal mit 10000 Hz zu codieren. 20 ms würden bei dieser Kodierung 200 Samples entsprechen. Es werden nun die verbleibenden Stellen mit Nullen aufgefüllt.

Dabei treten allerdings Artefakte im Spektrum auf. Diese zeigen sich durch Lücken. Die Lücken können aber mit einer entsprechenden Funktion geschlossen werden. Als einfachste Methode bietet sich eine Art Weichzeichner an, der die aktuelle Position im Spektrum aus dem Mittelwert der umliegenden Spektralwerte errechnet.

2.1.6 Peakberechnung und spektrale Energie

Der Peak ist das einfachste zu berechnende Merkmal im Spektrum. Er ist aber auch sehr aussagekräftig. Der Peak bestimmt den größten Wert im Spektrum. Dieser Wert kann mit einer Maximum-Operation bestimmt werden. Er verändert sich nur sehr gering von Sprecher zu Sprecher und wird daher auch in den Merkmalsvektor aufgenommen werden.

Die spektrale Energie wird über ein drei geteiltes Spektrum errechnet. Der erste Teil ist der niederfrequente Bereich, der sich von 100-2000 Hz erstreckt. Der zweite ist der mittlere Teil. Er umfasst alle Frequenzen zwischen 2000-6000 Hz und der dritte Teil betrifft den oberen Frequenzbereich von 6000-10000 Hz. Als weiteren spektralen Energiewert wird die Summe aus allen drei Bereichen gebildet. Diese Unterteilung macht eine spätere Verarbeitung einfacher. Da sich bei Sprache die meiste Energie im niederfrequenten Bereich befindet, braucht meist nur dieser ausgewertet werden. Die spektrale Energie kann auch als Filter verwendet werden.

2.1.7 Spectral-Rollover-Frequency-Filter

Laut Kim, et al. [Kim05] gibt die spektrale Rollover-Frequenz (SRF) eine Grenzfrequenz an. Unterhalb dieser Frequenz befindet sich ein angegebener Prozentsatz der gesamten spektralen Energie. In der Regel setzt man die Grenze auf 85 %. Bei Sprache bzw. stimmhaften Lauten liegt die SRF bei 1500 bis 3000 Hz. Musik oder Hintergrundgeräusche weisen eine viel höhere SRF auf. Die SRF wird wie folgt errechnet:

$$\sum_{k=0}^{K_{roll}} |S(k)| = 0.85 \sum_{k=0}^{N_{FT}/_2} |S(k)|$$

S(k) gibt das Spektrum für ein Fenster an, K_{roll} die Ordnung des Spektrums bis zur Rollover-Frequenz und N_{FT} die Ordnung des Spektrums.

Die ROF gibt Aufschluss über harmonische Signalteile und Rauschen. Im Allgemeinen ist die ROF bei harmonischen Signalen wesentlich kleiner als beim Rauschen. Mit Hilfe der Rollover-Frequenz können ähnlich wie bei der ZCR stimmhafte Laute von Geräuschen oder Musik unterschieden werden. Um andere Filter zu unterstützen, kann zusätzlich die SRF als Filter verwendet werden.

2.1.8 Tief- und Hochpassfilter

Die bekanntesten Audiofilter sind der Tief- und Hochpassfilter. Bei dem Tiefpassfilter passieren tiefe Frequenzanteile, das heißt der hochfrequente Teil wird gefiltert. Bei dem Hochpassfilter ist es genau umgekehrt. Es gibt verschiedene Arten dieser Filter. Die erste Art ist gleichzeitig die Einfachste. Bei dem Tiefpassfilter werden die Frequenzanteile ab einer angegebenen Frequenz abgeschnitten bzw. auf 0 gesetzt:

$$S(n) = \begin{cases} S(n), & wenn\ n \cdot FA < TF \\ 0, & wenn\ n \cdot FA > TF \end{cases}$$

S(n) ist das Spektrum eines Fensters wobei *n* ein Element des Spektrums ist. *FA* ist die Frequenzauflösung und *TF* die Trennfrequenz. Die zweite Art beschreibt eine lineare Abnahme der Frequenzenergie in einem Trennfrequenzbereich. Das heißt von einer Anfangstrennfrequenz bis zu einer Endtrennfrequenz nimmt die Frequenzenergie linear ab, danach ist sie 0:

$$S(n) = \begin{cases} S(n), & wenn\ FA < TFA \\ S(n) \cdot \left(\frac{n}{TFE - TFA} + \frac{TFA}{TFE - TFA}\right), & wenn\ FA > TFA\ \wedge FA < TFE \\ 0, & wenn\ FA > TFE \end{cases}$$

Bei der linearen Abnahme der Frequenzenergie steht TFA für die Anfangsfrequenz des Trennbereiches und TFE für das Ende, n ist die aktuelle Frequenz. Der Vorteil der letzten Methode ist, dass keine Kanteneffekte bei der weiteren Verarbeitung auftreten. Dieser Filter wird verwendet um kritische Bereiche bzw. Bereiche ohne nutzbare Information zu eliminieren.

2.1.9 Allgemeine Cepstralanalyse

Zur Berechnung der Merkmalsvektoren wird die Cepstralanalyse verwendet. Diese versucht die Periodizität der Grundfrequenz von der Impulsantwort zu trennen. Dabei hilft eine Logarithmierung des Frequenzspektrums nach Mel:

$$m = 1127.0148 \cdot \log\left(1 + \frac{f}{700Hz}\right)$$

Der lineare Anstieg der Frequenz bis 1 kHz ist bei der Berechnung der Grundfrequenz von Vorteil. So treten bei Harmonien im Spektrum regelmäßige Abstände zwischen Magnitudenerhöhungen auf. Dieser Abstand beschreibt die Grundfrequenz. Sie kann mit dem Mel-Spektrum somit innerhalb der ersten 1 kHz ermittelt werden. Zur Berechnung der Impulsantwort wird der Bereich über 1 kHz betrachtet. Das Cepstrum lässt sich mit der Formel

$$c(q) = \log\left|S[\omega]\right|^2 \cdot \cos(\omega \cdot q) \cdot d\omega.$$

Die Cepstralkoeffizienten werden im Wesentlichen durch folgende Schritte berechnet. Das ist erstens die Unterteilung des Eingabesignals in "Windows" (z. B. Hamming-Fenster um Kanteneffekte zu vermeiden). Dabei sind überlappende Fenster üblich. Zweitens wird die diskrete Fouriertransformation jedes einzelnen Fensters verwendet. Drittens werden die Fourierkoeffizienten logarithmiert. Dieser Schritt wurde durch die Einsicht motiviert, dass Lautstärke vom menschlichen Ohr in etwa logarithmisch wahrgenommen wird. Des Weiteren wird dadurch die Multiplikation von Anregungssignal und Impulsantwort in eine Addition transformiert. Viertens wird die Anzahl der Frequenzbänder (z.B. 256) durch Zusammenfassen (auf z.B. 40) reduziert. Und fünftens verwendet man eine abschließende Dekorrelation durch die besprochene Diskrete Cosinustransformation. Ursprünglich wurden die logarithmierten Fourierkoeffizienten invers fouriertransformiert. Die Anregungsfrequenz ist dann ein einzelner hoher Peak und leicht zu erkennen bzw. herauszufiltern.

2.1.10 Ableitung des Cepstrum

In den bisherigen Ausführungen wurden die Analyseergebnisse für jeden Signalabschnitt isoliert betrachtet. Die Beziehungen zwischen aufeinander folgenden

Blöcken blieben in den Merkmalsvektoren bisher unberücksichtigt und ihre Nutzung den nachfolgenden Klassifikationsverfahren vorenthalten. Beim Betrachten eines zeitlichen Merkmalverlaufes der ersten beiden Cepstralkoeffizienten c_1 und c_2 ensteht eine starke zeitliche Variabilität. In den zeitlichen Veränderungen der Merkmale scheinen Informationen über die spektrale Variation enthalten zu sein, die beispielsweise aus der Verknüpfung aufeinander folgender Merkmale bestimmt werden können. Da die einfache Ableitung der Koeffizienten eher den Charakter eines Rauschens besitzt, verwendet man eine Regression über einen längeren Zeitraum. Diese Näherung bezeichnet man als Delta-Cepstralkoeffizienten und wird mit

$$\Delta c_n(t) = \frac{\sum_{k=1}^{K} k \cdot h_k \cdot c(t+k)}{\sum_{k=1}^{K} h_k \cdot k^2}$$

beschrieben. Untersuchungen zum Einfluss der Fensterlänge ergaben, dass eine Fensterlänge von insgesamt ca. 160 - 200 ms die besten Resultate liefert.

2.1.11 Bestimmung der Merkmalsvektoren

Der Merkmalsvektor lässt sich laut [JZi97] aus den gewonnen Cepstralkoeffizienten und den Delta-Cepstralkoeffizienten zusammenstellen. Des Weiteren wird der Energieverlauf als Merkmal in den Vektor aufgenommen. Der Energieverlauf ist die erste zeitliche Ableitung der Energie des aktuellen Fensters. Man bezeichnet das Ergebnis als Delta-Energiekoeffizient. Aus der ersten Ableitung gewinnt man nun die zweite Ableitung, die ausschließlich für den Energieparameter berechnet wird. Diese zweite Ableitung wird Delta-Delta-Energiekoeffizient genannt. Da der Merkmalsvektor auf 26 Komponenten beschränkt wird, legt man folgende Belegung fest:

- 12 Cepstralkoeffizienten $(c_0 - c_{12})$

- 12 robuste Delta-Cepstralkoeffizienten $(\Delta c_0 - \Delta c_{12})$

- 1 robuster Delta-Energiekoeffizient (ΔE)

- 1 robuster Delta-Delta-Energiekoeffizient $(\Delta\Delta E)$

Jeder Merkmalsvektor entspricht einem Fenster. Bei einer Fensterlänge von 20 ms und einem überlappenden Fenster wird alle 10 ms ein Merkmalsvektor bestimmt. Damit erhält man aus einer Sekunde Sprachsignal 100 Merkmalsvektoren mit jeweils 26 Komponenten.

2.1.12 Spektrale Distanzmaße und Merkmalsnormierung

Bei der Erkennung durch Kurzzeitspektren werden im Allgemeinen lokale Distanzmaße zwischen Merkmalsvektoren verwendet [JZi97] . Verbreitet sind die Ähnlichkeitsmaße $(l_1 - Norm)$ oder der quadrierte euklidische Abstand $(l_2 - Norm)$. Die Norm des Ähnlichkeitsmaßes wird auch als City-Block-Abstand bezeichnet und in der Form

$$d(x_i, y_j) = \sum_{k=1}^{K} w_k^2 \cdot |x_k(i) - y_k(j)|$$

dargestellt. Insbesondere wird die $l_1 - Norm$ eingesetzt, um eine Quadrierung auf üblichen Prozessoren mit erhöhtem Rechenaufwand zu umgehen. Die $l_2 - Norm$ wird mit

$$d(x_i, y_j) = \sum_{k=1}^{K} w_k^2 \cdot (x_k(i) - y_k(j))^2$$

errechnet.

Für eine Klassifikation werden Wichtungswerte ermittelt und in die Koeffizienten eingerechnet,

$$d(x_i, y_j) = \sum_{k=1}^{K} (w_k x_k(i) - w_k y_k(j))^2 = \sum_{k=1}^{K} (x_k(i) - y_k(j))^2 \quad .$$

Das spart Rechenzeit bei der Ermittlung der Distanzmaße. Zur Festlegung der Wichtungswerte wird sich zurzeit noch an bekannte Untersuchungen zur Merkmalsstatistik orientiert.

Die Varianzen der Werte von Cepstral- und Delta-Cepstralkoeffizienten sind umgekehrt proportional zum Quadrat der Koeffizientenordnungen. Um nun die höheren Koeffizientenordnungen stärker in die Erkennung einfließen zu lassen, werden die Varianzen gewichtet. Dies geschieht mit Hilfe der Varianznormierung,

$$w_k = \frac{1}{\sigma_k^2} \quad \text{mit } \sigma_k^2 = E(c_k - \mu_k)^2,\ k = 1,...,K\,.$$

Es werden μ_k als Mittelwert, σ_k^2 als Varianz aus einer langfristigen Beobachtung von Sprachdaten, und *E* als Erwartungswert bezeichnet.

Ein besser und schnellere Methode ist Gewichtung in Form einer Sinusfunktion mit

$$w_k = 1 + h \sin\left(\frac{\pi k}{l}\right), \; k = 1,...,8 \text{ und } h = 6, \; l = 12.$$

Für niedrige Indexwerte verhalten sich alle drei Normierungsarten etwa gleich: Sie dämpfen den Einfluss der Koeffizienten mit hoher Varianz. Die Sinusfunktion bewertet die höchsten Koeffizientenordnungen nicht mehr so hoch wie die beiden anderen Normierungsarten. Bei einem längeren Merkmalsvektor mit $k > 8$ ist die Länge des sinusförmigen Wichtungsfensters anzupassen, um denselben Effekt zu erhalten.

2.2 Vektorquantisierung

Zum Training und zur Klassifizierung der Sprecher wird die Vektorquantisierung verwendet. Die Vektorquantisierung ist ein wortunabhänger Erkenner. Den Kern der Vektorquantisierung bildet SYSDATA-Algorithmus. Wobei es sich hier um eine veränderte Variante des k-Means Algorithmus handelt. Klassifiziert wurde mit Hilfe der $l_2 - Norm.$

2.2.1 SYSDATA

SYSDATA kann als Erweiterung des LBG-Algorithmus als auch des k-means verwendet werden. Es vermeidet durch eine zweite Iteration das Problem der Initialisierung der Codevektoren.

1. **Initialisierung:** Es wird ein Codebuch aus nur einem Codevektor erstellt. Dieser Codevektor ist der Zentroid des gesamten Satzes an Trainingsvektoren. Man bezeichnet ihn als Klassenzentrum.

2. **Erweiterung des Codebuchs:** In dem Schritt wird die Codebuchgröße durch Splitten der Zentroide verdoppelt. Ein Störvektor wird berechnet von einem Codevektor abgezogen und auf seinen Doppelgänger aufaddiert.

3. **Hauptiteration:** Es werden nun die Codebücher mit Hilfe von k-means neu berechnet, um den besten Satz der Codevektoren zu ermitteln.

4. **Terminierung:** Die Schritte 2 und 3 werden solange wiederholt, bis das Codebuch die gewünschte Größe erreicht hat.

Da bei der Verdopplung der Zentroide ein sehr kleiner Störvektor aufaddiert bzw. abgezogen wird, entstehen zwei neue unterschiedliche Zentroide, die die vorhandene Menge teilen können. Durch ständige Teilen und Neuberechnen der Zentroide werden die Bereiche optimal geteilt und zugewiesen.

2.2.2 Klassifikation

Zur Klassifizierung von Testsequenzen kann man die Gesamtverzerrung der Observation in Bezug auf die Codebücher aller Klassen bestimmen. Die Testsequenz wird einer Klasse $\hat{S}$ zugeordnet. Die Gesamtverzerrung D_k kann nun anhand der Gleichung (3.5) verwendet werden, um die wahrscheinlichste Klasse $k = 1...S$ zu bestimmen, die die Testsequenz klassifiziert:

$$\hat{S} = \arg\min_{1 \le k \le S} D_k(X)$$

Dies setzt aber voraus, dass bekannt sein muss, wann eine Sequenz eines Sprechers beginnt und wann sie endet. Da dies in manchen Anwendungsgebieten nicht vorhersehbar ist, muss auf eine erweiterte Methode zurückgegriffen werden. Zur Abgrenzung der Testsequenzen können die mit Hilfe der Sprachpausendetektion ermittelten Sprachpausenmarken verwendet werden. Das sich zwischen zwei Marken befindliche Sprachsignal kann mit dem obigen Verfahren klassifiziert werden. Sollte das Material, welches sich zwischen den nächsten Pausen befindet, ähnliche Ergebnisse erzielen, kann es mit der vorhergehenden Sequenz verbunden und erneut klassifiziert werden. Diese etwas zeitaufwendigere Methode bringt unter Umständen zuverlässigere Trefferquoten hervor.

3 Fazit

In dieser Arbeit wurde versucht die Sprache auf das Wesentliche zu reduzieren, um Störgeräusche zu eliminieren. Dabei wurden alle Signalteile die keine stimmhaften Laute sind aus dem Signal entfernt. So konnten Musik, Hintergrundgeräusche sowie Plosivlaute und stimmlose Laute im Audiomaterial durch die Filter eliminiert werden. Mit Hilfe dieser Methoden ist es möglich die Qualität der Klassifizierung durch bewährte Methoden zu verbessern und den Rechenaufwand zu verringern. Es muss untersucht werden ob für andere Methoden eine Filterung von stimmlosen Lauten nötig beziehungsweise sogar hinderlich ist, wenn das Trainingsmaterial optimal ist.

4 Zukünftige Arbeit

In der Zukunft soll ein umfangreiches Trainings- und Evaluierungsset, aus verschiedenen Videobeiträgen unterschiedlicher Fernsehsender, erstellt werden auf diesem klassischen Methoden getestet werden sollen. Weiterhin werde ich an Methoden forschen die einen komplett neuen Ansatz verfolgen. Es handelt sich um die Frequenz- und Cepstralmodulierung. Bei dem Signale auf ihre Langzeitmuster untersucht werden. Es ist dadurch möglich spektrale Muster in zweidimensionaler Form darzustellen.

Bei der Frequenzmodulation wird zusätzlich zum Frequenzspektrum ein Spektrum einzelner Frequenzbänder ermittelt. Dieses zweidimensionale Muster lässt sich nun auch mit Bildverarbeitungsalgorithmen bearbeiten. So ist es möglich durch Maskierung bestimmter Bereiche gleichzeitig sprechende Sprecher heraus zu filtern. Weiterhin wäre es möglich die Frequenzmodulationen der Sprecher als zweidimensionales Muster für gaußsche Mischverteilungsmodelle zu verwenden. Unter Zuhilfenahme der Texterkennung von Bauchbinden kann die Sprechererkennung zusätzlich erweitert werden. Ein weiterer Vorteil bei diesen Methoden ist die entstehende zweidimensionale Darstellung, auf der sich unter Umständen herkömmliche Bildverarbeitungsalgorithmen anwenden lassen. Denkbar ist auch eine Verwendung für dieses Verfahren auf Cepstralkoeffizienten. Zur Klassifizierung sollen Gaußsche Mischverteilungsmodelle eingesetzt werden. Die eine weiche Abschätzung zu anderen Sprechern ermöglichen. Desweiteren wird die Sprechererkennung mit der Texterkennung kombiniert um genauere Ergebnisse zu erzielen. Dabei sollen beide Prozesse voneinander profitieren können.

5 Literaturverzeichnis

[Kim05] Kim, H.-G., Moreau, N., & Sikora, T. (2005). *MPEG-7 Audio And Beyond.* Chichester, West Sussex: John Wiley & Sons Ltd.

[LEG03] LEGAmedia. (2003). *Cepstralanalyse Algorithmen*. Abgerufen am 05. Dezember 2006 von Cepstralanalyse Algorithmen: http://www.oweiss.com/articles/sprachanalyse_04.htm

[LuL02] Lu, L., & Zhang, H.-J. (2002). Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis. In *Multimedia '02.* Juan-les-Pins, Frankreich: ACM.

[JZi97] Zinke, J. (1997). *Sprachanalyse, Grundverfahren zur Sperchererkennung*. Abgerufen am 14. November 2006 von Sprachanalyse, Grundverfahren zur Sperchererkennung: http://www.fh-freiberg.de/fachbereiche/e2/telekom-labor/zinke/digiaudi/digiaudi.htm

Metadatenstandards und –formate für audiovisuelle Inhalte

Jens Kürsten

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

jens.kuersten@informatik.tu-chemnitz.de

Zusammenfassung: Der vorliegende Artikel gibt einen Überblick zu bestehenden Metadatenstandards für audiovisuelle Daten. Neben den etablierten Standards zur semantischen Beschreibung der Inhalte werden Formate betrachtet, die für B2B oder B2C Anwendungen in der Fernsehproduktion verwendet werden. Die Untersuchung soll klären, welche Metadatenformate für eine Verwendung zum Programmaustausch zwischen lokalen Fernsehstationen, den Einsatz in einem vernetzten Archiv oder zur Weiterverwertung von produzierten Inhalten in unterschiedlichen Distributionskanälen sinnvoll einsetzbar sind.

Schlagwörter: Metadatenstandards, Multimedia Retrieval, Broadcasting

1 Motivation

Im InnoProfile Projekt sachsMedia[1] werden unter anderem Methoden für die automatisierte und automatische Annotation von audiovisuellen Inhalte erforscht und prototypisch umgesetzt. Diese sollen in Zukunft für eine zur Suche in einem lokalen oder vernetzten Archiv verwendet werden. Für eine funktionell überzeugende Recherche in audiovisuellem Material sind die zur Suche verwendeten Metadaten der Schlüssel. Dies können neben den angesprochenen, automatisch erzeugten Metadaten auch alle zusätzlichen Daten sein. Dazu zählen während der Fernsehproduktion erzeugte intellektuell vergebene Metadaten genauso wie klassische, deskriptive Metadaten mit rein technischem Inhalt.

Während der Umstellung auf digitale Produktion wurden auch bei großen Fernsehstationen meist proprietäre Daten- und Metadatenformate eingesetzt. Diese waren anfangs in hohem Maße an deren spezifischen Produktionsworkflow angepasst und damit nicht generell für alle Produzenten verwendbar. Da dies jedoch für einen reibungslosen Austausch von audiovisuellen Inhalten erforderlich ist, wurde um die Jahrtausendwende mit der Entwicklung von standardisierten Metadatenformaten für die Fernsehproduktion begonnen. Dieser Artikel soll die wesentlichen Entwicklungen darstellen. Dabei wird explizit auf Standards und Formate eingegangen, bei denen deskriptive,

[1] Gefördert durch: Unternehmen Region, Die BMBF Innovationsinitiative Neue Länder

inhaltbeschreibende Metadaten im Vordergrund stehen. Ziel der Gegenüberstellung ist es, Formate und Standards zu identifizieren, die einen Programmaustausch unter lokalen Fernsehanbietern innerhalb einer vernetzten Plattform ermöglichen. Ebenso soll bestimmt werden, welche Formate für eine Recherche-Anwendung in einem vernetzten Archiv praktikabel sind.

Nachfolgend werden bestehende Metadatenstandards und –formate zur Beschreibung audiovisueller Inhalte kurz vorgestellt und zu einander in Beziehung gesetzt. In Abschnitt 2 werden standardisierte Formate beschrieben, während in Abschnitt 3 Metadatenformate zum Austausch in der Fernsehproduktion erläutert sind. In Kapitel 4 wird die Notwendigkeit von standardisierten Regularien zur intellektuellen Beschreibung von audiovisuellen Inhalten diskutiert. Abschließend werden neben der Zusammenfassung die Einsatzmöglichkeiten der Metadatenformate für eine kooperative Fernsehproduktion in Lokalsendern aufgezeigt.

2 Metadatenstandards

Zur Beschreibung von audiovisuellen Medien existieren heute zahlreiche Metadatenstandards und –formate, meist zur Beschreibung eines speziellen Medientyps oder auch für die Nutzung innerhalb einer konkreten Anwendung. In diesem Abschnitt werden die bedeutendsten Metadatenstandards zur semantischen Beschreibung von audiovisuellen Medien kurz vorgestellt und in Bezug zueinander gesetzt.

2.1 Dublin Core

Der bekannteste Standard für Metadaten ist das Dublin Core Metadata Element Set auf das sich im Rahmen des OCLC/NCSA Metadata Workshop 1995 in Dublin/Ohio geeinigt wurde. Es besteht aus 15 Kernelementen, die mehrfach standardisiert beschrieben[2,3,4] sind. Durch die Definition eines Standards allein wird die Interoperabilität der Metadaten jedoch nicht gewährleistet. Dies wird erst durch die Definition von generalisierten Anwendungsprofilen möglich, die dann wiederum in ihrer entsprechenden Domäne Gültigkeit haben.

Der Dublin Core Standard hat sich vor allem zur Beschreibung von textuellen Ressourcen durchgesetzt und findet daher insbesondere in Rechercheanwendungen im Web Anwendung. Die Interoperabilität wird aufgrund der zahlreichen definierten Anwen-

[2] ISO 15836:2009, Information and documentation - The Dublin Core metadata element set

[3] RFC 5013, The Dublin Core Metadata Element Set

[4] NISO Z38.95, The Dublin Core Metadata Element Set

dungsprofile erreicht. Andererseits sorgt das relativ kleine Vokabular mit nur 15 Elementen auch für ein entsprechend großes Anwendungsspektrum.

2.2 MPEG-7

Mit der Entwicklung des Vorgängers von MPEG-7, dem Multimedia Content Description Interface, wurde 1996 auf einem Workshop der MPEG Arbeitsgruppe in Tampere begonnen. Wesentliche Ziele dabei waren die effiziente Identifikation relevanter Informationen, effizientes Management und eine größtmögliche Interoperabilität zwischen Anwendungen für Multimedia-Daten [Man02]. Die Veröffentlichung von MPEG-7 als ISO/IEC Standard[5] erfolgte in den Jahren 2002 und 2003. Eine komplette Beschreibung des Standards ist in [Mar04] gegeben.

Für die Verwendung des Standards in einer vernetzten Recherche-Anwendung für Fernsehbeiträge spielt dessen Interoperabilität die zentrale Rolle. Diese wird durch die Entwicklung der Anwendungsprofile und deren weitere Einschränkung auf Ebenen (Level) ermöglicht. Die Generalität und Komplexität von MPEG-7 macht die Definition dieser Profile und Ebenen für eine konkrete Anwendung unabdingbar. Dies führt jedoch dazu, dass mehrere standardkonforme Möglichkeiten zur Beschreibung von gleichem Inhalt möglich sind [Bai07]. Ferner ist daher auch eine Interoperabilität auf semantischer Ebene nicht vollständig gewährleistet.

Vor allem aufgrund seiner Komplexität, aber auch wegen den angedeuteten Problemen in Bezug auf die Interoperabilität, wird MPEG-7 in praxisnahen Anwendungen für die Recherche in audiovisuellem Material kaum eingesetzt. Dennoch hat er sich im wissenschaftlichen Bereich durchgesetzt, wird jedoch häufig nur in konkreten Teilbereichen, wie bspw. der Verarbeitung von Audiodaten[6] oder für die Extraktion von Bilddeskriptoren[7], verwendet.

2.2.1 MPEG-7 Detailed Audiovisual Profile (DAVP)

Um den beschriebenen Problemen für die Beschreibung von audiovisuellen Inhalten zu entgegnen wurde das MPEG-7 Detailed Audiovisual Profile [Bai07] entwickelt. Als Hauptanwendungsgebiete des Profils werden audiovisuelle Archive, Bild- und Videodatenbanken sowie die Produktion audiovisueller Inhalte genannt. Es enthält daher die wesentlichen Deskriptoren zur Beschreibung von Audio- und Videoinhalten (MPEG-7 Standard, Teil 3 und 4). Eine wesentliche strukturelle Eigenschaft des Profils ist die Trennung der Beschreibung für audiovisuelle Inhalte auf Basis des Medientyps, d. h.

---

[5] ISO 15938-1 bis -10, Multimedia Description Interface

[6] MPEG-7 Audio-Encoder, http://mpeg7audioenc.sourceforge.net

[7] Caliph & Emir, http://sourceforge.net/projects/caliph-emir

die Deskriptoren für Audio-, Video- und Bilddaten werden in separaten Beschreibungen abgelegt.

Aufgrund der genannten Eigenschaften ist MPEG-7 unter Verwendung des DAVPs für die Archivierung von Fernsehbeiträgen durchaus geeignet. Auf Basis der Extraktion unterschiedlicher Merkmale können diese dann zur semantischen Recherche innerhalb eines Archivs genutzt werden, welches sowohl fertige Produktionen als auch Rohmaterial beinhaltet.

2.3 DMS-1 im Material eXchange Format (MXF)

Das standardisierte MXF Dateiformat[8] ist im eigentlichen Sinne kein Metadatenstandard sondern ein flexibles Containerformat für den Produktionsprozess im Fernsehumfeld. Die Relevanz des Formats für die hier gegebene Problematik ergibt sich daraus, dass ein einheitliches Containerformat ideale Möglichkeiten für einen einfachen Dateibasierten Austausch im Produktionsprozess bietet. Neben diesem Fakt wird ein standardisiertes Metadatenmodell verwendet und durch die direkte Integration der eigentlichen Daten ist das Format unabhängig von der eingesetzten Kompression [Wel06].

Zur Beschreibung von Inhalten wird empfohlen das Descriptive Metadata Scheme 1 (DMS-1)[9] zu verwenden. Es ist in drei Frameworks unterteilt, die jeweils eine Ebene der Produktion widerspiegeln. Namentlich sind dies: a) das Production Framework, welches Metadaten zu einem gesamten Medium enthält, b) das Clip Framework, was sich auf einen fortlaufenden audiovisuellen Inhalte, aber nicht das vollständige Medium bezieht und c) das Scene Framework, welches Metadaten zum Inhalt einer Szene enthält. Das DMS-1 Schema ist für eine freie Erweiterung ausgelegt und kann damit für spezielle Anwendungen erweitert werden.

Das Material Exchange Format ist, wenn man aktuelle Entwicklungen betrachtet, zum Standardformat in der Fernsehproduktion geworden. Dies wird auch dadurch begünstigt, dass sich in das Format jegliche Metadatenformate integrieren lassen, wie bspw. das in Abschnitt 3.1 beschriebene Format BMF. Allerdings setzt die Nutzung dieser Metadatenformate sowohl Konformität auf Anwendungsebene bzgl. MXF selbst als auch zu dem jeweils eingebetteten Metadatenstandard der am Programmaustausch beteiligten Partner voraus. Ein positiver Fakt ist die frei verfügbare MXF-Bibliothek[10], die eine Programmierschnittstelle zur Nutzung des Standards bietet und damit für eine gute Verbreitung von MXF sorgt.

---

[8] SMPTE 377M - MXF File Format Specification

[9] SMPTE 380M – DMS-1 Standard Set of Descriptive Metadata

[10] http://sourceforge.net/projects/mxflib

2.4 P/Meta

P/Meta ist eine Bibliothek von einheitlich genutzten Begriffen im Umfeld der Fernsehproduktion. Sie enthält Datentypen und –strukturen, die zur Identifikation, zur technischen Beschreibung und dem mit der Produktion verbundenem Rechtemanagement dienen [EBU07]. P/Meta wurde von der EBU[11] und der SMTPE[12] innerhalb einer gemeinsamen Arbeitsgruppe erstellt und ist hauptsächlich zum Programmaustausch während und nach der Produktion vorgesehen. Die aktuelle Version 2 des Schemas wurde 2007 veröffentlicht. Eine wesentliche Besonderheit des P/Meta Formates ist, dass der sprachübergreifende Programmaustausch in Form von Ländercodes unterstützt wird. Dadurch wird eine einfache Übersetzung der Deskriptoren sowie der eigentlichen Werte möglich.

Die P/Meta Bibliothek ist eine semantische Ebene mit standardisierten Begriffen für die Fernsehproduktion. Der wesentliche Vorteil von P/Meta ist die Flexibilität durch die Möglichkeit der Definition von Anwendungssets. Allerdings ist dies ebenso das Hauptmanko in Bezug auf Interoperabilität und damit gleichzeitig auf die Verbreitung und Nutzung als Austauschformat.

2.5 TV Anytime

TV Anytime ist ein Metadatenstandard, der für das Anwendungsszenario Personal Video Recorder (PVR) entwickelt wurde. Das Format wurde durch das 1999 gebildete TV Anytime Forum spezifiziert und 2003 durch die ETSI[13] als Standard[14] veröffentlicht. Im TV Anytime Standard sind vier Hauptkategorien definiert: Content Description Metadata (inhaltsbeschreibende Metadaten), Instance Description Metadata (Metadaten zum Dienst und/oder Programm), Consumer Metadata (Informationen zu Benutzerpräferenzen und –history) und Segmentation Metadata (Informationen zu zeitlichen Einteilungen des audiovisuellen Materials) [Man02]. Neben den klassischen beschreibenden Metadaten für die audiovisuellen Inhalte werden im Standard Metadaten zum Konsumverhalten des Nutzers definiert. Daher wird mit der Nutzung von TV Anytime personalisiertes Fernsehen möglich.

Für die Recherche in einem vernetzten Beitragsarchiv spielt der Standard eine untergeordnete Rolle. Die Möglichkeiten zur Nutzung des Konsumverhaltens ist vor allem aus Sicht der Dienstanbieter von großer Bedeutung. Dies schließt die Fernsehprodu-

---

[11] European Broadcasting Union, http://www.ebu.ch

[12] Society of Motion Picture and Television Engineers, http://www.smpte.org

[13] European Telecommunications Standards Institute, http://www.etsi.org

[14] TS 102 822-1 bis -9, http://www.etsi.org/website/technologies/tvanytime.aspx

zenten ebenfalls mit ein. Daher sollte der Standard im Kontext des InnoProfile Projektes sachsMedia unbedingt Berücksichtigung finden.

3 Metadatenformate in der Fernsehproduktion

Für den standardisierten Austausch von sendefertigen Fernsehbeiträgen in B2B Anwendungen existieren ebenfalls einige Metadatenformate. Nachfolgend werden einige dieser Datei- und Metadatenformate vorgestellt. Neben den schon beschriebenen Metadatenstandards wurden Formate zur Anwendung im B2C-Bereich entwickelt. Dabei geht es um Metadaten, die der Nutzer entweder zur Suche in den audiovisuellen Daten nutzen kann oder die im als Zusatzinformationen präsentiert werden während er die Inhalte konsumiert. Daher wird zusätzlich auf das Format RSS-TV für Internetfernsehen eingegangen.

3.1 Broadcast Metadata Exchange Format (BMF)

Das Format BMF wurde am IRT[15] als standardisiertes Austauschformat für Metadaten in der Fernseh- und Rundfunkproduktion entwickelt [Ebn05]. Um dieses Ziel bestmöglich zu erreichen wurde eine Analyse von Anwendungsfällen in der Fernsehproduktion vorgenommen, welche den Ausgangspunkt für die zu definierenden Deskriptoren und das letztliche Schema bildete. Es wurde ein Klassenmodell erzeugt, welches als Austauschschema für Metadaten zwischen Komponenten der Datei-basierten Produktion fungiert. BMF beschreibt nicht, wie die Speicherung der Informationen in den einzelnen Systemen zu erfolgen hat. Daher lässt sich das Format sowohl in MXF integrieren oder separat als zusätzliche Datei austauschen.

Das vollständige Klassenmodell von BMF ist sehr umfangreich und komplex. Für eine konkrete Anwendung können jedoch die ausschließlich benötigten Teile des Modells implementiert werden. Vorteil des Modells ist die vollständige Abbildung des Produktionsprozesses. Für eine Anwendung zum Austausch von Programmen ist es vor allem dann sinnvoll, wenn Informationen zu Bestandteilen des Produktionsprozesses eine Rolle spielen. Dies ist für den Programmaustausch unter lokalen Fernsehanbieter nur sehr bedingt der Fall.

3.2 PBCore

PBCore wurde von der Public Broadcasting Metadata Initiative (PBMI) entwickelt und liegt als XML-Schema seit Ende 2008 in Version 1.2 vor [Whi03]. Wie der Name bereits vermuten lässt, handelt es sich dabei, um ein in Anlehnung an Dublin Core entwi-

[15] Institut für Rundfunktechnik, http://www.irt.de

ckeltes Schema für die Broadcast-Welt. In Version 1.2 beinhaltet das Schema 61 Elemente, die in 15 Containern organisiert und in vier inhaltliche Klassen (Content Classes) eingeordnet sind. Diese Klassen sind im Einzelnen: Intellectual Content (inhaltsbeschreibende Metadaten), Intellectual Property (Metadaten zum Copyright), Instantiation (technische Metadaten) und Extensions (zusätzliche Metadaten).

Die Kompaktheit des Schemas, die gute Dokumentation in Form von XML-Schemata und die freie Zugänglichkeit sind ideale Faktoren für eine weite Verbreitung des Formates aber auch für eine gute Interoperabilität auf Anwendungsebene. Deshalb sollte das Format von einem recherchierbaren Beitragsarchiv mindestens in Form einer Export-Schnittstelle unterstützt werden.

3.3 RSS-TV

RSS-TV[16] an sich ist kein Metadatenstandard sondern viel mehr eine Erweiterung von RSS für Web-basierte Mediendienste. Die Spezifikation[17] liegt seit März 2008 in der Version RSS-TV 2.3 vor. Ziel der Spezifikation ist es den Austausch von Metadaten für IP-TV Anwendungen zu gewährleisten. Es besteht laut der Spezifikation des Formats der Bezug zum ESG aus dem DVB Standard, der entscheidende Unterschied liegt jedoch im Datenaustausch in beide Richtungen auf Basis des IP-Protokolls.

Die RSS-TV Spezifikation ist sehr kompakt und an Web-basierte Anwendungen angepasst. Dies fördert einerseits die Verbreitung und ebenso die Interoperabilität zumindest im Medium Internet. Im Deutschen IPTV Verband wurde im August 2008 beschlossen RSS-TV als Beschreibungsformat für IPTV-Inhalte einzusetzen und zu etablieren [Str08], damit auf der Ebene der Web-TV Anbieter ein einheitliches Austauschformat verwendet werden kann. Dies schürt die These, dass ein kompaktes Metadatenformat in der Wirtschaft eher angenommen wird, als ein komplexer und flexibel erweiterbarer Standard. Aus diesem Grund sollte auch dieses Format, zumindest in Form einer Schnittstelle in ein vernetztes Beitragsarchiv integriert werden.

4 Vorgaben zur Erstellung intellektueller Metadaten

Neben der Strukturierung der Metadaten zu audiovisuellen Inhalten nach vorgegeben Standards muss auch die inhaltliche Erfassung, insofern sie vollständig intellektuell oder semi-automatisch durchgeführt wird nach vorgegebenen Regeln erfolgen. Nur so wird gewährleistet, dass ein ständig wachsendes Archiv auch einfach zugänglich bleibt. Die Problematik fast unüberschaubarer Bestände mit audiovisuellem Material stellt sich

[16] http://www.rss-tv.org

[17] http://www.rss-tv.org/documentation.html

primär natürlich eher großen Fernsehanstalten, wie den öffentlich-rechtlichen Sendern. Diese haben aufgrund ihrer Größe und der damit verbundenen enormen Reichweite die Möglichkeit ihrer Archive intensiv intellektuell zu pflegen. Um dabei möglichen Problemen zu entgegnen, die bspw. durch unterschiedliches verwendetes Vokabular der Dokumentare oder auch eine nicht-standardisierte Vorgehensweise bei der Archivierung entstehen können, wurde das Regelwerk Mediendokumentation verfasst [DRA08]. Die Ziele des Regelwerks sind:

- eine formale Beschreibung von Fernsehproduktionen und –sendungen
- Vorgabe von Richtlinien für die Inhaltsbeschreibung
- Vorgabe von Richtlinien zur Feststellung der Archivwürdigkeit

Die Problematik eines aus Sicht der Größe unüberschaubar großen Archivs stellt sich den Lokalfernsehsendern (noch) nicht. Dennoch sollten bei der Entwicklung einer Archiv-Anwendung die Grundgedanken des Regelwerks bedacht werden. Denn in der täglichen Arbeit im Lokalsender befassen sich nahezu alle Mitarbeiter mit der Archivierung des Materials, meist die von ihnen selbst erstellten Beiträge. Die Mitarbeiter haben nur in den seltensten Fällen eine dokumentarische Ausbildung, was das Problem der intellektuellen Annotation ohne standardisierte Regeln verschärft.

5 Zusammenfassung und Ausblick

Die Gegenüberstellung der im vorliegenden Artikel aufgeführten Metadatenstandards und –formate lässt zwei wesentliche Schlussfolgerungen zu. Einerseits sind die etablierten Standards wie MPEG-7, DMS-1 (MXF), P/Meta und TV Anytime sehr generisch angelegt, allerdings werden sie vermutlich genau aus diesem Grund mit Ausnahme von TV Anytime nicht einheitlich eingesetzt. Zudem überschneiden sich die Elemente der Standards auch erheblich [Mas08]. In realen Anwendungen wird, wenn überhaupt standardisierte Metadatenschema eingesetzt werden, nur auf Untermengen dieser Formate zurückgegriffen.

Im praktischen Einsatz haben sich in der Vergangenheit kompakte Formate wie Dublin Core durchgesetzt. Dies legt den Schluss nahe, dass für einen vernetzten Datenaustausch in der Fernsehproduktion insbesondere kompakte Formate wie PBCore und RSS-TV berücksichtigt werden sollten. Für die automatische Annotation und die entsprechenden Schnittstellen in einem Multimedia Retrieval System sollten die etablierten Standards MPEG-7 und TV Anytime genutzt werden.

Aus diesen Gründen ist es sinnvoll ein Konverter-Tool zu entwickeln, welches die sich überschneidenden Deskriptoren der hier aufgeführten Standards ineinander überführbar macht. Zusätzlich sollte das Tool möglichst generisch angelegt sein, damit zukünftige Formate und Deskriptoren später leicht integriert werden können.

6 Literaturverzeichnis

[Bai07] Bailer, W., Schallauer, P., & Neuschmied, H. (2007). *MPEG-7 Detailed Audiovisual Profile*. Abgerufen am 11. Mai 2009 von http://iiss039.joanneum.at/cms/fileadmin/mpeg7/files/mpeg7davp0.6.pdf

[DRA08] DRA. (2008). *Regelwerk Mediendokumentation - Fernsehen.* Abgerufen am 11. Mai 2009 von http://rmd.dra.de/arc/doc/REM_RDK_64.pdf

[Ebn05] Ebner, A. (2005). *Austausch von Metadaten - Broadcast Metadata exchange Format, BMF.* Abgerufen am 11. Mai 2009 von http://www.irt.de/IRT/referate/eoe/jb04/09_metadaten.pdf

[EBU07] EBU. (2007). *P_Meta 2.0 - Metadata Library, EBU - Tech Report 3295v2*. Abgerufen am 11. Mai 2009 von http://tech.ebu.ch/docs/tech/tech3295v2.pdf

[Man02] Manjunath, B. S., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Interface.* Wiley & Sons.

[Mar04] Martinez, J. S. (2004). *MPEG-7 Overview*. Abgerufen am 11. Mai 2009 von http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[Mas08] Maschauer, S., & Kosch, H. (2008). *Strukturierter Vergleich aktueller Multimedia Metadatenstandards.* Abgerufen am 11. Mai 2009 von http://www.multimedia-metadata.info/Software%20and%20Tools/comparison-of-metadata-standards

[Str08] Strzebkowski, R., Schulz, V., Kohle, N., Leykum, M., & Schultz, C. (2008). *Workshop "Metadaten-Standard für WebTV / IPTV".* Berlin: Deutscher IPTV Verband.

[Wel06] Wells, N., Morgan, O., Wilkinson, J., & Devun, B. (2006). *Introduction to MXF: Understanding the Material eXchange Format.* Butterwoth Heinemann.

[Whi03] White, A. M., Baker, A., Bloss, M., Burrows, P. E., Efthimiadis, E. N., Brooks, M., et al. (2003). PB core --- the public broadcasting metadata initiative: progress report. *Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice---metadata research & applications.* Seattle: Dublin Core Metadata Initiative.

Entwurf einer Service-orientierten Architektur als Erweiterung einer Plattform zum Programm-Austausch

Jens Kürsten

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

jens.kuersten@informatik.tu-chemnitz.de

Zusammenfassung: Der vorliegende Beitrag beschreibt ein Konzept für eine Webservice Architektur zur Recherche in audiovisuellen Inhalten. Dazu wird auf die Rahmenbedingungen einer bestehenden Plattform und den Workflow der Produktion in Lokalsendern eingegangen. Das Ziel ist die Entwicklung eines standardisierten Archivierungsprozesses und die Erweiterung der bestehenden Plattform zum Programmaustausch zur Wieder- und Weiterverwertung von produziertem Material.

Schlagwörter: Information Retrieval, Archivierung, Webservices

1 Einleitung

Aufgrund der Umbrüche in der Medienlandschaft innerhalb des letzten Jahrzehnts stehen kleine und mittelständische Lokal-Fernsehsender vor neuen Herausforderungen. Neben der wachsenden Konkurrenz um Werbekunden durch lokale Nachrichten-Anbieter wie Lokalzeitungen ist die komplette Umstellung auf digitale Produktion und die damit verbundenen Investitionen in Technik das schwerwiegendste Problem. Dieser Problematik wurde mit dem Projekt „Veranstaltervernetzung“, initiiert durch den Sendernetz e.V.[1], entgegnet. Dabei gab es zwei wesentliche Ziele. Einerseits die Vernetzung der Sender mit dem Zielen der Kostensenkung in der Produktion und der Erhöhung der Reichweite der einzelnen Veranstalter. Andererseits sollten die zahlreich vorhandenen Kopfstellen vernetzt werden, um den logistischen Aufwand der Sender zu minimieren und damit in Zukunft tagesaktuelles Programm zu ermöglichen. Als wesentliches Produkt des Projektes wurde die Programmbörse[2] entwickelt. Dabei handelt es sich um ein System zum Austausch von sendefähigem Material der partizipierenden Veranstalter. Aufbauend auf dieser Grundidee werden im InnoProfile Projekt sachsMedia[3] Methoden zur semantischen Suche in den audiovisuellen Beiträgen konzipiert und prototypisch umgesetzt. Die Verfahren sollen dabei möglichst automatisiert arbeiten,

---

[1] http://www.sendernetz.tv

[2] http://www.programmboerse.tv

[3] Gefördert durch: Unternehmen Region, Die BMBF Innovationsinitiative Neue Länder

um den Aufwand für eine intellektuelle inhaltliche Annotation des Videomaterials zu sparen oder mindestens zu minimieren. Damit können die bestehenden Archive, die unter den bisher gegebenen Bedingungen nur mit erheblichem Aufwand recherchierbar sind, wieder zugänglich gemacht werden. Um die entwickelten Verfahren später im praktischen Einsatz nutzen zu können wird eine flexible Architektur benötigt, die sich problemlos in die bestehende Plattform der Programmbörse integrieren lässt. Ferner soll die entwickelte Architektur ebenso für weitere potenzielle Anwender nutzbar sein.

Nachfolgend wird ein Konzept für eine Service-orientierte Architektur vorgestellt. Basierend auf einer Anforderungsanalyse bestehend aus dem Workflow der Lokalsender und der bereits existierenden Austauschplattform Programmbörse werden die Kernkomponenten des Konzepts beschrieben. Im Anschluss daran wird auf ein prototypisch umgesetztes Recherche-Frontend eingegangen. Ferner werden Erweiterungsmöglichkeiten und Anwendungsszenarien skizziert. Abschließend wird neben einer Zusammenfassung ein Ausblick auf zukünftige Anwendungen des Konzepts gegeben.

2 Konzept Service-orientierte Architektur

Ziele der Entwicklung des nachfolgend vorgestellten Konzepts sind die flexible Nutzung der im InnoProfile Projekt sachsMedia entstehenden Methoden zur automatischen Annotation von audiovisuellen Inhalten und der entwickelten Algorithmen zur Recherche, sowie die prototypische Einbindung praktikabler Verfahren in die Austauschplattform Programmbörse. Um dies zu gewährleisten, werden die bestehende Plattform, sowie der Workflow der Lokalsender analysiert und mit den wissenschaftlichen Frameworks zur inhaltlichen Analyse und dem Retrieval in Bezug gesetzt.

2.1 Produktionsworkflow im Lokalfernsehen

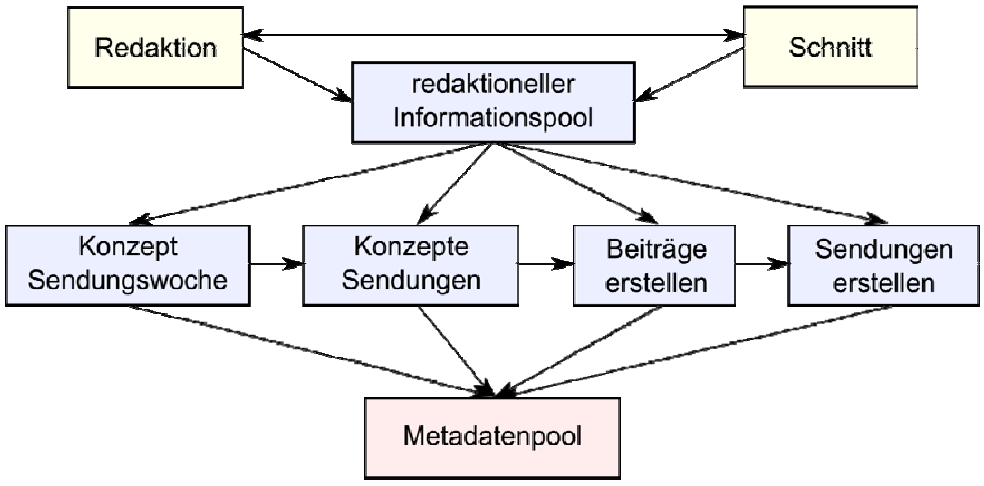


Abbildung 1: Metadaten im Workflow der Produktion im Lokalfernsehen

Im Produktionsprozess eines Beitrags oder einer ganzen Sendung entstehen in kreativer Arbeit strukturierte Daten für die audiovisuellen Inhalte. Der Workflow in diesem Prozess ist in Abbildung 1 dargestellt. Aus einem mehr oder minder strukturiertem Informationspool werden in diesem Prozess Konzepte für Beiträge erstellt. Diese wiederum werden produziert und in einer Sendung oder einer Sendungswoche zusammengestellt.

Anspruch eines zu entwickelnden Archivsystems ist einerseits die Unterstützung der redaktionellen Arbeit, die meist auf unstrukturierten Informationen basiert und andererseits die standardisierte Erfassung und Strukturierung der im Produktionsprozess erzeugten Metadaten.

2.2 Austauschplattform „Programmbörse“

Im Rahmen des Projektes „Veranstaltervernetzung“ wurde durch die HMS OHG[4] die Plattform Programmbörse realisiert. Sie besteht im Wesentlichen aus drei Komponenten: dem Nutzerfrontend zur Recherche und Selektion von Sendungen oder Beiträgen, einer Backend-Anwendung mit Datenbank zur Verwaltung der Metadaten und dem Transfer-Client, welcher zum Up- und Download der audiovisuellen Inhalte dient.

Die Recherche innerhalb des Programms der teilnehmenden Veranstalter erfolgt auf Basis der intellektuell vergebenen Metadaten beim Upload in die Plattform. Technisch erfolgt die Suche im Volltext der Metadaten in der Datenbank. Dies bietet Vorteile in Bezug auf die Antwortzeit in der Rechercheoberfläche, geht aber mit Effektivitätseinbußen im Vergleich zu Verfahren des Information Retrieval einher.

Der Austausch des Programms erfolgt direkt über den Transfer-Client, der auf einem Kommunikations-PC bei jedem teilnehmenden Veranstalter installiert ist. Der tatsächliche Austausch kann innerhalb der Plattform auf zwei Ebenen erfolgen. Zum einen als klassische Börse, bei der ein Veranstalter die Rechte zur Nutzung des Materials eines Anderen käuflich erwirbt und zum anderen als Programmverteiler, bei der die gewünschten Inhalte kostenfrei (abgesehen vom entstehenden Internet- bzw. Netzwerk-Traffic) ausgetauscht werden.

Für eine möglichst transparente und rückwärts-kompatible Erweiterung der Programmbörse ist daher eine Architektur auf Basis von Webservices ideal, da keine grundlegenden Änderungen am bestehenden System nötig sind und schrittweise Neuerungen integriert werden können.

---

[4] http://www.hms-dev.de

2.3 Funktionale und technische Anforderungen

Zur Umsetzung des entworfenen Konzepts werden nachfolgend grundlegende Voraussetzungen beschrieben. Auf funktionaler Ebene sind dies:

Inhaltsbasierte Recherche im redaktionellen Produktionsprozess: Für die redaktionelle Arbeit eines Fernsehproduzenten soll es möglich sein sowohl auf Basis vorhandener Metadaten als auch mit Hilfe automatisch erstellter inhaltsbasierter Metadaten im bestehenden Archiv nach existierenden Beiträgen zu suchen. Die Datenbasis selbst kann dabei das eigene Archiv des Senders, aber auch das vernetzte Archiv mehrerer Sender bilden.

Umfassende Erfassung von Metadaten im redaktionellen Prozess: Alle intellektuell erzeugten Metadaten müssen während des Produktionsprozesses erfasst, strukturiert und zusammen mit dem fertigen Beitrag abgelegt werden. Nur so kann eine weitere Verwertbarkeit auch jenseits einer erneuten Nutzung im redaktionellen Prozess gewährleistet werden. Dazu muss der Workflow der Produktion analysiert und darauf aufbauend ein entsprechendes System zur Erfassung der Metadaten mit Speisung ins Recherchesystem für das Archiv entworfen werden.

Konverter Modul für Metadatenformate: Eine Analyse bestehender Metadatenformate [Kür09] zeigt, dass für jegliche Arten der Weiterverwertung von produzierten Fernsehbeiträgen Metadaten in für das Verwertungsmedium angepassten Standards und Formaten vorliegen sollten. Daher ist ein Tool zur Umwandlung in verschiedene Zielformate unumgänglich.

Korrektur und Unterstützung einer automatischen Klassifikation: Die automatische Klassifikation von Inhalten basiert auf einem lernenden System, welches insbesondere in einer Trainingsphase intellektuelle Unterstützung benötigen wird. Hierzu ist es erforderlich eine Schnittstelle zu schaffen, die es einem Nutzer ermöglicht das System zu unterstützen und entsprechendes Feedback für Inhalte zu geben, die vom Klassifikationssystem als schwer kategorisierbar identifiziert wurden.

Technisch erfolgt die Umsetzung des Konzepts als Webservice. Hierfür hat sich die OpenSource-Entwicklung Apache Tomcat[5] als Quasi-Standard durchgesetzt und wird daher für die Realisierung der Dienste verwendet. Der eigentliche Datenaustausch über die einzelnen zu implementierenden Services erfolgt dabei mittels SOAP (Simple Object Access Protocol). Zur Erzeugung der SOAP-basierten Webservices wird das Apache Axis[6] Framework verwendet, da es eine unkomplizierte Generierung von Schnittstellen zu entsprechendem Java-Code ermöglicht. Die Realisierung eines Frontends

[5] http://tomcat.apache,org

[6] http://ws.apache.org/axis2

kann in einer Skriptsprache, wie PHP, Perl oder Python, zur Erzeugung von dynamischen Webinhalten erfolgen. Eine modulare, prototypische Umsetzung in PHP ist in Abschnitt 3.3 beschrieben.

Die beschriebenen technischen Eckdaten der prototypischen Umsetzung des Konzepts ermöglichen es die entworfene Architektur sowohl lokal in einem oder mehreren Lokalsendern als dezentrale Lösung, als auch zentral in einem Szenario zur Erweiterung der Programmbörse zu betreiben. Zusätzlich wird durch die Verwendung von etablierten OpenSource-Komponenten eine flexible Erweiterbarkeit der Architektur gewährleistet.

3 Komponenten der Architektur

In Abbildung 2 ist ein vereinfachtes Schema der Service-orientierten Architektur dargestellt. Die beiden zentralen Komponenten der Architektur, das Xtrieval Framework und das Annotation Framework sind Entwicklungen, die an der Professur Medieninformatik realisiert worden sind und in den folgenden Abschnitten kurz beschrieben werden. Die Komponente Media Storage wird zur Verwaltung der audiovisuellen Daten verwendet.

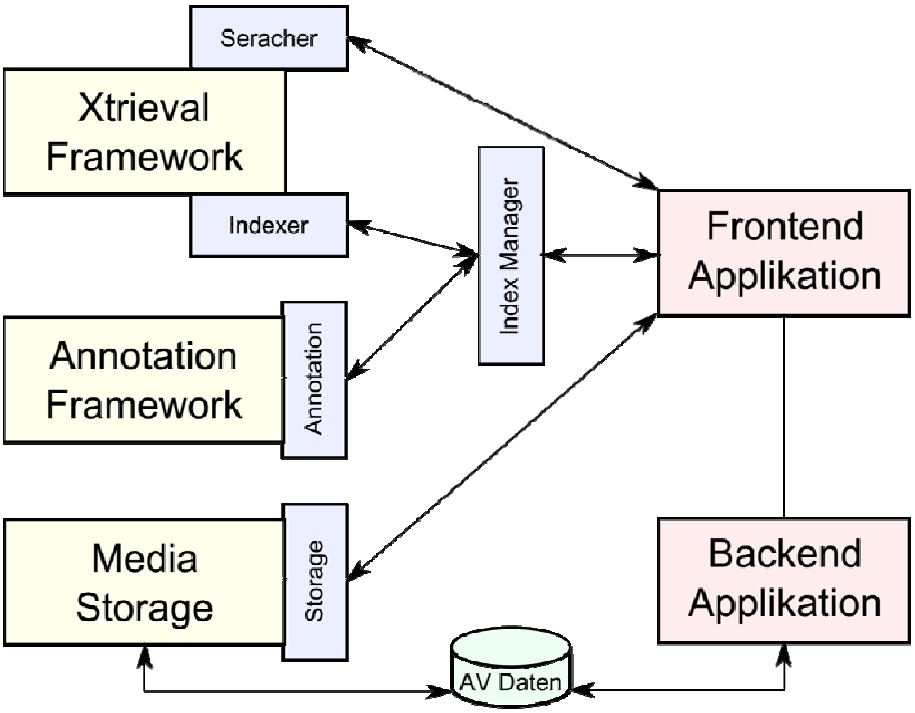


Abbildung 2: Schematische Übersicht Webservice-Konzept

Die blau dargestellten Elemente sind eigenständige Webservices, die jeweils die Kernfunktionalität des dahinter stehenden Frameworks zur Verfügung stellen. Auf der rechten Seite ist beispielhaft eine externe Applikation dargestellt, die die verfügbaren

Dienste nutzen kann. Das dargestellte Schema ist an die bestehende Plattform Programmbörse (siehe Abschnitt 2.1) angepasst.

Um der Anforderung gerecht zu werden, die im InnoProfile Projekt sachsMedia entwickelten Methoden prototypisch in die Programmbörse zu integrieren, ist zwischen dem Backend der Programmbörse und dem Rechencluster der Professur Medieninformatik ein Datentransfer realisiert worden. Dieser ermöglicht zum einen die Verarbeitung der Inhalte der Börse mit dem Annotation Framework. Andererseits wird an dieser Stelle die Verknüpfung der Inhalte mit der Identifikation im Index für die Recherche vorgenommen. Dies ist notwendig, damit das Frontend der Programmbörse weiterhin mit den Identifikatoren der audiovisuellen Inhalte des Backends arbeiten kann.

3.1 Annotation Framework

Zur Realisierung der prototypischen Methoden zur automatischen Annotation der audiovisuellen Inhalte wurde ein Framework entwickelt. Es vereint demzufolge die implementierten Methoden zur Inhaltsanalyse, die unabhängig oder in Kombination von Bild- und Toninhalten erfolgen kann. Eine detaillierte Beschreibung dieser Komponente ist in [Rit09] gegeben.

3.2 Recherche-Framework Xtrieval

Die Komponente Xtrieval Framework stellt die Funktionalität für die Suche zur Verfügung. Ursprünglich für die Evaluation von Methoden des Text-basierten Information Retrieval entwickelt, wird das Framework für die hier beschriebene Architektur erstmals im praktischen Einsatz verwendet. Eine vollständige Beschreibung des Xtrieval Frameworks ist in [Wil08] gegeben. Im Rahmen des internationalen Evaluationsforum CLEF[7] konnte mehrfach mit hervorragenden Ergebnissen die Qualität der implementierten Verfahren nachgewiesen werden [Kür08a], [Wil06]. Neben dem klassischen Text Retrieval wurde das Framework inzwischen ebenso zum Question Answering, zur Klassifikation von Videos und für inhaltsbasiertes Bild Retrieval eingesetzt. Auch hier sind die Ergebnisse [Kür08b], [Kür08c], [Wil06] im internationalen Vergleich beachtlich. Das Framework selbst unterliegt einer ständigen Weiterentwicklung. So sind in den letzten Jahren neben der ursprünglich eingesetzten Retrieval API Lucene[8], die vornehmlich in der Praxis für Recherche-Anwendungen eingesetzt wird, zusätzlich Schnittstellen zu den renommierten wissenschaftlichen APIs Lemur Toolkit[9] und Terrier [Oun07] geschaffen worden. Eine umfangreiche Evaluation dieser beiden Schnittstellen steht noch aus.

---

[7] http://www.clef-campaign.org

[8] http://lucene.apache.org

[9] http://www.lemurproject.org

Für den Einsatz des Frameworks im hier beschriebenen Konzept werden vor allem Verfahren zur Klassifikation und zur Recherche auf inhaltlicher Ebene entwickelt und getestet. Daher wird die Webservice-Schnittstelle zur Recherche in audiovisuellen Inhalten im Entwicklungsprozess ebenfalls einer fortlaufenden Weiterentwicklung unterliegen. Um dennoch einen praktischen Betrieb der Schnittstelle zu ermöglichen, müssen alle Erweiterungen rückwärts-kompatibel gestaltet werden.

3.3 Prototypische Frontend-Applikation

Nachfolgend wird kurz auf eine prototypische Umsetzung einer Frontend-Applikation für das erstellte Webservice-Konzept zur Recherche eingegangen [Müc08]. Das Frontend ist eine logisch von den Webservices getrennte PHP-Anwendung. Nach dem Model-View-Controller-Architekturmuster entworfen, bietet es Flexibilität in Hinsicht auf Datenquelle (Searcher bzw. Indexer) sowie Darstellung (HTML, XML o. ä.).

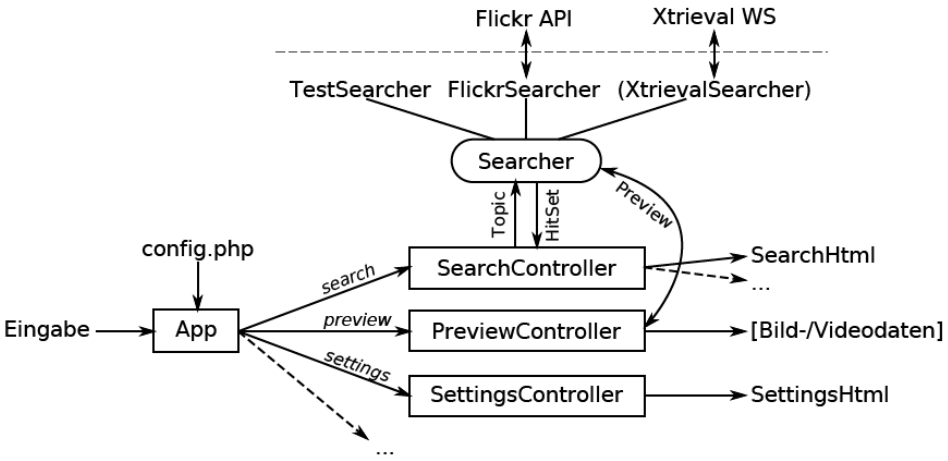


Abbildung 3: Datenfluss des Frontend Prototyps

Den Kern der Anwendung bildet die Klasse App, welche die Verwaltung des internen Zustands sowie den Datenfluss kontrolliert. Anhand des Aktionsparameters übergibt sie die Kontrolle dem jeweils zuständigen Controller, welcher wiederum mit dem entsprechenden Modell interagiert und die Daten einem View zur Darstellung zur Verfügung stellt (siehe Abbildung 3).

Im Falle des SearchController wird das Modell durch eine vom flexiblen Searcher-Interface abgeleiteten Klasse repräsentiert, welche für die Generierung von Suchergebnissen zuständig ist. In einem ersten Prototyp wurde eine exemplarische Implementie-

rung eines Searchers für die API des Webdienstes Flickr[10] genutzt. Zusätzlich wurde auch eine Schnittstelle zum Xtrieval Framework implementiert.

Die Formulierung einer Suchanfrage an ein Searcher-Objekt erfolgt in Form des Datentyps Topic, welches in Anlehnung an das Xtrieval Framework eine Liste von Datenfeldern ist. Ergebnisse werden als HitSet zurückgegeben, einer Liste von Hit-Objekten, von denen jedes jeweils ein gefundenes Dokument repräsentiert. Ebenso wird vom Searcher eine Anfrage einer Vorschau (Preview) zu einem spezifischen Dokument unterstützt. Diese wird durch den PreviewController, einem speziell für Ausgabe von Dokumentvorschauen konzipierten Controller, in mit einem Hash versehenen Dateien lokal auf dem Server zwischengespeichert.

3.4 Erweiterungsmöglichkeiten

Das in diesem Abschnitt beschriebene Modell stellt nur eines von mehreren denkbaren Anwendungsszenarien dar. Diese erfordern geringfügige Änderungen und Erweiterungen, sind jedoch nicht nur denkbar sondern als weitere Entwicklungsstufen des Konzepts bereits in Planung. Dazu gehören:

- Entwicklung eines Services zur Generierung von inhaltsbasierten Metadaten für audiovisuelle Medien
- Anbindung innovativer, multimedialer Rechercheoberflächen mit vorheriger Nutzerevaluation
- Umsetzung eines Metadaten Konverters als Service zur besseren Weiterverwertung von Programminhalten

4 Zusammenfassung und Ausblick

Im vorliegenden Artikel wurde ein Webservice-Konzept zur Recherche in audiovisuellen Inhalten vorgestellt. Das Modell kann als Erweiterung der Programmbörse oder auch als eigenständige Archivanwendung implementiert werden. Der Fokus für die Weiterentwicklung des Konzepts liegt in den folgenden drei Bereichen: a) Implementierung von Webservices für praktikable automatische Annotations- und Klassifikationsverfahren, b) wissenschaftliche Untersuchung innovativer Recherche-Oberflächen für audiovisuelle Inhalte und c) Entwicklung eines generischen Tools zur Konvertierung von Metadaten in verschiedene etablierte Formate für weitere Distributionskanäle.

[10] http://www.flickr.com

5 Literaturverzeichnis

[Kür09] Kürsten, J. (2009). *Metadatenstandards und -formate für audiovisuelle Inhalte.* Chemnitz: dieser Band.

[Kür08c] Kürsten, J., Kundisch, H., & Eibl, M. (2008). QA Extension for Xtrieval: Contribution to the QAst track. *Working Notes for the CLEF 2008 Workshop, 17-19 September.* Arhus.

[Kür08b] Kürsten, J., Richter, D., & Eibl, M. (2008). VideoCLEF 2008: ASR Classification based on Wikipedia Categories. *Working Notes for the CLEF 2008 Workshop, 17-19 September.* Aarhus.

[Kür08a] Kürsten, J., Wilhelm, T., & Eibl, M. (2008). Extensible Retrieval and Evaluation Framework: Xtrieval. *LWA: Lernen - Wissen - Adaption,* (S. 107-110). Würzburg.

[Müc08] Mücklisch, S. (2008). Dokumentation - Xtrieval Webservice Frontend. TU Chemnitz, Fakultät für Informatik, Projektarbeit.

[Oun07] Ounis, I., Lioma, C., Macdonald, C., & Vassilis, P. (2007). Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. *Novatica/UPGRADE Special Issue on Next Generation Web Search* (S. 49-56). CEPIS.

[Rit09] Ritter, M. (2009). *Visualisierung von Prozessketten zur automatischen Shot Detection.* Chemnitz: dieser Band.

[Wil08] Wilhelm, T. (2008). Entwurf und Implementierung eines Frameworks zur Analyse und Evaluation von Verfahren im Information Retrieval. TU Chemnitz, Fakultät für Informatik, Diplomarbeit.

[Wil06] Wilhelm, T., & Eibl, M. (2006). ImageCLEF 2006 Experiments at the Chemnitz Technical University. *LNCS vol. 4730* (S. 739-743). Alicante: Springer Verlag.

Untersuchungen zu semantischem Retrieval von Bildern mit Hilfe von MPEG7 anhand einer Beispielapplikation

Daniel Pötzinger

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

poetzinger@googlemail.com

Zusammenfassung: In der heutigen Welt spielen multimediale Inhalte eine immer größere Rolle und die Anzahl verfügbarer Inhalte steigt immens. Zur sinnvollen Erschließung dieser Daten erhöht sich der Bedarf nach semantischem Retrieval. MPEG7 als Beschreibungsformat Multimedialer Daten bietet mit seinen Beschreibungselementen eine umfassende Basis für semantisches Retrieval. Anhand einer Beispielapplikation werden die grundlegenden Abläufe und Funktionsweisen einer semantischen Suche nach Bildern demonstriert.

Schlagwörter: MPEG7, semantische Suche, Bildretrieval

1 Einleitung

In der heutigen Welt spielen multimediale Inhalte eine immer größere Rolle: moderne Kommunikations- und Produktionstechnologien haben dazu beigetragen, dass wir eine immer größer werdende Menge von Inhalten haben. Diese wachsenden Datenmengen sinnvoll speichern und erschließen zu können, stellt eine wichtige Herausforderung dar.

Ist man auf der Suche nach bestimmten Inhalten ist zurzeit immer noch text- bzw. schlagwortbasierte Suche die häufigste Wahl. Um jedoch auch in Zukunft die Fülle der Daten möglichst optimal nutzen und erschließen zu können, ist es wichtig auch semantisch höher angesiedelte Suchanfragen stellen zu können.

MPEG7 bietet dafür die benötigten Beschreibungselemente. Im Rahmen der Arbeit wird anhand einer Bildersuche gezeigt, wie man diese nutzen kann und welche Wege zur Performanzoptimierung einer Suchanfrage bestehen.

2 MPEG7

Formal wird der Standard auch „Multimedia Content Description Interface“ (ISO/IEC 15938) genannt und er behandelt, im Unterschied zu den vorangegangenen Standards, nicht die Komprimierung oder technische Auslieferung multimedialer Daten. Vielmehr standardisiert MPEG-7 die Beschreibung verschiedener Typen multimedialer Inhalte unabhängig von deren Repräsentation oder Speicherung.

Eine MPEG-7 Beschreibung stellt also eine Repräsentation der Informationen eines multimedialen Inhaltes dar. Die Beschreibung reicht dabei von low-level Beschreibungen, wie beispielsweise die Angabe der Farbverteilung, bis hin zu high-level Beschreibungen, wie beispielsweise die Angaben von Ort, Zeit und Personen.

2.1 Visuelle Descriptoren

Ein Hauptziel der visuellen Deskriptoren in MPEG-7 ist die standardisierte Beschreibung von visuellen (low-level) Eigenschaften. Diese Low-Level Beschreibungen können genutzt werden, um Bilder oder Videos nur auf der Basis nicht textueller Beschreibungen des Inhalts zu vergleichen zu filtern und zu durchblättern.

Im Anschluss sind einige der in der Beispielapplikation verwendeten Descriptoren kurz beschrieben.

2.1.1 Dominat Color Descriptor

DominantColor D erlaubt die Angabe einer kleinen Anzahl dominanter Farbwerte mit ihren statistischen Häufigkeiten, Verteilungen und Varianz. Er bietet damit eine kompakte und intuitive Repräsentation von Farben in einer Region oder einem Bild.

Der D nutzt die Descriptoren ColorSpace und ColorQuantization um den verwendeten Farbraum zu beschreiben.

Der DCD besteht im wesentlichen aus einer Menge dominanter Farben im Bild zusammen mit der relativen Häufigkeit der Farben, uns stellt damit eine recht inuitive Farbbeschreibung eines Bildes dar.

In der XML Repräsentation sieht ein DCD wie folgt aus:

```
<Image>
 <VisualDescriptor xsi:type="DominantColorType">
  <SpatialCoherency>0</SpatialCoherency>
  <Value>
   <Percentage>1</Percentage>
   <Index>21 2 14</Index>
  </Value>
 ....
</VisualDescriptor>
....
```

2.1.2 Color Layout Descriptor

ColorLayout D beschreibt das Layout der repräsentierenden Farbe auf einem, über eine Region oder Bild gelegten, Netz.

Die Repräsentation basiert auf Koeffizienten der DCT einer ZickZack gescannten YCbCr Repräsentation des Bildes. Es ist ein sehr kompakter und für Such- und Filteraufgaben sehr effizienter D.

```
<VisualDescriptor xsi:type="ColorLayoutType">
<YDCCoeff>15</YDCCoeff>
<CbDCCoeff>21</CbDCCoeff>
<CrDCCoeff>43</CrDCCoeff>
<YACCoeff5>28 16 16 16 19</YACCoeff5>
<CbACCoeff2>0 16</CbACCoeff2>
<CrACCoeff2>1 16</CrACCoeff2>
</VisualDescriptor>
```

2.1.3 Edge Histogram Descriptor

Der EHD repräsentiert die räumliche Verteilung von Ecken in einem Bild und stellt damit eine Beschreibung zur Verfügung, welche insbesondere bei unregelmäßigen und heterogenen Texturen effizient ist.

Für den EHD wird das Bild in 4x4 Segmente unterteilt, und für jedes so entstandene Subimage werden 5 Werte für die unterschiedlichen Ecktypen gespeichert. So dass man auf 80 zu speichernde Werte kommt (4 x 4 x 5)

Demnach sieht die XML Representation des EHD aus:

```
<VisualDescriptor xsi:type="EdgeHistogramType">
<BinCounts>1 2 5 6 ... 7 2 3 4 5 7</BinCounts>
</VisualDescriptor>
```

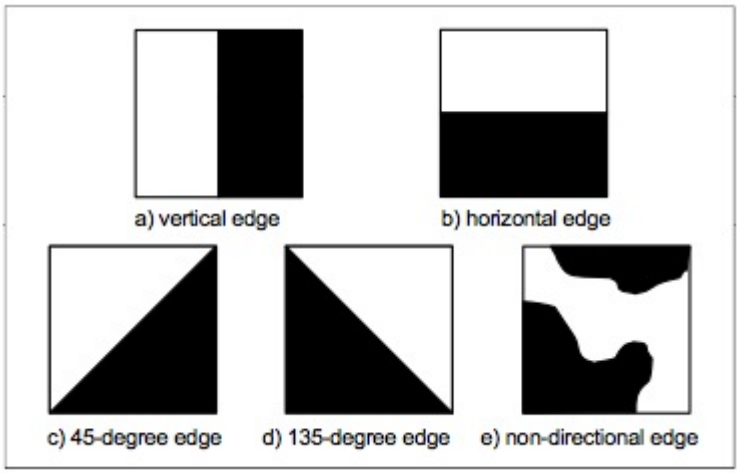


Abbildung 1: Die 5 verschiedenen Ecktypen des EHD (MPEG-7)

Vorstellung der Beispielapplikation

Die entwickelte Beispielapplikation demonstriert, wie man, auf Basis von visuellen Eigenschaften wie Farbe oder Farbverteilung, nach Bildern suchen kann. Weiterhin erlaubt die Applikation theoretisch die Verknüpfung von beliebigen Suchen, durch einen generellen Ansatz Suchanfragen auszuwerten.

Die Implementation der Applikation erfolgte als FLOW3 Paket, FLOW3 ist ein neues PHP Framework mit modernen Konzepten wie Domain Driven Design und Dependency Injection.

2.2 Vorstellung des Domain Models der Suche

Die Applikation wurde nach Domain Driven Design entworfen, im Kern der Anwendung gibt es Klassen die den Problembereich der semantischen Suche abbilden. Der Entwurf des Suchpacketes erfolgte beispielsweise nach folgender allgemeiner Sicht:

Eine Suche besteht aus einer Suchanfrage (*Query*), welche nach verschiedenen Strategien an verschiedene Suchmaschinen (*Search Engine*) weitergereicht wird. Jede Suchmaschine gibt eine Menge (*Result Set*) von Treffern (*Hit*) zurück, welche nach einer bestimmten Strategie kombiniert werden. Der suchende Client bekommt letztendlich wieder eine Menge von Treffern zurück.

Eine Query selbst kann aus mehreren verschiedenen Filtern bestehen. Die Suchmaschine ist entsprechend verantwortlich, die Filter entsprechend auszuwerten.

Für die Beispielapplikation kommt ein Index zum Einsatz, der die relevanten Eigenschaften eines Multimedia-Dokumentes in einer relationalen Datenbank speichert und sich SQL bedient. Dieser Index soll nun genauer vorgestellt werden um im nächsten Kapitel Performanzoptimierungen am Beispiel dieses SQL Indexes zu betrachten.

2.3 Das Userinterface

Semantische Suchanfragen intuitiv zu formulieren, ist schwierig. Dies wird am Beispiel einer allgemeinen *Query by Example*-Anfrage deutlich. Wenn ein Nutzer nach ähnlichen Bildern sucht, können für Ihn recht unterschiedliche Dinge relevant sein:

1. Lediglich die Farben im Bild
2. Die Farbverteilung kann wichtig oder unwichtig sein
3. Die Farbe kann uninteressant sein und stattdessen der Focus auf dargestellten Objekten und Motiven liegen
4. Nur ein Ausschnitt des Bildes ist relevant

Das hier gezeigt Beispiel erlaubt es Suchanfragen zu stellen, in dem man zu einer optionalen Textsuche verschiedene semantische Filter hinzufügen kann. Ein solcher Filter kann beispielsweise „dominate Farbe“ oder „Ort des Fotos“ sein.

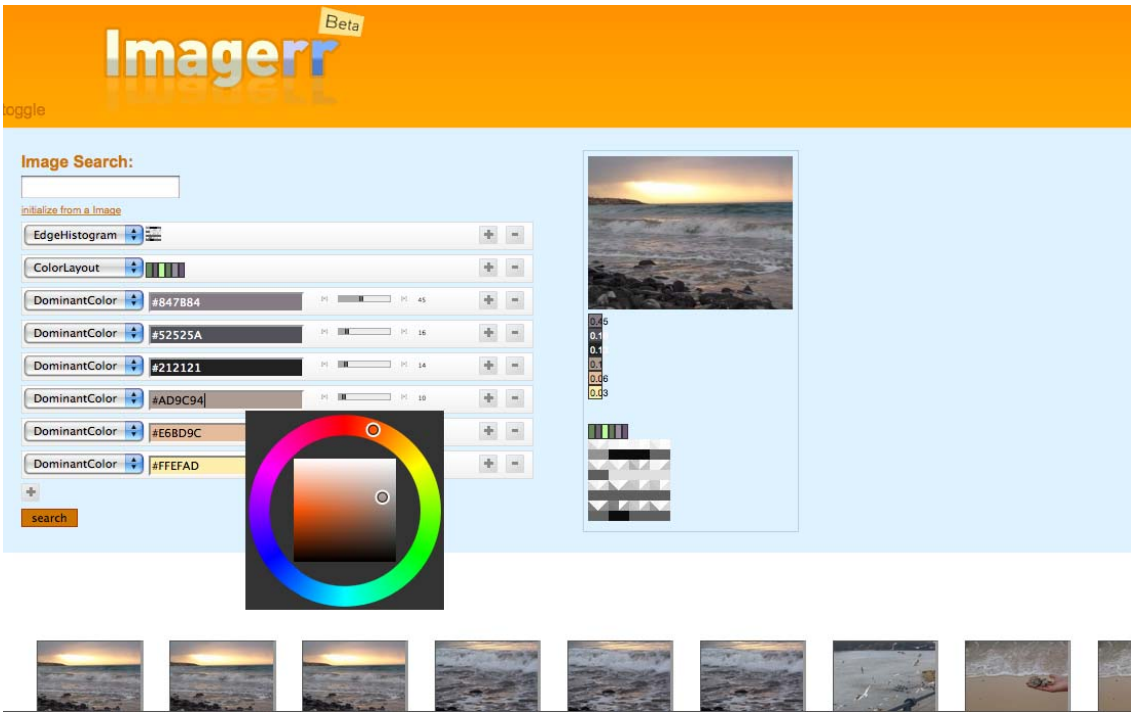


Abbildung 2: Screenshot einer Beispielsuche anhand eines gegebenen Bildes

Die GUI stellt dem User verschiedene Filter zur Verfügung, welche er beliebig hinzufügen oder löschen kann. So gibt es beispielsweise einen Suchfilter „DominantColor“ bei dem der User eine Farbe auswählen kann sowie die erwartete Gewichtung im Bild.

Es ist außerdem möglich, die Suchfilter anhand eines gegebenen Bildes zu initialisieren. Dazu kann der User ein Bild angeben, welches im Hintergrund an den Server gesendet und ausgewertet wird – an Hand der Eigenschaften des Bildes werden die Suchfilter initialisiert – der Nutzer kann vor dem absenden der Suchanfrage diese Filter verändern.

2.4 Performanceoptimierung

Das Grundproblem einer Abfrage wird recht schnell deutlich: Um die Dokumente mit dem geringsten Abstand zu ermitteln, werden die Abstände zu jedem indexierten Dokument in der Indextabelle berechnet. Hier liegt der prinzipielle Ansatz einer Optimierung: Die Menge von Dokumenten, für die das Abstandsmaß berechnet werden muss, sollte möglichst günstig und optimal eingeschränkt werden. Dabei ist das Ziel bei dieser Eingrenzung möglichst keine potentiellen Hits auszuschließen, andererseits aber

die Menge soweit einzuschränken, dass möglichst wenige Dokumente für die eigentliche Auswertung übrig bleiben.

Eine allgemeine Herangehensweise sei an dem Beispiel einer einfachen Umkreissuche erklärt. Eine nicht optimierte SQL Anfrage für eine Suche nach Orten in der Umgebung einer gegebenen Koordinate könnte in etwa so aussehen:

```
select *, SQRT(POW(x-2000,2)+POW(y-2500,2)) as distance
from test order by distance limit 0,100
```

Die Menge der nötigen Abstandsberechnungen kann man durch eine einfache Umquadratsuche bereits deutlich optimieren:

```
select *, SQRT(POW(x-2000,2)+POW(y-2500,2)) as distance
from test where x<3000 AND x>1000 AND y < 3500 AND y >
1500 order by distance limit 0,100
```

Das finden der optimalen Eingrenzung, lässt sich in einem getrenntem Schritt mit geringem Aufwand feststellen.

Unter der Annahme, dass man die Daten vorab bereits nach Umquadraten geclustert und indexiert hat, kann man die SQL Abfrage weiter optimieren:

```
select *, SQRT(POW(x-2000,2)+POW(y-2500,2)) as distance
from test where cluster=100 order by distance limit 0,100
```

Überträgt man diese Herangehensweise auf die Bildsuche so ist das Ziel die Ergebnismenge anhand der Filter in der Anfrage bereits vor der Berechnung des Abstandsmaßes sinnvoll einzugrenzen und somit zu einer besseren Skalierung der Suche zur Anzahl der indexierten Dokumente zu kommen.

3 Literaturverzeichnis

[MPEG7] B.S. Manjunath; Philippe Salembier; Thomas Sikora, Introduction to MPEG7, John Wiley \& Sons LTD, 2003.

[FLOW3] Robert Lemke, http://flow3.typo3.org/about/principles/

Dynamische Distribution personalisierten Mobilfernsehens in hybriden Netzen

Albrecht Kurze, Robert Knauf und Arne Berger

Technische Universität Chemnitz
Fakultät für Informatik
Professur Medieninformatik

{albrecht.kurze,robert.knauf,arne.berger}@informatik.tu-chemnitz.de

Zusammenfassung: Der Trend zum personalisierten Medienangebot und zum zeitunabhängigen Konsum individuell interessierender Dienste nimmt zusammen mit den Wünschen nach Mobilität und Ubiquität stetig zu. Ein nachhaltiger Netzausbau im Bereich der Punkt-zu-Punkt-Kommunikation unterstützt dies. Jedoch sind Gleichzeitigkeitseffekte und somit Kapazitätsengpässe nach wie vor an der Tagesordnung. Rundfunknetze bieten in diesen Situationen die Möglichkeit zur Netzentlastung und zusätzlich zur Verbreitung lokalitätsabhängiger Inhalte. Dieser Beitrag erstellt ein mögliches Anwendungsszenario, welches die Distribution personalisierter Inhalte durch ein koordiniertes Zusammenspiel in hybriden Netzen realisiert. Die Betrachtung erfolgt erstens senderseitig, wobei Dienste und Inhalte semantisch beschrieben und katalogisiert zum Bezug bereitgestellt werden, zweitens empfängerseitig, wobei Inhalte adäquat gefiltert und relevanzsortiert zum Abruf angeboten werden, und drittens auf Kommunikationsebene, wobei eine effiziente Nutzung des Broadcastkanals als Zusatz zur Punkt-zu-Punkt-Verbindung organisatorisch ermöglicht werden soll.

Schlagwörter: Personalisierung, Metadaten, Mobil-TV, Rundfunk, DVB, UMTS, LTE, ESG, Hybrid-Netz, Streaming

1 Individualität, Mobilität, Ubiquität

Die Evolution terrestrischer Mobil-Kommunikationsnetze steht mitten im Umbruch. Mobile IP-Datennetze finden zunehmend Anwendung im Endkundenbereich [TNS09]. Der Anspruch, Zugang zu multimedialen Inhalten zu erhalten, weitet sich von den heimischen Arbeitsplätzen in die Ortsunabhängigkeit aus [JvTD05]. Internet und Medienkonsum werden ubiquitär, also allgegenwärtig. Paketübertragungsstandards wie EDGE, UMTS oder HSPA stehen flächig zur Verfügung [For09] und bieten nominell genügend Bandbreite, um auf Abruf Inhalte flüssig, in ausreichender Qualität auf Displays mobiler Endgeräte darzustellen. Die Erfahrung uninteressante Inhalte ausblenden zu können, zusam-

men mit der gefühlten Verknappung zur Verfügung stehender Zeit und geringen Aufmerksamkeitsspannen, führen zu einem steigenden Bedarf an aktiv zu rezipierenden Medien. Herkömmliches – linear ausgestrahltes – Fernsehprogramm verliert an Attraktivität für zunehmende Zuschauerschichten [Pow09].

Der Faktor der persönlichen Zeitverknappung fällt jedoch ebenfalls bei der Suche nach gewünschtem Inhalt ins Gewicht [JvTD05]. Das Internetangebot ist nur schwer zu überschauen – nicht zuletzt durch große Mengen nutzergenerierter Inhalte. Werkzeuge oder Portale, welche Inhalte vorgefiltert, kategorisiert oder katalogisiert zum Abruf anbieten, sind in diesem Fall gefragt. Die Gewährleistung der (Vor-)Selektion, ob sie nun anbieter- oder nutzerseitig stattfindet, obliegt allein dem Content Provider. Dieser trägt Sorge dafür, dass eine ausreichende Metadatenbasis vorliegt. Neben der Semantik spielen hierbei Informationen bspw. zu Ort, Zeit oder Dringlichkeit eine Rolle, welche der Auswahl interessierender Inhalte dienlich sind. Die Repräsentation der Metadaten hat in entsprechenden Ontologien zu erfolgen, anhand derer sich der Nutzer oder dessen automatischer Filterassistent mit dem Medienangebot abgleichen kann. An solche strukturierten Daten können Lernverfahren anknüpfen, welche Zusammenhänge aus empfangenen Metadaten und anschließender Nutzerinteraktion protokollieren und daraus ein Nutzerinteressenprofil abzuleiten versuchen. Mit einer solchen Datenbasis ist es denkbar, neue abrufbare Inhalte speziell auf den einzelnen Konsumenten abgestimmt hinsichtlich vermuteter Relevanz einzuschätzen. Ein entsprechend fähiges Endgerät wird zum „Inhaltsberater“ und stellt im Idealfall den aktuell interessantesten Beitrag direkt auf Knopfdruck zur Verfügung. Einem solchen Systementwurf widmet sich Abschnitt 2 dieser Arbeit.

Nach der Klärung der Frage, was der Rezipient zu einem bestimmten Zeitpunkt schauen möchte, gilt es in Abschnitt 3 den Verbreitungsweg zu diskutieren, über den personalisierte Inhalte transportiert werden. Prinzipiell eignen sich Punkt-zu-Punkt-Verbindungen bestens, individuelle Anfragen zu bedienen. Trotz zur Verfügung stehender Bandbreiten sind diese zum aktuellen Stand jedoch nicht ausreichend, um gleichzeitig eine große Zahl von Einzelverbindungen in einer Funkzelle mit den erwarteten Qualitätsmerkmalen ausstatten zu können. Kommende Übertragungsstandards wie LTE versprechen hier Abhilfe [rGPP06]. Jedoch werden auch hier entsprechende Engpässe zu erwarten sein, wenn Qualitätsanforderungen und Nutzungszahlen steigen oder Versorgungslücken auftreten. Neben Kurzformaten ist im mobilen Fernsehsektor auch ein Bedarf an längeren Beiträgen zu erwarten, so dass sich ein Mobil-TV-Betrieb nur schwerlich ausschließlich unter Punkt-zu-Punkt-Verbindungsbedingungen vorstellen lässt [Pow09].

An dieser Stelle stehen mobilen Paketdatendiensten etablierte Broadcastnetze zur Seite, die in der Lage sind, aus Gleichzeitigkeitseffekten Nutzen zu ziehen und Bandbreiten-

engpässe effizient zu vermeiden. Digitaler Fernsehempfang hat sich in mobilen Szenarien durch DVB-T zu einem etablierten Medienkanal entwickelt [dL08]. Darauf aufbauend soll der DVB-H-Standard die Anknüpfung an IP-Strukturen bieten sowie Charakteristiken mobiler Endgeräte Rechnung tragen (geringer Energiebedarf durch Time Slicing, effiziente Vorwärtsfehlerkorrektur zur Empfangssicherung im mobilen bzw. urbanen Umfeld, geeignete Bild- und Tonformate) [ETS04]. Marktpolitisch betrachtet hat der DVB-H-Standard seinen Einstieg in Deutschland bislang leider nicht geschafft. Anders jedoch in weiteren Staaten. Der italienische Mobilfunkbetreiber 3Italia bspw. bietet mobiles Fernsehen auf DVB-H-Basis ohne Extrakosten mit großem Erfolg an und in Korea wurden bereits 2006 vom durchschnittlichen Nutzer täglich etwa 60 Minuten mobiles Fernsehen konsumiert. In DVB-H eingebettete Standards wie IP Datacast bieten Raum für Protokolle und Signalisierungsmechanismen, welche zur kooperativen effizienten Mediendistribution mit Paketdatendiensten genutzt werden können. Für die Endnutzer soll es keinen merkbaren Unterschied geben, ob Medien per Uni- oder Broadcastverbindung empfangen werden. Termini und Auswahlmöglichkeiten der Übertragungsverfahren bleiben im Hintergrund bzw. gänzlich verborgen [Pow09].

Das Projekt *sachsMedia* an der Technischen Universität Chemnitz beleuchtet in seiner Forschungstätigkeit mehrere der oben genannten Aspekte mit dem Ziel der Unterstützung lokaler Medienanbieter. Im Themenbereich Annotation und Retrieval werden Verfahren zur automatisierten oder assistierten Merkmalsextraktion aus Medienströmen und die Strukturierung der erhaltenen Metadaten erarbeitet. Im Bereich Graphical User Interfaces stehen Usability von Annotationssystemen sowie endgeräteseitige Interaktionen im Interessenfokus. Im dritten Themenbereich Distribution werden Netzwerkstrukturen, Steuer- und Streamingmechanismen für eine zukunftstaugliche Medienverteilung beleuchtet.

2 Personalisierung durch Interessenbewertung

Die Auswahl des TV-Programms am stationären Fernsehgerät erfolgt aktuell zufallsbasiert (Zapping) oder durch gezielte Off-Screen- (Programmzeitschrift) oder On-Screen-Selektion (Electronic Program Guide). Für letztere wird wenig Nutzereingabe benötigt, da lediglich in Auswahllisten geblättert wird. Eine Schnittstelle zur weitergehenden Informationseingabe ist nicht vorgesehen. Anders als das TV-Gerät verfügt das Mobiltelefon jedoch über ein akzeptiertes Eingabe-Interface für relativ komplexe Textinformationen. Ebenso wie der Benutzer interessiert ist, sein Mobiltelefon zu individualisieren, ist er gewohnt, auf personalisierte Inhalte bspw. beim (mobilen) Surfen oder Musikhören zurückzugreifen [JvTD05]. In der Erlebniswelt des Video-On-Demand-Konsums sieht

sich der Nutzer allgemein anpassbaren Zugangsoptionen gegenüber. Er wählt aus Videostreams seiner Bekannten, seiner Region oder seines Interessensgebietes.

Aus der zunehmenden Akzeptanz dieser Gegebenheiten ist abzuleiten, dass Nutzer ihren Konsum zukünftig verstärkt auf individuell interessierende Beiträge zu selbstgewählten Zeitpunkten beschränken. Folglich ist das Forschungsziel der Autoren die Entwicklung eines Anwendungsszenarios, das es ermöglicht, unterwegs oder zu Hause auf personalisierte Medieninhalte zuzugreifen. Mittels Synergieeffekten aus Mobilfunk- und Broadcasttechnologie sowie eines ständig aktuell gehaltenen Beitragsstroms soll sichergestellt werden, dass bspw. die Abendnachrichten auch wirklich abends konsumiert werden können. Ein präzises Nutzerprofil soll zudem gewährleisten, dass auf dem Mobiltelefon nur Sendungen abrufbereit gehalten werden, die den Nutzer auch wirklich interessieren [ZHH07].

2.1 User Profiling

Flexibel gestaltbare Interaktionsplattformen mobiler Geräte wie bspw. Smartphones und deren akzeptierte Nutzerschnittstellen machen eine solche Profilbildung durch Datenakkumulation aus mindestens den folgenden Bereichen möglich:

1. Langzeitinteressen (Analyse der vom Nutzer eingegebenen grundsätzlichen Interessen)
2. Charakteristiken des Medien-Konsumverhaltens (kontinuierliche Datensammlung und -auswertung aus der Nutzerinteraktion)
3. Kurzzeitinteressen, temporär (Analyse zeitnaher Nutzerinteraktion)
4. Kurzzeitinteressen, lokal (stattfindende Ereignisse und interessante Orte in der aktuellen Umgebung des Nutzers)

Die Profilbildung und -speicherung ist zunächst vorrangig lokal auf dem Endgerät vorgesehen. Vorteil dabei ist vordergründig der Schutz persönlicher Daten. Beschreibungsdaten – also Tags – werden dem eingehenden Datenstrom entnommen und in einer Datenbank gesammelt. Eine zentrale Instanz verarbeitet die gewonnenen Daten aus der erfolgten Nutzerinteraktion, um die Tags (neu) zu bewerten. Betrachtet werden dabei Suchanfragen, Selektionen, Abbrüche oder Modifikation der Wiedergabe. Ebenso werden ortsbezogene Daten wie GPS-Positionen ausgewertet.

Die inhaltliche Strukturierung von Nutzerprofilen und darauf basierenden personalisierten Fernsehprogrammen steht aktuell im Fokus der Untersuchungen. Es gilt unterschied-

liche Anwendungsfälle der Inhaltsrezeption zu definieren und in Fallstudien zu evaluieren. Mittels Interviews sollen Daten, die in die Profile einfließen, und sinnvolle Gruppierungsmöglichkeiten der Inhalte identifiziert werden.

2.2 Media Profiling

Um Medienströme personalisiert anbieten zu können, bedarf es Klassifizierungsmethoden. Medienobjekte werden dabei mit Metadaten angereichert, welche sich möglichst an etablierte Beschreibungsstandards halten sollten. Mit MPEG-7 liegt ein solcher in strukturierter Form vor [Gro99] und bietet auf Basis vielzähliger Unterklassen Beschreibungsmöglichkeiten für sämtliche mediale Szenarien. Eine Katalogisierung inklusive textueller Inhaltsbeschreibungen ist im Bereich des Digitalfernsehens mit EPG (z.B. DVB-Event Information Table) oder ESG (bei DVB-H) ebenfalls standardisiert. Um Metadaten strukturiert – also maschinell – auswerten und verarbeiten zu können, bedarf es Richtlinien wie die des TV-Anytime- oder des RSS-Beschreibungsstandards. Diese definieren neben Titel, Genre und Beschreibungsfeldern Content Identifier zur eindeutigen Benennung der Quelle. Mittels dieser Werkzeuge ist es möglich, die inhaltliche sowie bezugstechnische Verbindung zwischen Anbieter und Nutzer herzustellen. Da diese Arbeit in der Folge ein Zusammenspiel zwischen Uni- und Broadcastübertragung fokussiert, wird die Bereitstellung der Inhaltskatalogdaten auf Basis der Punkt-zu-Punkt-Verbindung unter Nutzung von RSS realisiert. Gründe sind die geringen Datenmengen, die punktgenaue Möglichkeit zur Anforderung bei Bedarf und eine mögliche Parallelverwendung im stationären Bereich.

Für den Aufbau des Inhaltskataloges ist senderseitig eine strukturierte Datenhaltung nötig. Sämtliche Metadaten wie inhaltliche Tags, Geo- und Dringlichkeitsinformationen sind in einer gemeinsamen Datenbasis zu organisieren. Diese wird redaktionell bespielt und bietet Schnittstellen zur automatischen und assistierten Inhaltsannotation. Um eine schnelle RSS-Generierung zu bewerkstelligen, werden Indizes gehalten, die nach kategorierelevanten Schlüsseln strukturiert sind. Mittels einfacher Transformation kann aus den Indizes ein stets aktueller Satz an RSS-Katalogdaten per Webserver zum Abruf bereitgestellt werden. Content Identifier werden stets mitgeführt, um die Anknüpfung zur originären Metadatenbasis aufrecht zu erhalten.

2.3 Media Matching

Idealerweise kann durch Auswahl von Medienobjekten in Abhängigkeit des vorliegenden Nutzerprofils der mobile TV-Konsum wesentlich präziser auf die Nutzerinteressen

zugeschnitten werden. Angebotene Inhalte werden anhand mitgeführter Metadaten mit Nutzerdaten verglichen, gefiltert und sortiert. Idealerweise sollte der Beitrag, welcher den Nutzer kongruent zum Profil am meisten interessiert, direkt auf dem Endgerät abspielbereit gehalten werden. In einem Moment, in dem der Nutzer Medien konsumieren möchte, sollte durch minimale Interaktion deren Inhalt im Vollbildmodus präsentiert werden.

Aufmerksamkeitsspannen sind je nach Konsumintention oder Zeitbudget verschieden. Für kurzzeitige Zerstreuung empfehlen sich Beitragslängen von etwa 60-120 Sekunden. Bei länger währendem Interesse erfolgt der Aufruf des nächsten Beitrages, der dem Nutzerinteresse am nächsten kommt. Bei kontinuierlichem Konsum setzt sich somit ein vollständig zusammenhängendes, jedoch personalisiertes Fernsehprogramm aus mehreren Kurzbeiträgen zusammen. Zusätzlich ist eine Generierung mehrerer Individual-Programme möglich, die sich aus einer den Nutzerinteressen entsprechenden Kategorisierung ableitet. Analog dazu sollten Inhalte für längerfristige Aufmerksamkeitsspannen angeboten werden, da ebenso von einem Bedarf nach klassischen audiovisuellen Formaten ausgegangen werden muss [Pow09]. Eine Koexistenz von On-demand und Live-/Linear-Inhalten ist im Rahmen des Anwendungsszenarios zu realisieren.

3 Synchronisierte Mehrwege-Distribution

In bestimmten Situationen, z.B. bei Events großen öffentlichen Interesses, kann ein Ansturm – also viele kurz aufeinanderfolgende Anfragen – auf damit verbundene Inhalte leicht zu einer Überlastsituation führen. Einen Engpass bildet im mobilen Nutzungsszenario z.B. die Übertragungskapazität einer Mobilfunkzelle, da im herkömmlichen Fall selbst bei gleichzeitigem Abruf gleichen Inhalts zu jedem Nutzer eine Individualverbindung (Unicast) aufgebaut werden muss. Im Gegenzug bieten Broadcast-Netze wie DVB-T/-H attraktive Bandbreitenreserven, um populäre Inhalte großflächig und gleichzeitig zu verteilen. Der Aspekt des im vorigen Abschnitt diskutierten assistierten, personalisierten Medienkonsums soll jedoch ebenso einbezogen werden. Es gilt einen Konsens aus Verteilung massiv gefragter und individueller Verteilung spezialisierter Medienobjekte zu finden. Dies bedeutet eine Verknüpfung personalisierter Mediendistribution mit einer Symbiose aus Uni- und Broadcastarchitekturen.

Sowohl für die Sender- als auch die Empfängerseite können sich dadurch Vorteile ergeben. Die Broadcast-Nutzung für populäre Inhalte führt zu einer beidseitigen Kostenentlastung, da weniger Individual-Traffic über Mobilfunknetze transportiert werden muss. Selbst für Abonnenten günstiger Datentarife oder gar Flatrates ist eine solche Lösung interessant, da in genannten Überlastsituationen bisher die erwartete Quality of Experience

(QoE) nicht erreicht wird [Pow09]. Anbieterseitig profitiert man von niedrigeren Serverlasten, da weniger gleichzeitige Verbindungen nötig sind. In Summe sinkt die benötigte Bandbreite bei optimaler Nutzung des Broadcast-Netzes.

Für die Übertragung in einem Hybrid-Streaming-Szenario eignen sich bedingt durch hohe und kontinuierliche Datenraten besonders Videoinhalte. Der Umstand, dass sich mobile Sehgewohnheiten eher auf kurze Beiträge oder Live-Events konzentrieren, kommt einer personalisierten Hybrid-Sendestrategie entgegen. Anbieter bspw. von News-Portalen oder auch von lokalem Fernsehen erzeugen eine Vielzahl solcher (tages-)aktuellen Berichte, meist in Form kurzer Videoclips. Insbesondere für Inhaltsanbieter ist dieses Format günstig, bedenkt man technische Randbedingung wie Looping, Zwischenspeicherung oder eine hohe Nutzerfluktuation. Selbst das Playout klassischen linearen Fernsehens ist prinzipiell nur eine Aneinanderreihung einzelner Beiträge – allerdings mit vom Anbieter festgelegter Reihenfolge und Startzeit. Aus einer Vielzahl kurzer Beiträge lässt sich aber auch eine auf individuelle Vorlieben angepasste Zusammenstellung generieren.

Für die nachfolgenden Ausführungen gilt die Annahme der „80-20-Regel“, d.h. eine Vielzahl der Aufrufe gelten nur einem kleinem Teil des angebotenen Inhalts. Dieser verteilt sich demnach in „Mainstream“ (z.B. Top 10/100, Most viewed, Top rated) – den sogenannten Fat-Tail – und eine vergleichsweise hohe Anzahl nur relativ selten betrachteter Inhalte – den sogenannten Long-Tail [Pow09].

3.1 Hybrid-Streaming

Die Terminologie Hybrid-Streaming soll verdeutlichen, dass verschiedene Übertragungskanäle, -techniken, -verfahren und -protokolle in einem Szenario gemeinsam genutzt werden können. Die Übertragung per Internet Protocol bildet dabei den kleinsten gemeinsamen Nenner.

Individuelle Punkt-zu-Punkt-Verbindungen erfolgen typischerweise über Mobilfunknetze. Die Netze der 2.5 (GPRS/EDGE) und 3. Generation (UMTS/HSPA) arbeiten dabei paketorientiert. Daneben existiert mit WLAN eine weitere etablierte Drahtlosübertragungstechnik. Für die hier untersuchten terrestrischen Broadcast-Netze wird vom DVB-T- bzw. DVB-H-Standard ausgegangen. Die Professur Medieninformatik der Technischen Universität Chemnitz betreibt Sendetechnik beider Verfahren. Mit zwei in einem Single Frequency Network (SFN) synchronisierten Senderstandorten können so Indoor- bzw. mobile Outdoor-Szenarien in den Campusarealen sowie Teilen der Chemnitzer Innenstadt praktisch untersucht werden.

Broadcast-Netze eignen sich für eine großflächige Versorgung. Auf diesem Weg realisiertes Multicast-IP-Streaming ermöglicht so gleichzeitig eine Vielzahl entsprechend ausgerüstete Empfänger zu erreichen. Trotz der Hybrid-Möglichkeit soll sowohl für eine Unicast- als auch eine Broadcast-Übertragung – soweit wie möglich – Rückwärtskompatibilität erhalten bleiben. Dies bedeutet die Konzipierung einiger Funktionsblöcke so, dass sie auch ohne die Hybrid-Komponente funktionieren. Angestrebt wird dazu eine Kompatibilität im Bereich genutzter Broadcast-Netze zur klassischen Linear-TV-Übertragung nach DVB-T- bzw. DVB-H-Standard. Hierzu empfiehlt sich die Nutzung von Standardprotokollen, die ggf. geringfügig erweitert werden müssen. Für das gesamte Hybrid-Streaming-Szenario sind sowohl Anpassungen schon vorhandener bzw. etablierter Komponenten als auch die Einführung neuer, zusätzlicher Elemente nötig.

3.2 Anpassung mobiler Endgeräte

Da ein wie in Abschnitt 2 beschriebener Inhaltskatalog ständigen Aktualisierungen unterworfen ist, sollte der Zugriff auf diesen bevorzugt online erfolgen. Eine solche Unicast-Übertragung ist als unkritisch einzustufen, da mit jedem einzelnen Abruf normalerweise nur relativ geringe Datenmengen übertragen werden müssen.

Nach Bezug und Aktualisierung des Katalogs und einer erfolgten Verarbeitung im Media Matching-Prozess (s. Abschnitt 2.3) liegen im Endgerät eine oder mehrere gewichtete Listen verfügbarer Inhalte vor. Diese werden optisch ansprechend dargestellt: entweder innerhalb einer eigenen Benutzeroberfläche oder bspw. als aufbereitete HTML-Seite mit eingebetteten Links auf Medieninhalte, die durch Content-IDs/URIs eindeutig bezeichnet sind. Als Laufzeitumgebung bietet sich die auf den meisten modernen Mobiltelefonen verfügbare Java Micro Edition (Java-ME) an. Dabei können mittels Erweiterungsschnittstellen (JSR APIs) sinnvolle und für den Zugriff auf Empfangshardware bzw. Mobiltelefon-Basisfunktionalität nötige Komponenten einbezogen werden.

In die Geräte-Firmware sind ebenfalls oft Medienplayer mit Unterstützung gängiger audiovisueller Formate integriert. Zugriff auf deren Funktionalitäten kann bspw. über die weit verbreitete Java-MMAPI erfolgen [JCP06]. Verglichen damit sind bisher nur relativ wenige Geräte für den Empfang von Broadcast-Inhalten nach den Standards DVB-T bzw. DVB-H verfügbar. Selbst die Geräte, die serienmäßig den Empfang von linearem Live-TV mittels in der Firmware vorinstallierter Software ermöglichen, bieten keine standardisierten Möglichkeiten, um Software von Drittanbietern Zugang zum empfangenen Broadcast-Datenstrom zu gewähren. Zwar existiert mittlerweile mit JSR-272 eine

standardisierte DVB-H-Zugriffs-API für Java-ME, doch wurde diese bislang noch nicht durch die Mobiltelefonhersteller implementiert. [JCP08]

Der Broadcast-Empfang selbst muss deshalb noch gesondert untersucht werden. Bis dahin sind Testimplementierungen auf Basis von PC-Hardware zu realisieren. Diese bieten gut dokumentierte Schnittstellen zum Zugriff auf Broadcast-Hardware und -Datenströme. Alternativ verbleibt noch die Nutzung anderer IP-fähiger Netze, wie z.B. WLAN, auf entsprechend ausgestatteten Mobilgeräten.

3.3 Anpassung Media-Control-Server

Der Media-Control-Server (MCS) verwaltet senderseitig die zur Wiedergabe bereitgehaltenen Inhalte und stellt den aktuellen Inhaltskatalog zum Abruf bereit. Zusätzlich werden Daten zum Status einzelner Verbindungen (Sessions) gehalten. Dazu zählen auch Statusdaten zum Broadcast-Empfang der einzelnen Clients.

Ausgehend von diesen Daten erfolgt die Entscheidung, ob eine Medienübertragung per Unicast-Verbindung, über welche auch Inhaltsanforderungen vonstatten gehen, oder per Multicast über den zusätzlichen Broadcast-Kanal erfolgt. Abhängig von dieser Entscheidung übernimmt der MCS anbieterseitig das Ressourcenmanagement für die Übertragungswege. Im Unicast-Fall reagiert er entsprechend der Anforderungen des Clients, im Multicast-Fall über das Broadcast-Netz auf Basis der verwalteten Statusdaten der empfangenden Clients. Dabei sind verschiedene Übertragungsstrategien denkbar, siehe entsprechender Abschnitt.

Etablierte Protokolle aus der Realtime Protocol-Suite (RTP) bieten dafür gute Ausgangspositionen, sowohl für die Unicast- als auch Multicast-Nutzung. Das Realtime Streaming Protocol (RTSP) stellt sowohl mit den Requests *DESCRIBE*, *SETUP* und *TEARDOWN* die nötige Funktionalität zur Session-Steuerung als auch mit *PLAY* und *PAUSE* zur Medien-Steuerung bereit [SRL98]. Mit einem zusätzlichen Header-Feld, das den Broadcast-Empfangsstatus signalisiert, ist eine rückwärtskompatible Erweiterung des Protokolls leicht realisierbar.

Für weitere Anwendungsszenarien ist zur Etablierung einer Media-Session auch eine SIP-Signalisierung denkbar.

3.4 Anpassung der Medienübertragung

Im Fall einer Unicast-Verbindung soll RTP genutzt werden. Somit ist auch weitgehende Kompatibilität zu anderen nicht hybridstreaming-fähigen Clients gegeben, ohne dass

für diese gesonderte Medienserver bereitstehen müssen. Im Fall von Multicast über ein Broadcast-Netz sind im Zusammenhang mit optimierten Übertragungsstrategien verschiedene Transportprotokolle in Betracht zu ziehen. Für eine angestrebte – zumindest teilweise – Rückwärtskompatibilität zum DVB-H-/IP-Datacasting-Standard ist im Multicast-Hybrid-Fall ebenfalls ein Medientransport per IP/UDP/RTP interessant. [ETS06]

Auch wenn Datenströme verteilt über heterogene Netze wie Mobilfunk und DVB transportiert werden, muss die Wiedergabe beim Empfänger in der zeitlich korrekten Reihenfolge erfolgen. Geht man von einer Folge inhaltlich nicht zusammenhängender Beiträge aus, ist eine einfache Anpassung realistisch: Für die Synchronisation können die vorhandenen Möglichkeiten im RTP-Header genutzt werden [SCFJ03]. Für eine Abstimmung der Datenpakete beim Transport mittels einer einheitlichen RTP-Signatur bedarf es durch den MCS einer gezielten Wahl von *Sequence Number*, *Timestamp* und *Synchronization Source* (SSRC). Statt der üblichen initialen Zufallswerte pro Session, sollen diese Werte für alle Sessions eines spezifischen Medieninhaltes an deren Beginn gleich gewählt werden. Die Reihenfolge von Paketen, die über verschiedene Netze unter Anwendung heterogener Strategien eintreffen, ist somit eindeutig und leicht wiederherzustellen.

Im zweiten Fall soll davon ausgegangen werden, dass ein zusammenhängender Datenstrom übertragen wird. Somit muss eine ständige Synchronisation der RTP-Timestamps erfolgen. Dies kann durch RTCP-Sender Reports mit einer gemeinsamen Zeitbasis für alle Sessions erfolgen [SCFJ03]. Dazu müssen empfängerseitig die jeweiligen Zuordnungen zwischen Sequenz-Nummer, RTP-Timestamp, Coding-Timestamp und Normal Play Time (NPT) zueinander in einer Tabelle abgebildet werden. Die gemeinsame senderseitige Zeitbasis ist dabei von besonderer Bedeutung, da nur so ein lückenloses Zusammenfügen von Medienströmen aus scheinbar verschiedenen Quellen bzw. aus Kanälen mit zueinander unterschiedlichen Übertragungsstrategien ermöglicht wird. Im Idealfall geschieht dies bis auf Frame-Basis genau.

Unabhängig von der Synchronisationsmethode ist die notwendige Bedingung, dass unabhängig von dem durch den MCS festgelegten Transportnetz die gleiche Paketisierung der Mediendaten genutzt wird. Nach Umrechnung der Zeitstempel etc. müssen RTP-Pakete, die im lokalen Out-Buffer auf die gleiche Signatur abgebildet werden, auch das exakt gleiche Medienfragment/Sample enthalten.

3.5 Zwischenkomponente „Proxy“

Wie zuvor bereits erläutert, sind auf Empfängerseite verschiedene Modifikationen nötig. Eine notwendige Komponente zwischen lokalem Medienplayer und senderseitigem

Medienserver ist ein Proxy, um das Streaming für vorgegebene geräteinterne Prozesse transparent zu halten. Sowohl Medienkontrolle als auch -transport müssen darüber geführt werden.

Dem Proxy kommen verschiedene Funktionen zu: Dem internen Mediaplayer gegenüber agiert er als RTSP-/RTP-Server. Dabei nimmt er RTSP-Befehle entgegen und liefert Mediendaten an lokale UDP-Ports. Die aufbereiteten internen URI des Inhaltskataloges formt er in die entsprechende externe URI-Repräsentationen um. Dem externen MCS gegenüber agiert der Proxy als RTSP-Client, der Steuer-Requests zum Etablieren einer Session und zur Medienwiedergabe generiert.

Zusätzlich übernimmt der Proxy noch die Kontrolle des Broadcast-Empfangs. Durch geeignete Methoden, wie z.B. zusätzlichen RTSP-Headern, wird dem MCS bei Inhaltsanforderungen zusätzlich der Broadcast-Empfangsstatus mitgeteilt. Der Empfang relevanter Datenpakete per Broadcast-Netz wird ebenfalls vom Proxy initiiert und überwacht. Empfangene Pakete werden unabhängig vom Übertragungsweg zwischengespeichert. Dieser Vorgang unterteilt sich in einen In-Buffer für empfangene Daten und einen Out-Buffer, der für die interne Wiedergabe durch den Medienplayer genutzt wird. Für die empfängerseitige Synchronisation der Unicast-Daten und Multicast-Daten ist ebenfalls der Proxy zuständig.

Der Proxy soll selbständig zwischen Unicast- und Broadcast-Empfang wählen können. Er sollte also bei einer vorhandenen Möglichkeit, die Multicast-Datenströme des Broadcast-Netzes zu empfangen, diese nutzen und die Unicast-Verbindung pausieren bzw. beenden. Im Umkehrfall, z.B. bei sich verschlechternden Broadcast-Empfangsbedingungen, muss er ein rechtzeitiges automatisches Fallback auf Unicast bei leerem In-Buffer beherrschen. Der Out-Buffer muss somit stets Daten für eine unterbrechungsfreie Wiedergabe bereithalten. Zudem soll per geräteinterner Kommunikation mit der steuernden Personalisierungs-Komponente die clientseitige Auswertung der Nutzungsinteraktion (Pausieren, Vorspulen, vorzeitiger Abbruch der Wiedergabe etc.) ermöglicht werden. Für die Auswertungsunterstützung ist der Proxy somit auch in Fällen reinen Unicast-Streamings nötig.

3.6 Übertragungsstrategie

Eine Variante der Mediennutzung ist On-Demand, also genau das Gewünschte zur gewünschten Zeit, was oft „Sofort!“ heißt. Diesen Anwendungsfall können Unicast-Netze gut abdecken. Im Unicast-Fall spricht nichts gegen die Nutzung üblicher und im mobilen Umfeld erprobter Übertragungsstrategien wie linearem Streaming mit Übertragung in

Echtzeit mit geringer Vorpufferung um kleine Schwankungen der Übertragungsbandbreite ausgleichen zu können. Ein vorauseilendes Laden unter Nutzung der gesamten verfügbaren Bandbreite macht hingegen nicht unbedingt Sinn, auch bei Nicht-Live-Inhalten, da ein Client die Wiedergabe vorzeitig Abbrechen könnte (Zapping-Effekt). Möglichkeiten wie wahlfreier Zugriff (Vor- / Rückspulen etc.) werden auch beim Echtzeit-Streaming, z.B. auf Basis von RTSP/RTP unterstützt.

Das Broadcasting linearen Live-TVs, z.B. gemäß den Vorgaben des DVB-H-Standards, ist der Trivialfall, der allerdings nur eingeschränkt die Nutzerwünsche repräsentiert. Alle Nutzer erhalten den selben Inhalt, zu quasi fest vorgegeben Zeiten. Selbst bei planmäßigen Wiederholungen der Beiträge in einem Playout-Karussell verpassen Späteinsteiger ohne Vorabspeicherung den Anfang eines Beitrags, andere Nutzer hingegen müssten bis zum Beginn eines für sie interessanten Beitrags unter Umständen lange warten.

Um sich den Nutzerwünsche anzunähern, ist im Broadcast-Fall eine fortschrittlichere Übertragungsstrategie notwendig. Wenn viele Nutzer sich innerhalb eines relativ kurzen Zeitfensters für den selben Inhalt interessieren, z.B. für die Nachrichten um Acht, dann kann dieser Gleichzeitigkeitseffekt zu einer sinnvollen Near-Video-on-Demand (NVoD) Nutzung des Broadcast-Netzes führen. Ausgehenden von diesen Nutzerverhaltensmustern muss der richtige Zeitpunkt zum Wechseln vom Unicast zum Broadcast gewählt werden. Besonders bei Live-Events ist innerhalb kurzer Zeit mit einem starken Anstieg, allerdings auch Abfallen des Nutzerinteresses zu rechnen, Ansätze zur Analyse finden sich z.B. in [TM09].

Übertragungsstrategien, die auf der Zerlegung des Medienstroms in Segmente basieren, lassen sich in verschiedene Gruppen unterteilen. Eine Gruppe bilden Schemata, die auf der Bildung von Blöcken ansteigender Größe und der parallelen Übertragung in mehreren Kanälen mit gleichen Bandbreitenanteilen beruhen. Zu dieser Gruppe gehören z.B. Pyramid- und Skyscraper-Strategien. Eine weitere Gruppe charakterisiert sich durch die Zerlegung in Segmente gleicher Länge, die aber mit unterschiedlichen Datenraten transportiert werden. Dieser Gruppe wird als Harmonic Broadcasting klassifiziert. Die Vorteile beider Verfahren nutzen Strategien auf Basis des Pagoda-Broadcasting. Der Medienstrom wird in Segmente und die Übertragungskanäle in Zeitschlitze jeweils konstanter Länge zerteilt. Durch eine verschachtelte Zuordnung der Mediensegmente auf die Zeitschlitze können so ohne extreme Bandbreitenspitzen und mit begrenzter Zahl paralleler Kanäle noch attraktive Einstiegswartezeiten erzielt werden. [PCL99]

Ein technische Analyse und Performance-Evaluation zu den verbreitetsten Broadcast-Schemata findet sich ebenfalls in [PCL99]. Bei einer genügend großen Anzahl möglicher

„Unterkanäle“, mit nur jeweils geringem Anteil der Gesamtbandbreite, sind demnach Schemata zu bevorzugen, die eine Vielzahl Segmente parallel übertragen.

Client-Centric-Approach (CCA) und CCA+ [NCW09] verfeinern die stufenweise Segmentierung und Verschachtelung noch weiter, werden aber auch nicht den eingangs formulierten Zielen einer weitgehenden Rückwärtskompatibilität zu Empfängern ohne spezielle Anpassungen, die nur lineares Streaming im Basiskanal empfangenen können, gerecht.

Welche Übertragungsstrategie sich in der Praxis unter diesen Vorgaben bewährt, muss noch im Detail untersucht werden. Ob sich ein „festes“ Schema überhaupt eignet, ist zum jetzigen Zeitpunkt noch nicht klar. Eventuell kann nur mit einem vollkommen dynamischen Ressourcenmanagement auf Seite des Anbieters eine optimale Entlastung des Unicast-Netzes und optimale Auslastung des Broadcast-Netzes erreicht werden. Ein möglicher Ausgangspunkt könnte eine feste Anzahl linearer Basiskanäle mit weiteren adaptiven Unterkanälen sein. Der Wechsel der Inhalte in den Basiskanälen und flexibel hinzugefügter Datensegmente zielt auf beste Ausnutzung der Broadcast-Kapazität ab, in dem Sinn, dass möglichst die größte Zahl Nutzer vom Broadcast-Kanal profitieren soll.

Um den nicht exakt gleichen Einstiegszeitpunkten gerecht zu werden, sind unabhängig von einem speziellen Schema Wiederholungen verschiedener Inhaltssegmente unvermeidlich. Zudem muss der Empfänger prinzipiell immer in der Lage sein, mehrere Datenströme parallel zu empfangen und zu verarbeiten sowie ggf. eine größere Datenmenge zwischenzuspeichern. Je nach Verfahren kann dies ein Vielfaches der Bandbreite sein, die für ein lineares Echtzeit-Streaming gebraucht würde.

Inwieweit sich ein Mix aus fortschrittlichen DVB-H-Techniken wie Time Slicing, also der Übertragung in Bursts, und ergänzenden Zusatzkanälen mit Segmentwiederholungen vereinbaren lässt, muss noch abgeklärt werden.

Selbst mit fortschrittlicher Übertragungsstrategie im Broadcast-Kanal und Vorabspeicherung kann sich für die Nutzer eine gewisse Wartezeit bis zum Wiedergabestart ergeben. Um die anfänglichen Wartezeiten zu überbrücken / zu vermeiden sind verschiedene Varianten denkbar:

1. Wiedergabe schon vorgespeicherter Inhalte
2. Falls diese nicht vorhanden sind: Start des Unicast-Streamings
3. Einblenden eines kurzen Werbeclips, der per Broadcast-Kanal übertragen wird, ggf. nach minimaler Wartezeit

4. Streaming eines (personalisierten) Werbeclips per Unicast-Verbindung

So wird Zeit bis zum Empfang des eigentlich interessanten Inhalts gewonnen, ohne dass der Nutzer auf einen leeren Bildschirm starren muss. Daraus können sich auch erste Ansatzpunkte für eine kommerzielle Verwertung ergeben.

4 Offene Fragen

Durch weitere (empirische) Untersuchungen müssen die Nutzerwünsche und das Nutzerverhalten noch detailliert belegt werden. Aus welchen Variablen soll sich ein Nutzerprofil genau zusammensetzen? Wie sollen diese Variablen zueinander gewichtet werden? Welche mittlere Wartezeit bis zum Start der Wiedergabe eines gewünschten Inhalts wird von der Mehrzahl der Nutzer noch akzeptiert? Gibt es im Fall von kontinuierlichen Sendungswiederholungen (Karrussellübertragung) z.B. eine gaußsche Normalverteilung für den Zeitpunkt des Wiedergabestarts auf dem Endgerät oder vorwiegend Früh- bzw. Späteinsteiger? Zappen die Nutzer sehr oft, d.h. werden viele Inhalte nur kurz angespielt? Kann man die Übertragungsstrategie darauf anpassen?

Entscheidend für die hybride Distribution wird die Verfügbarkeit geeigneter Endgeräte sein. Neben den für die Distribution spezifischen Aspekten wie Übertragungs- und Steuerprotokollen etc. muss ebenfalls die Seite der Medienproduktion betrachtet werden. Welche Produktionsvorgaben für Beiträge sind an die Content Provider zu richten (Beitragsdauer, Beitragsformat, Verschlagwortung)? Auch müssen für die Finanzierung eines entsprechenden Systems noch geeignete Ansätze konkretisiert werden, wie bspw. o.g. Werbeeinblendungen.

5 Zusammenfassung und Ausblick

In diesem Artikel wurde mit dem Ansatz eines personalisierten Mobilfernsehens ein mögliches Einsatzszenario hybrider Netze vorgestellt. Ausgehend von Nutzerinteressen erfolgt eine Vorauswahl bzw. persönliche Zusammenstellung verfügbarer Medien. Unter einer einheitlichen Nutzeroberfläche werden dabei sowohl populäre Fat-Tail-Inhalte als auch spezielle Long-Tail-Inhalte präsentiert. Das Konzept sieht sowohl die individuelle Unicast-Nutzung als auch die Multicast-over-Broadcast-Nutzung vor. Ausgehend von vorhandenen Komponenten auf Empfänger- und Anbieterseite wurden nötige Systemanpassungen vorgestellt. Zusätzlich wurde mit dem Proxy-Konzept eine Möglichkeit

zur Nutzung vorhandener Komponenten, wie z.B. den integrierten Medienplayern, gezeigt, ohne an diesen selbst Veränderungen vornehmen zu müssen. Das Nutzungsszenario geht von Unicast-Streaming sowie fortschrittlichen Übertagungsstrategien im Broadcast-Kanal aus. Weitere Untersuchungen müssen die noch offenen technischen als auch auf Nutzerakzeptanz und -verhalten bezogenen Fragen klären.

Literatur

[dL08] Gemeinsame Stelle Digitaler Zugang der Landesmedienanstalten. Digitalisierungsbericht 2008. Technical report, Arbeitsgemeinschaft der Landesmedienanstalten in der Bundesrepublik Deutschland (ALM) und Gemeinsame Stelle Digitaler Zugang (GSDZ), September 2008.

[ETS04] ETSI. ETSI EN 302 304: Transmission System for Handheld Terminals (DVB-H). European Standard, European Telecommunications Standards Institute, 2004.

[ETS06] ETSI. ETSI TS 102 472: IP Datacast over DVB-H: Content Delivery Protocols. Technical Specification, European Telecommunications Standards Institute, 2006.

[For09] UMTS Forum. Fast Facts. Technical report, UMTS Forum, 2009. Abruf am 13.05.2009.

[Gro99] MPEG-7 Requirements Group. MPEG-7: Context, Objectives, and Technical Roadmap, V.12. Technical report, MPEG-7 Requirements Group, July 1999.

[JCP06] JCP. JSR 135: Mobile Media API. Java Spezifikation Request, Java Community Process (JCP), 2006. Abruf am 15.05.2009.

[JCP08] JCP. JSR 272: Mobile Broadcast Service API for Handheld Terminals. Java Spezifikation Request, Java Community Process (JCP), 2008. Abruf am 15.05.2009.

[JvTD05] Ivar Jørstad, Do van Thanh, and Schahram Dustdar. The Personalization of Mobile Services. In *IEEE International Conference on Wireless And Mobile Computing, Networking And Communications, 2005*, pages 59–65, 2005.

[NCW09] Ashwin Natarajan, Ying Cai, and Johnny Wong. An enhanced client-centric approach for efficient video broadcast. *Multimedia Tools Appl.*, 43(2):179–193, 2009.

[PCL99] Jehan-francois Paris, Steven W. Carter, and Darrell D. E. Long. A hybrid broadcasting protocol for video on demand. In *Multimedia Computing and Networking Conference*, pages 317–326, 1999.

[Pow09] Jerry Power. How Carriers and Suppliers Are Evolving to Provide Next Generation Mobile TV Service. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2009.

[rGPP06] 3rd Generation Partnership Project. UTRA-UTRAN Long Term Evolution (LTE) and 3GPP System Architecture Evolution (SAE). Technical report, 3rd Generation Partnership Project, 2006.

[SCFJ03] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC, Internet Engineering Task Force (IETF), 2003. IETF RFC3550.

[SRL98] H. Schulzrinne, A. Roa, and R. Lanphier. Real Time Streaming Protocol (RTSP). RFC, Internet Engineering Task Force (IETF), 1998. IETF RFC2326.

[TM09] Saurabh Tewari and Satish Menon. On Resource Provisioning in Hybrid Peer-to-Peer Live Streaming Systems. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2009.

[TNS09] TNSInfratest. Mobilfunknutzung und Nutzungsabsichten 2009. Studie, TNSinfratest & E-Plus Gruppe, 2009.

[ZHH07] Yan-quan Zhou, Ying-fei Hu, and Hua-can He. Learning User Profile in the Personalization News Service. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2007*, pages 485–490, 2007.

Evaluation of an Image and Music Indexing Prototype

Peter Dunker, Ronny Paduschek, Christian Dittmar, Stefanie Nowak

and Matthias Gruhne

Fraunhofer Institute for Digital Media Technology IDMT

{dkr,pdk,dmr,nwk,ghe}@idmt.fraunhofer.de

Abstract: This paper describes a technical solution for automated semantic indexing of music and images for a media archive environment. The indexing is based on a multi-modal low-level feature extraction and semantic high-level feature classification such as mood, genre, daytime or visual scene types. The classification on both, the audio and the visual information is based on a generic machine learning core architecture. A combination and cleansing process validates for improving the classification results. This paper presents the technical realization of a prototype and its corresponding evaluation. Finally, the practical relevance of this technology results, based on the findings of the evaluation is discussed.

Keywords: multi-modal media indexing, media archives indexing, music retrieval, image retrieval

1 Introduction

In the last years the amount of digital content in TV archive environments as well as in user generated media archives increased dramatically. Especially, the management of these archives and retrieval of certain media items gets more and more complicated. New techniques for automated indexing, organizing and searching in media archives with less manual effort for annotation of individual media items are required. We propose a technology which allows the automatic content-based annotation of music and images with descriptive information such as mood, genre, music-color or visual scene types like city, beach or sunset. Based on that core technology we present a prototype that enables searching and navigating in a media archive. Finally, we present evaluation results of the proposed system and discuss the benefit of this technology.

2 Media Indexing

The media indexing process is based on a audio visual feature extraction, a classification and a combination and cleansing step as depicted in Figure 1. The complete process as

well as the evaluation is realized in a generic and multi-modal machine learning framework utilizing various feature extraction algorithms. The modules subsequent to the feature extraction, e.g. the classification which consists of a training and classification part, are developed independently of the targeted media type. The final semantic annotations of the image and music data are stored in a MPEG-7 XML document, which could easily be changed to other meta data formats, e.g. TV Anytime [PS00]. The following sections describe the components of the indexing workflow.

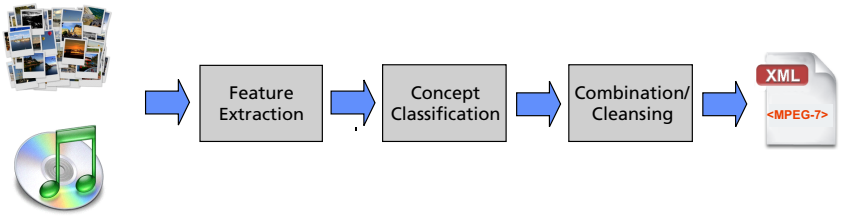


Figure 1: Cross-Media Semantic Indexing

2.1 Feature Extraction

Several visual low-level features are applied as input for the image classification. Most of the features are derived from the MPEG-7 visual descriptors [ISO01], e.g. *color layout*, *color structure*, *scalable color*, *edge histogram*, *dominant color* and *color temperature*. Further incorporated low-level features are: a *haar-wavelet energy feature* and a *blur detection factor*. In addition to the content-based features, the following EXIF information are used as meta features: *exposure time*, *F-number*, *focal length*, *focal length in 35 mm*, *ISO speed* and *flash*.

The following audio descriptors [ISO01]are used: *Log Loudness*, *Norm Loudness*, *Mel-Frequency Cepstral Coefficients*, *Audio Spectrum Envelope*, *Spectral Centroid*, *Spectral Crest Factor*, *Spectral Flatness Measure*, *Zero Crossing Rate* and *Enhanced Pitch Class Profiles (EPCP)* as described in [Lee06]. Besides, this set of low-level features, several specialized mid-level representations [DBG07] have been developed. These mid-level features range from simple modulation coefficients computed over low-level features, to autocorrelation-based rhythmic patterns. Additionally, histograms of note and chord candidates, that have been derived from a EPCP-based chromagram, were utilized.

2.2 Concept Classification

For the classification of audio and image concepts a Gaussian Mixture Model (GMM) classifier was applied. The decision for a category was performed by choosing the category with the highest probability and an additional test against an experimental elaborated threshold. For the image classification, three independent GMMs are trained, two with content-based feature sets and one with EXIF data only.

For the semantic classification of images, we chose the following scene concepts: scene I: *Indoor* and *Outdoor*, scene II: *Day*, *Night* and *Sunset*, and scene III: *Architecture*, *Beach* and *Snow*. The music classification based on the low- and mid-level features exhibits the following semantic categories: *Genre*, *Color* and *Music Texture*.

In addition, for both media types the following mood aspects are classified: *Valence*, *Arousal* and the four mood entities *Aggressive*, *Euphoric*, *Calm* and *Melancholic*. The underlying mood model was already introduced in [DNBL08].

2.3 Combination and Cleansing

With reference to visual classification, we created different classifier models during the preceding training process for each visual category. For instance, the classifiers are separately trained with a pre-estimated feature selection, extracted from EXIF data or content-based features. The achieved classification results of each model of a specific category are aggregated in different modalities, e.g. at least two models have to vote for the same category to annotate the media with this specific category. With this approach imprecise assertions are largely excluded through cleansing of weak classification results. In addition, the assignment of a media item to a category is validated by an experimentally estimated threshold for each class.

3 Prototype

Based on the proposed core technology components, a prototype was build to demonstrate the functionalities. The prototype combines a solution for several multi-modal search tasks. The prototype processes the media files and stores the MPEG-7 meta data

output with the semantic annotations and corresponding confidence values in a proprietary XML database core, which enables a combined search across image and music annotations. Figure 2 illustrates the image view of the prototype with a table of the indexed data. The columns of the table refer to the semantic categories of the media items that are arranged in the table rows. A two dimensional visualization based on mood char-

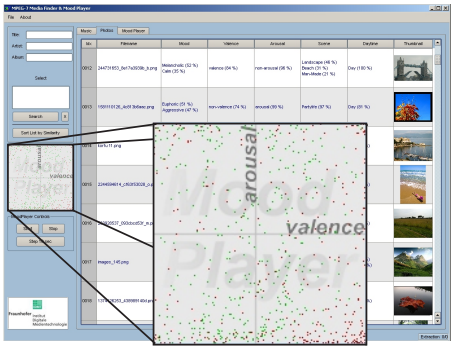


Figure 2: Screenshot of the media indexing prototype.

acteristics of all media items could be achieved as depicted in Figure 2, by incorporating a range search on the confidence values of the Valence and Arousal categories. Each green dot represents a music item, each red dot represents an image item. This allows a simple selection of a subset of items by marking a region in this two dimensional plane. Finally, a search interface is integrated to define cross-modal semantic queries such as *"I'd like soft, bright pop music and a slideshow of sunset photos"*.

4 Evaluation

The classification modules are trained with audio and image data from various sources. The audio training and test data (about 600 songs) were collected from commercial audio CDs or web-portals like Last.fm by three experienced musicians. The image data set was compiled from the photo community Flickr by using different keywords per search and a manual review process. The scene set consists of about 8000 and the mood set of about 400 images. The benchmark utilizes a Monte-Carlo test, which randomly changes the disjunctive training and test sets in order to compute average classification results. For the image classification, different aggregation strategies of multiple GMMs were compared. The presence of at least two independent decisions out of the three GMMs increased the overall accuracy.

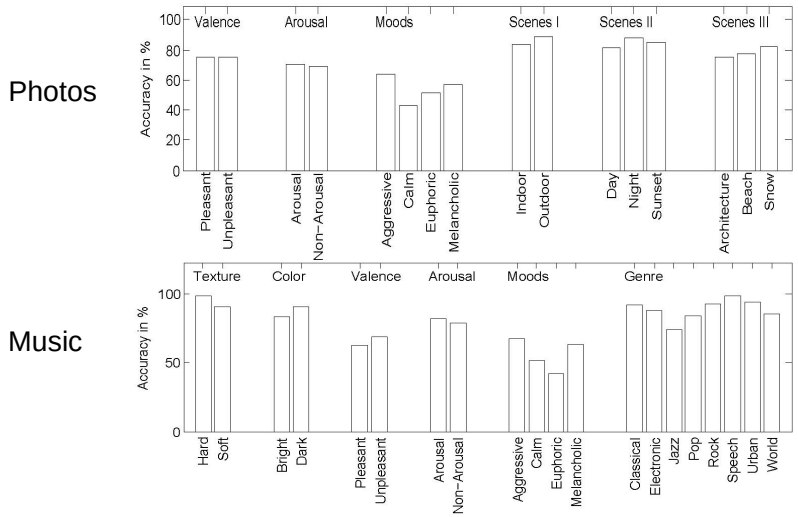


Figure 3: Audio and Image Evaluation Results

As shown in Figure 3, the different semantic categories perform quite well. The scene concepts were examined separately. The results over all scene classes average to 82.62%. Speech poses the best performing class in genre detection, whilst Jazz and Pop perform worse. The mood-related classifiers performed worst in both domains, which is caused by the subjective nature of moods and the resulting difficulty to obtain well-defined ground truth data.

5 Conclusions and Further Work

In this paper, we proposed a prototypic system for multi-modal indexing on music and images with a semantic interface for mood based media search. The evaluation of all semantic annotations show that selected categories perform very well (e.g. genre and scene) while especially the mood descriptions lack on accuracy. Nevertheless, the proposed system offers a beneficial set of semantic description that could help to simplify search and retrieval in media archives. A precise pre-selection of significant features adjusted to each category separately could contribute to a further improvement of the results. Future work is also focused on post-processing methods, e.g. an intelligent combination of results, which could be particularly realized by using real-world knowledge.

The next working steps on the proposed technology concentrate on the enrichment of semantic categories as well as on the transfer of the existing algorithms to the video domain.

Acknowledgements

This work has been partly supported by grant No. 01MQ07017 of the German research program THESEUS funded by the Ministry of Economics and Technology and partly supported by the German research project GlobalMusic2One funded by the Federal Ministry of Education and Research (BMBF-FKZ: 01/S08039B).

References

[DBG07] C. Dittmar, C. Bastuck, and M. Gruhne. Novel mid-level audio features for music similarity. In *Proc. of the Intern. Conf. on Music Communication Science (ICOMCS), Sydney, Australia*, 2007.

[DNBL08] P. Dunker, S. Nowak, A. Begau, and C. Lanz. Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In *Proc. of the Int. Conf. on Multimedia Information Retrieval (ACM MIR), Vancouver, Canada*, 2008.

[ISO01] ISO/IEC. ISO-IEC/JTC1 SC29 WG11 Moving Pictures Expert Group, Information Technology - Multimedia Content Description Interface. 2001.

[Lee06] K. Lee. Automatic chord recognition from audio using enhanced pitch class profile. In *Proc. of the Intern. Computer Music Conference (ICMC), New Orleans, USA*, 2006.

[PS00] S. Pfeiffer and U. Srinivasan. TV Anytime as an application scenario for MPEG-7. In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 89–92. ACM New York, NY, USA, 2000.

Aspekte inhaltlicher Modellierung von Musikdokumenten in digitalen Archiven

Michael Rentzsch und Frank Seifert

Technische Universität Chemnitz
Fakultät für Informatik
Professuren Informationssysteme und Softwaretechnik, Datenverwaltungssysteme

{mren,fsei}@cs.tu-chemnitz.de

Zusammenfassung: Aktuelle Methoden des Music Retrievals sind nicht in der Lage, Musikstücke wiederzuerkennen, die stärkeren Variationen unterzogen wurden, wie dies bei Live-Einspielungen oder Improvisationen oft der Fall ist. Ursache ist das fehlende Wissen über musikalische Funktion und Zusammenhänge. Im Rahmen des Projekts Musikalische Datenbanken wird versucht, ein Modell zur Speicherung dieses Wissens zu finden und zu implementieren. Damit sollen Methoden gefunden werden, die das Wiedererkennen eines Musikstücks auch dann ermöglichen, wenn es z. B. als Improvisation mit veränderter Stilrichtung vorliegt.

Schlagwörter: Music Information Retrieval (MIR),
Semantische Modellierung, Mustererkennung

1 Einleitung

Spätestens seit sich perzeptuelle Komprimierungsverfahren für Musikdateien durchsetzten – der bekannteste Vertreter dieser Verfahren ist sicherlich MPEG-1 Audio Layer 3, kurz MP3 – haben sich große Mengen an Musikdaten in verschiedenen Formen, sei es als öffentliche Sammlungen im Internet, als Musikkollektionen in Sendeanstalten oder als private Sammlung eines Benutzers auf dem heimischen Computer angesammelt. Zu diesen reinen Musikarchiven kommen dabei noch Videos, deren Inhalte in den meisten Fällen mit Musikstücken unterlegt sind.

Verstärkt wurde dieser Trend in den letzten Jahren durch den Einsatz von Abspielgeräten für komprimierte Musikdateien, den Verkauf von Musik über das Internet durch Anbieter wie Apple oder Amazon und – als neuester Trend – das Aufkommen reiner Internetradiostationen und microblog-ähnlichen Diensten für Musik, z. B. blip.fm. So wurden im Jahr 2008 schon 20 % der Musikverkäufe in digitaler Form beim Kunden ausgeliefert – zum Vergleich: Im Jahr 2006 waren es noch 10 % [DMR2009].

Datenmengen dieser Größenordnungen können nur schwer gespeichert und effektiv angefragt werden. Nach wie vor werden große Teile der zur Verfügung stehenden Musikdaten nur über Metadaten (Titel, Komponist, Interpret, etc.) verwaltet und identifiziert [Lee2004]. Diese Art der Speicherung (und Abfrage) wird von heutigen Computersystemen, etwa relationalen Datenbanksystemen, perfekt unterstützt. Eine

inhaltsorientierte Suche ist damit aber nicht möglich. Die Anfrage nach einem Stück, das „so ähnlich“ wie ein anderes klingt, kann mit diesen Mitteln nicht beantwortet werden.

Mit der inhaltsorientierten Verwaltung und Suche von Musikdaten beschäftigt sich das *Music Information Retrieval* (MIR). Einige Teilgebiete des MIRs [DOWN2003] sind:

- Retrieval (Query By Humming, Polyphones Retrieval)
- Wiedererkennung von Musik (Audio-Fingerprinting, u. a.)
- Automatische Transkription
- Rhythmus- und Harmonieanalyse
- Segmentierung
- Notenverfolgung (Score Following)
- Ähnlichkeitsanalyse und Visualisierung

Besonders die ersten beiden Disziplinen sind für die Verwendung in digitalen Archiven interessant. Beim Query By Humming (QBH) wird eine Anfrage an eine Sammlung von Musikdokumenten durch Summen oder Pfeifen einer Melodie gestellt. Zurückgegeben werden die Musikstücke, die die vorgegebene Melodie – unter Einbeziehungen von Toleranzschwellen – enthalten. Beim Audio-Fingerprinting werden kleine, eindeutige Mustercodes (analog zu Fingerabdrücken) aus Audiodokumenten generiert, die dazu dienen, dasselbe Musikstück schnell wiedererkennen zu können. In den letzten Jahren wurden große Fortschritte in diesen beiden Bereichen erzielt [Dann2003, Dann2004, Wang2008] und [Cano2002, Mitro2006, Balu2007].

Verfahren, die QBH oder Audio-Fingerprinting verwenden, sollten natürlich gegen kleine Fehler in den Anfragedaten, z. B. Ungenauigkeiten beim Summen der gesuchten Melodie oder Rauschen im Hintergrund einer Audioaufnahme, robust sein. Dies wird in gewissem Maße von aktuell existierenden Methoden umgesetzt. Dabei können viele Fingerprinting-Verfahren sogar Dokumente korrelieren, die in verschiedenen Sampleraten vorliegen.

Leider sind diese Verfahren, besonders Audio-Fingerprinting, im Regelfall darauf beschränkt, ausschließlich dieselbe Instanz eines Musikstückes zu erkennen. Liegt ein Musikstück in einer stärker veränderten Variante, z. B. einer Live-Einspielung oder einer Improvisation, vor, ist eine (Wieder-) Erkennung nicht mehr möglich. Ursache sind die verwendeten Vergleichsalgorithmen und das fehlende Wissen über musikalische Zusammenhänge und Funktionen der gespeicherten und zur Suche herangezogenen Muster.

Im Rahmen des Projektes *Musikalische Datenbanken* wird deshalb versucht, ein Modell für die Repräsentation von Musik zu finden und umzusetzen, das musikalische Muster um Wissen über Zusammenhänge und Funktion in einem Gesamtwerk erweitert. Mit Hilfe des Wissens und geeigneten Vergleichsmethoden sollen Algorithmen

geschaffen werden, die auch live eingespielte und improvisierte Varianten, wie sie z. B. im Jazz üblich sind, von Musikstücken erkennen.

Im folgenden Abschnitt erläutern wir das Modell zur Repräsentation von Musik, das im Rahmen des Projekts entwickelt wurde. Danach beschreiben wir den Erkennungsprozess unter Verwendung dieses Modells. In den Abschnitten 4 und 5 gehen wir auf die praktische Umsetzung ein und erläutern bisher erreichte Ergebnisse. Den Abschluss des Beitrags bildet ein Ausblick auf weiterhin anstehende Arbeiten.

2 Hierarchische Modellierung musikalischer Merkmale

Um ein geeignetes Modell zur Repräsentation von Musik zu finden, muss man das menschliche Hören und Wahrnehmen analysieren. Hören wir ein Musikstück sehr häufig, können wir dieses „trainierte“ Stück bereits an einem sehr kleinen Ausschnitt wiedererkennen. Dieses Phänomen ist dann besonders ausgeprägt, wenn die einzelnen Abschnitte relativ einzigartig sind. Auf der anderen Seite werden wir dieses Musikstück auch in einer variierten Form, z. B. in einem anderen Genre oder mit verändertem Rhythmus, problemlos erkennen.

Die menschliche Wahrnehmung funktioniert also in dieser Hinsicht in zwei Richtungen, Top-Down und Bottom-Up. Ein Bottom-Up-Ansatz repräsentiert das Wiedererkennen von Musikstücken auf Grund bereits kleiner Ausschnitte. Eine Top-Down-Methodik modelliert das Identifizieren von bisher unbekannten Stücken über wiedererkannte Teile. Diese Eigenschaften des Wahrnehmungsprozesses werden durch unser Modell zur Repräsentation von Musik aufgenommen [Seif2008].

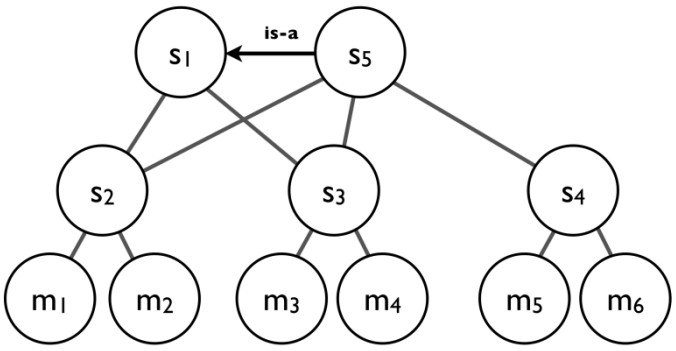


Abbildung 1: Zwei Templates in is-a-Beziehung

Ausgangspunkt ist eine hierarchische Beziehung zwischen verschiedenen Elementen zur Beschreibung von Musik. Jedes Stück wird zuerst durch eine in höchstem Maße generische Variante modelliert. Die entstehende Struktur wird als *Template* bezeichnet. Wir nehmen an, dass diese Variante unserer Erinnerung an das Musikstück entspricht. Weitere Instanzen dieses Stückes, die wir ebenfalls besonders gut kennen und deren

spezielle Eigenschaften, z. B. ein bestimmter Rhythmus, werden durch zusätzliche Templates dargestellt, die mit dem „Original“ in einer *is-a*-Beziehung stehen.

Zur Modellierung der elementaren Elemente von Musik, die neben Struktur und Kontext, die (Wieder-) Erkennung von Liedern und Werken ermöglichen, führen wir sogenannte *Generic Perceptual Music Patterns* (GPM-Pats) ein. GPM-Pats stehen für minimale, bedeutungstragende musikalische Muster, die als perzeptuelle Einheit wahrgenommen werden. Im Prinzip sollten diese Muster also in das menschliche Kurzzeitgedächtnis passen und nur wenige Sekunden dauern. Als mögliche Patterns sind Tonfolgen, Rhythmen bzw. zeitliche Abfolgen, Harmoniefolgen oder auch Klangfarben denkbar. Abbildung 2 zeigt zwei verschiedenartige GPM-Pats aus dem Jazztitel „So What“.

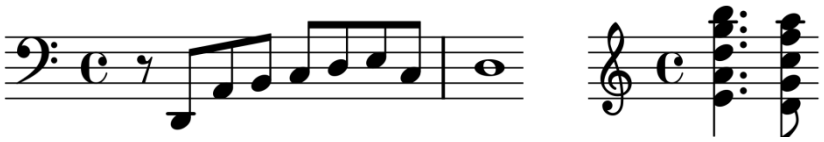


Abbildung 2: Verschiedenartige GPM-Pats

Den zweiten Bestandteil zur Beschreibung musikalischer Eigenschaften bildet der *Kontext*. Er modelliert implizites Wissen, z. B. über die harmonischen Zusammenhänge zwischen Melodiestücken. Eine gleiche Tonfolge kann in mehreren verschiedenen harmonischen Kontexten auftreten und nimmt dabei unterschiedliche Inhaltsformen an. Es ist wichtig zu bemerken, dass auch einfache Volkslieder, die häufig als monophone Melodien überliefert werden, über einen impliziten harmonischen Kontext verfügen.

Zuletzt werden die Patterns in ihrem musikalischen Kontext zu abstrakteren *Strukturen* zusammengefasst. Auf erster Ebene repräsentieren die Strukturknoten die zeitliche Abfolge von GPM-Pats. Die Strukturierung kann rekursiv auch auf höheren Ebenen erfolgen. Dabei können die Beziehungen (Kanten) durch zusätzliche Informationen über die Semantik – also die musikalische Bedeutung – der Unterelemente ergänzt werden. Mögliche Werte für die Bedeutung sind dabei full entity (Vollständiges Musikstück), independent unit (Sätze), structural units (Einleitung) und andere (vgl. [Seif2008]).

Durch die Ergänzung der Strukturen um Informationen zur Semantik können erkannte Abschnitte in einer Instanz in ihrer Ähnlichkeit zu einem Original-Template bewertet werden. Es können Ähnlichkeitsmaße definiert werden, die z. B. das Fehlen einer strukturellen Einheit höher bewerten als das Fehlen eines einzelnen Motivs.

3 Erkennung von Musikstücken mittels Hypothesenbildung

Um ein Musikstück zu identifizieren, muss es dem passenden Template zugeordnet werden. Dieser Zuordnungsprozess kann in drei Schritte gegliedert werden: (1) Erkennung aller GPM-Pats, die im Musikstück vorkommen, (2) Finden und Abstrahieren

passender Strukturknoten und (3) Auflösen von Konflikten bzw. Mehrdeutigkeiten. Im Folgenden werden die einzelnen Schritte kurz beleuchtet.

In der ersten Phase der Analyse müssen die elementaren musikalischen Ereignisse identifiziert werden, die in unserem Modell als GPM-Pats bezeichnet werden. Dabei spielen perzeptuelle Gesetzmäßigkeiten, wie das Gesetz der Nähe, eine wichtige Rolle. So kann bei einer Suche der Zeitraum für mögliche Ereignisse auf wenige Sekunden beschränkt werden. Wie Abschnitt 4 zeigt, führt diese Suche zu einer sehr großen Anzahl an Ergebnissen. Leider ist ein Ausschluss von Ereignissen als „nicht wahrgenommen“ nicht ohne Weiteres möglich, da z. B. bekannte Motive in einem musikalischen Zusammenhang auch dann wiedererkannt werden, wenn sie nur an unscheinbarer Stelle auftreten.

Ausgehend von einer Menge gefundener Patterns wird im nächsten Schritt versucht, unter Zuhilfenahme der Informationen aus den Templates, semantisch höherwertige Strukturen zu identifizieren. Auch hier können zeitliche Begrenzungen eingeführt werden, etwa eine Beschränkung des Suchraums auf 30 Sekunden. Sollten in dieser Phase Mehrdeutigkeiten auftreten, muss versucht werden, diese durch Informationen aus höheren Abstraktionsebenen aufzulösen. Grundregel ist dabei, dass die Auflösung zu einem Strukturknoten höherer Ebene der Auflösung zu einem Knoten niedrigerer Ebene vorzuziehen ist. Das Bilden von Strukturknoten führt nicht in allen Fällen zu (den richtigen) Ergebnissen. Gründe dafür können sein:

- Variationen musikalischer Patterns, z. B. durch Fehler oder Improvisationen,
- Quasi zufälliges Auftreten nicht-zusammenhängender Motive oder auch Strukturknoten niedriger Ebenen oder
- Variationen auf Ebene des Strukturknotens, etwas das Wiederholen eines Refrains.

Um den oben genannten Problemen zu begegnen verwenden wir einen *hypothesen-basierten* Ansatz. Unter Zuhilfenahme des Wissens über die bereits identifizierten Elemente werden Hypothesen aufgestellt, um welches Stück bzw. um welchen Abschnitt eines Stückes es sich handeln könnte (Bottom-Up). Ausgehend von dieser Information, kann darauf geschlossen werden, welche Patterns (oder größere strukturelle Einheiten) an bestimmten Positionen auftreten könnten (Top-Down). Es wird dann versucht, diese Hypothesen zu validieren. Abbildung 3, in der eine Variation des Volksliedes „Alle meine Entchen“ Verwendung findet, verdeutlicht diesen Vorgang. In diesem Beispiel wurden die Motive m_1 und m_2 identifiziert und dem Strukturknoten s_1 zugewiesen. Auf Grund des Wissens über den Knoten s_1 (besteht aus der Sequenz m_1 m_2 m_2) kann dann die Schlussfolgerung gezogen werden, dass die beiden nicht-identifizierten Takte eine Variante des Motivs m_2 sein könnten.

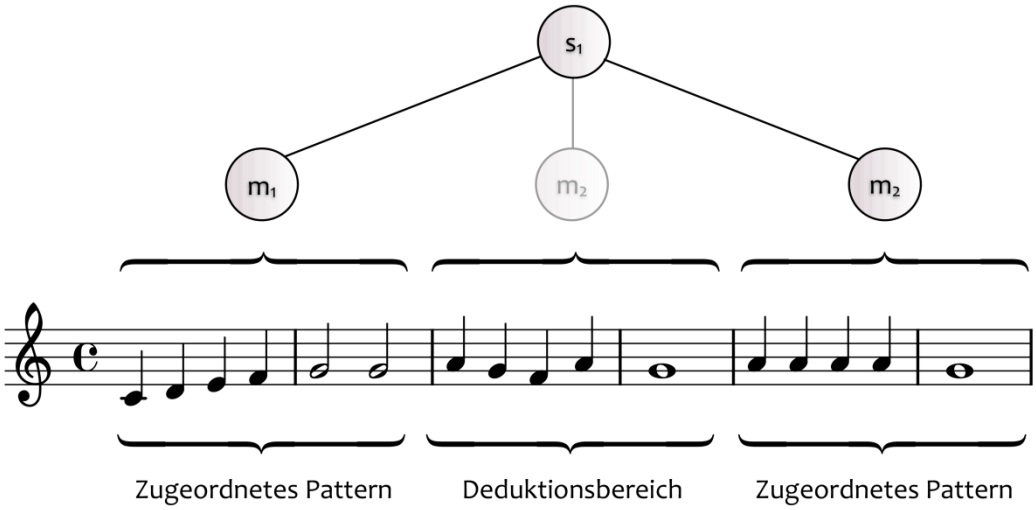


Abbildung 3: Deduktion bei der Musikerkennung

Wie die vorangestellte Abbildung zeigt, muss bei der *Validierung* der Hypothesen berücksichtigt werden, dass die zu suchenden Muster nicht mehr unverändert gefunden werden können. Deshalb findet an dieser Stelle eine Suche unter Einsatz von Ähnlichkeitsmaßen statt. Dabei werden besonders die musikalischen Merkmale Melodie, Rhythmus und Harmonie betrachtet.

4 Implementierung und praktische Probleme

Zur Validierung der Anwendbarkeit des vorgeschlagenen Modells zur inhaltsorientierten Repräsentation von Musik und zum Test der entwickelten Algorithmen wurden verschiedene Implementierungen durchgeführt. Alle Umsetzungen arbeiten dabei vorerst auf symbolischen Daten (MIDI), so dass musikalische Informationen explizit zur Verfügung stehen. Die Extraktion ebensolcher Informationen aus Audiodateien ist Gegenstand anderer Forschungen und soll an dieser Stelle nicht weiter erläutert werden. Zurzeit sind unsere Implementierungen in der Lage, melodiehafte GPM-Pats in Musikstücken zu suchen, wobei im Moment noch auf die Verwendung von Ähnlichkeitsmaßen verzichtet wird und solche Patterns nur unverändert wiedergefunden werden können.

Als Testdaten stehen uns die manuell erstellten – die computergestützte Extraktion von Motiven aus vorgegebenen Musikstimmen, als *Automatische Melodiesegmentierung* bezeichnet, liefert zurzeit noch nicht ausreichend befriedigende Ergebnisse [Ren2008, Orio2005] – Templates von 135 Stücken aus den Bereichen Volkslied, Jazz, Pop/Rock und Klassik zur Verfügung. Ihnen stehen ca. 1300 verschiedene Instanzen dieser Stücke gegenüber, die aus diversen Sammlungen von MIDI-Dateien gewonnen wurden

und deren Vielfalt von einfachen Variationen über die Originalstücke bis hin zu komplexen Arrangements in neuartigen Besetzungen reicht.

Ein erstes Problem bei der Erkennung von GPM-Pats stellt die exponentiell hohe Anzahl zu untersuchender Tonfolgen dar. Besonders bei komplexen, polyphonen Stücken kann eine vollständige Suche aller definierten Muster unverhältnismäßig viel Zeit in Anspruch nehmen. Diesem Umstand begegnen wir auf technischer Ebene, indem wir zuerst nach größeren Einheiten, die dann aus mehreren Patterns bestehen und charakteristischer sind, suchen. Außerdem verwenden wir ausgereifte Indexstrukturen und eine Beschränkung auf prinzipiell passende Melodieabschnitte um die Suche nach geeigneten Tonfolgen in linearer Zeit durchführen zu können.

In engem Zusammenhang mit der Anzahl zu untersuchender Tonfolgen steht die Größe der Ergebnismenge. Selbst bei relativ einfachen polyphonen Musikstücken werden viele Patterns (wieder-) gefunden, die perzeptuell nicht relevant sind, weil sie z. B. in einer Unter- oder Mittelstimme auftauchen, von anderen Tönen in räumlicher Nähe unterbrochen werden oder neben anderen, lauteren oder höheren Tönen nicht wahrgenommen werden. Abbildung 4 zeigt eine Analyse des Jazz-Standards „All of me“, die das Problem verdeutlicht. Die gefundenen musikalischen Motive sind dabei als Linien dargestellt.

Eine Analyse des 1. Klavierkonzerts von Sergei Rachmaninow in einer Bearbeitung für vier Hände liefert sogar 1.551.158 Ergebnisse.

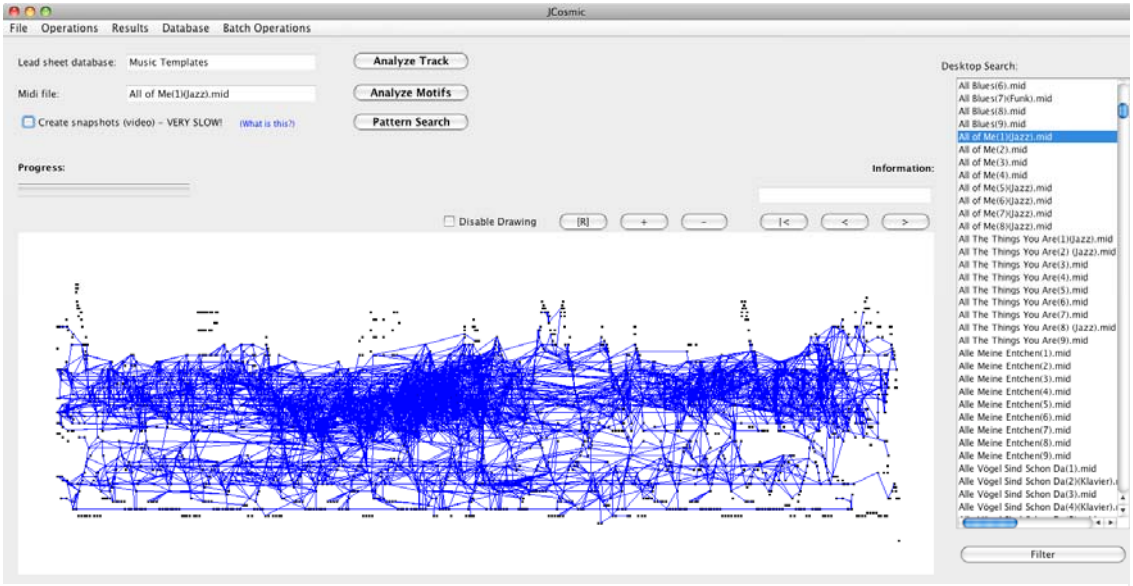


Abbildung 4: Analyse von „All of me“

Da die gefundenen Muster als Grundlage für die Hypothesenbildung im nächsten Schritt dienen, müssen folgerichtig auch hier sehr große Mengen bearbeitet werden. Zurzeit untersuchen wir deshalb, wie die Menge generierter Hypothesen sinnvoll

eingeschränkt werden kann, ohne „echte“ Ergebnisse zu verwerfen. Zusätzlich werden weitere technische Hilfsmittel ausprobiert, um die großen Datenmengen in akzeptabler Zeit verarbeiten zu können.

5 Vorläufige Ergebnisse

Zur Vorstellung konkreter Ergebnisse haben wir einen Teil unseres Bestands an Testdaten durch unsere Referenzimplementierung JCoSMIc (Java Comparing of Symbolic Musical Instances) analysieren lassen. Dazu haben wir 102 MIDI-Dateien ausgewählt, die einen Querschnitt durch unseren Datenbestand repräsentieren. Alle Dateien hatten freie Varianten von Musikstücken zum Inhalt, die unserer Applikation als Templates vorlagen.

Die Ergebnisse der Bearbeitung durch unser Referenzprogramm wurden analysiert und in folgende Kategorien eingeordnet: Template („Originalstück“) vollständig erkannt, Ausschnitte des Templates wurden erkannt und Template wurde nicht wiedergefunden. Tabelle 1 zeigt die Ergebnisse der Analyse.

Tabelle 1: Gruppierung der Testergebnisse

Anzahl Musikdokumente	102	100 %
Vollständig erkannte Templates	60	59 %
Teilweise erkannte Templates	92	90 %
Nichterkannte Templates	10	10 %

Im nächsten Schritt wurden die Dateien aus den einzelnen Gruppen noch einmal näher beleuchtet, um Aussagen über die Ergebniskategorien treffen zu können. Dabei wurden folgende Zusammenhänge deutlich:

- Musikstücke, in denen die „Originalmelodie“ zumindest in Ausschnitten unverändert und in annähernd korrektem Tempo auftritt, werden vollständig wiedererkannt.
- Musikstücke, in denen die Melodie nur in variierter Form auftritt, können noch nicht erkannt werden.
- Kommen leichte Strukturveränderungen, z. B. das Wiederholen eines Motivs, in den Dokumenten vor, so werden diese Veränderungen beim Zusammenfassen von Strukturelementen vernachlässigt, und das entsprechende Musikstück wird erkannt.
- Musikstücke mit gewichtigeren Strukturveränderungen, etwa dem Weglassen eines Abschnitts, können nicht erkannt werden.

Die Ergebnisse spiegeln den Entwicklungsstand der Implementierungen sehr gut wider. Gleichzeitig geben sie einen Hinweis auf die in den nächsten Schritten durchzuführen-

den Erweiterungen der Anwendungen, die im nächsten Abschnitt erläutert werden sollen.

6 Zusammenfassung und Ausblick

In Kapitel 2 haben wir ein inhaltsorientiertes Modell zur Beschreibung von Musikdokumenten beschrieben, das im Rahmen unseres Projekts entstanden ist. Einer der Vorteile dieses Modells ist die Möglichkeit, Hypothesen zur Unterstützung bei der Erkennung von Musikstücken aufzustellen, z. B. wenn einige Teile des Stückes bereits erfolgreich erkannt wurden, während andere Teile so stark variiert wurden, dass eine Wiedererkennung nicht ohne weiteres möglich ist.

Die Validierung solcher aufgestellter Hypothesen – von uns als Top-Down-Hypothesen bezeichnet – ist mit unseren prototypischen Implementierungen bisher nicht möglich. Sie setzt den Einsatz von Ähnlichkeitsmaßen beim Finden von GPM-Patterns voraus. Ausgehend von den grundlegen Elementen von Musik müssen diese Distanzmaße mindestens folgende Eigenschaften berücksichtigen:

- Melodie
- Rhythmus
- Harmonie
- Tempo
- Metrum.

Es existiert bereits eine Vielzahl unterschiedlicher Ansätze und Untersuchungen für bzw. über solche Ähnlichkeitsmaße für musikalische Muster, u. a. [Craw1998, Anag2000, Down1999, Camb2000, Lems2000, Engl2001].

Eine der nächsten Schritte wird es sein, diese existierenden Ansätze zu testen und zu bewerten, inwieweit sie für den Einsatz im Rahmen unserer Anwendungen geeignet sind. Ebenfalls ist zu überprüfen, ob eine Kombination verschiedener Algorithmen zu besseren Ergebnissen führt. Ausgehend von den Resultaten unserer Evaluierungen müssen ein oder mehrere Ähnlichkeitsmaße implementiert werden, um bei der Hypothesenvalidierung zum Einsatz zu kommen.

Auch die Suche nach Strukturelementen über identifizierten Motiven bzw. die rekursive Suche über anderen Strukturknoten muss robust gegen gewisse Variationen, z. B. das Wiederholen eines Motivs, sein. Bisher spiegelt sich dies in unseren Implementierungen nicht ausreichend wider. Treten solche Variationen auf, müssen sie zum Einen entdeckt und zum Anderen entsprechend ihrer Einwirkung bewertet werden. So kann das Wiederholen einer Einleitung zu einem Stück nicht mit dem Fehlen eines Refrains gleichgesetzt werden, obwohl mathematisch gesehen beide Fälle nur eine Änderungsoperation auf dem entsprechenden Graphen darstellen.

Die Ergebnisse, die in diesem Beitrag präsentiert wurden, beziehen sich nur auf einen Ausschnitt der uns zur Verfügung stehenden Testmenge. Diese Einschränkung war nötig, um relevante Aussagen über die durch unsere Anwendungen erzielten Resultate – abseits der Bestimmung statistischer Werte, wie True Positives usw. – treffen zu können. Sobald unsere Implementierungen die Möglichkeiten, die das zu Grunde liegende Modell bietet, voll ausschöpfen, müssen Tests unter Verwendung der gesamten Dokumentmenge durchgeführt werden. Dabei ist auch die Skalierbarkeit der Anwendungen zu prüfen, da bei digitalen Archiven das Vorhandensein sehr großer Datenmengen anzunehmen ist. Natürlich muss beachtet werden, dass der Fokus unserer Prototypen zurzeit noch nicht primär auf dem Aspekt der Effizienz liegt.

Mit unseren Prototypimplementierungen, die das entwickelte hierarchische Modell zur Modellierung von Musikdokumenten noch nicht vollständig umsetzen, konnten bereits vielversprechende Ergebnisse erzielt werden. Wir gehen davon aus, durch das Umsetzen eines hypothesenbasierten Erkennungsprozesses, der auf gut ausgewählte Ähnlichkeitsmaße zurückgreift, diese Ergebnisse noch verbessern zu können. Dieser Ansatz sollte eine inhaltsorientierte Identifikation von Musikstücken in digitaler Form ermöglichen, die gleichzeitig robust gegen verschiedene Arten von Variationen ist.

7 Literaturverzeichnis

[Anag2000] Anagnostopoulou, C., Hörnel, D., Höthker, K. Investigating the Influence of Representations and Algorithms in Music Classification. In: Computers and the Humanities, Vol. 35, S. 65-79, 2000.

[Balu2007] Baluja, S., Covell, M. Audio fingerprinting: Combining computer vision & data stream processing. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), S. 213-216, 2007.

[Camb2000] Cambouropoulos, E. Melodic Cue Abstraction, Similarity and Category Formation: A Computational Approach. In: Music Perception, Vol. 18, S. 347-370, 2000.

[Cano2002] Cano, P., Batlle, E., Kalker, T., Hatsma, J. A Review of Algorithms for Audio Fingerprinting. In: Workshop on Multimedia Signal Processing, S. 169-173, 2002.

[Craw1998] Crawford, T., Iliopoulos, C. S. und Raman, R. String Matching Techniques for Musical Similarity and Melodic Recognition. In: Computing in Musicology, Vol. 11, S. 73-100, 1998.

[Dann2003] Dannenberg, R. B., Birmingham, W. P., Tzanetakis, G., Meek, C., Hu, N., Pardo, B. The MUSART Testbed for Query-by-Humming Evaluation. In: Proceedings of the 4th International Symposium on Music Information Retrieval, S. 34-48, Baltimore, 2003.

[Dann2004] Dannenberg, R. B., Hu, N. Understanding search performance in query-by-humming systems. In: Proceedings of the 5th International Sym-

posium on Music Information Retrieval, S. 232-237, Barcelona, 2004.

[DMR2009] International Federation of the Phonographic Industry (IFPI). Digital Music Report 2009: New Business Models for a Changing Environment. IFPI, 2009.

[Down1999] Downie, J. S. Evaluating a simple approach to music information retrieval: conceiving melodic n-grams as text. London, Ont.: Faculty of Graduate Studies, University of Western Ontario, 1999.

[Down2003] Downie, J. S. Music information retrieval (Chapter 7). In: Annual Review of Information Science and Technology 37, ed. Blaise Cronin, S. 295-340. Medford, NJ: Information Today, 2003

[Engl2001] Hofmann-Engl, L. Towards a Cognitive Model of Melodic Similarity. In: Proceedings of the 2nd Annual International Symposium on Music Information Retrieval, S. 143-151, 2001.

[Lee2004] Lee, J. H., Downie, J. S. Survey Of Music Information Needs, Uses, And Seeking Behaviours: Preliminary Findings. In: Proceedings of the Fifth International Conference on Music Information Retrieval: ISMIR 2004, S. 441-446, Barcelona, 2004.

[Lems2000] Lemström, K. String Matching Techniques for Music Retrieval. Helsinki, Faculty of Science of the University of Helsinki, 2000.

[Mitro2006] Mitrovic, D., Eidenberger, H. Analysis of the Data Quality of Audio Features of Environmental Sounds. In: Journal of Universal Knowledge Management (JUKM), S. 4-17, 2006.

[Orio2005] Orio, N., Neve, G. Experiments on Segmentation Techniques for Music Documents Indexing. In Proceedings of the 6th International Conference on Music Information Retrieval, Proceedings of the International Symposium on Music International Retrieval, London, 2005.

[Ren2008] Rentzsch, M., Seifert, F., Hornfischer, C. und Schreiber, A. Melodic Segmentation on different Musical Genres. In: Proceedings of the 4th IEEE International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution, S. 3-9, Florence (Italien), 2008.

[Seif2008] Seifert, F., Rentzsch, M. Generic Music Identification by Hierarchic Modeling of Human Perception. In: Proceedings of the 4th IEEE International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution, S. 10-16, Florence (It), 2008.

[Wang2008] Wang, L., Huang, S., Hu, S., Liang, J., Xu, B.: An effective and efficient method for query by humming system based on multi-similarity measurement fusion. In: International Conference on Audio, Language and Image Processing, S. 471-475, Peking, 2008.

9 783000 278587
ISBN 9-78300-278587
ISSN 0947-5125 · Chemnitzer Informatik Berichte CSR-09-04

KMD

Bundesministerium für Bildung und Forschung

INNOPROFILE UNTERNEHMEN REGION Die BMBF-Innovationsinitiative Neue Länder

Informatik

Projektpartner

Arbeitsgemeinschaft Regionalfernsehveranstalter in Sachsen (ARiS) envia TEL GmbH *(Markkleeberg)* · HMS oHG *(Halle/Saale)* · Mugler AG *(Oberlungwitz)*

TU Chemnitz · Fakultät für Informatik · Strasse der Nationen 62 · 09107 Chemnitz