GMD Report 121

GMD –
Forschungszentrum
Informationstechnik
GmbH

Frank Nack, Wolfgang Putz

Semi-automated Annotation of Audio-Visual Media in News

December 2000

GMD –
Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
Germany
Telefon +49 -2241 -14 -0
Telefax +49 -2241 -14 -2618
http://www.gmd.de

Anschrift der Verfasser/Address of the authors:
Dr. Frank Nack
Wolfgang Putz
Institut für Integrierte Publikations- und Informationssysteme
GMD – Forschungszentrum Informationstechnik GmbH
Dolivostraße 15
D-64201 Darmstadtt
E-Mail: Frank.Nack@gmd.de
Wolfgang.Putz@gmd.de

ISSN 1435-2702

Abstract
This report considers the automated and semi-automated annotation of audiovisual media in a new type of production framework, A4SM (**A**uthoring **S**ystem for **S**yntactic, **S**emantic and **S**emiotic **M**odelling). We present the architecture of the framework, describe a prototypical camera, a handheld device for basic semantic annotation, and an editing suite to demonstrate how video material can be annotated in real time and how this information can not only be used for retrieval but also can be used during the different phases of the production process itself. We then outline the underlying MPEG-7 based content description structures of A4SM and discuss the pros and cons of our approach of evolving semantic networks as the basis for audio-visual content description.

Keywords: Media production, media annotation, news production, MPEG-7, news production tools

Kurzfassung
Dieser Bericht beschreibt die automatische und halbautomatische Annotation von audio-visuellen Medien in einer neuartigen Produktionsumgebung, A4SM (**A**uthoring **S**ystem for **S**yntactic, **S**emantic and **S**emiotic **M**odelling). Wir stellen die Architektur des Systems vor, beschreiben eine prototypische Kamera, ein Handgerät für elementare semantische Annotation und eine Edier-Umgebung, um zu zeigen, wie Videomaterial in Realzeit annotiert werden kann, und wie die Annotationen nicht nur für das Informationsretrieval, sondern auch während der verschiedenen Phasen des Produktionsprozesses selbst genutzt werden können. Anschliessend werden die zugrundliegenden MPEG-7 basierten Inhaltsbeschreibungsstrukturen von A4SM beschrieben und die Vor- und Nachteile unseres Ansatzes der dynamischen semantischen Netzwerke als Basis für audio-visuelle Inhaltsbeschreibung diskutiert.

Schlagworte: Medienproduktion, Medienannotation, Nachrichtenproduktion,, MPEG-7, Werkzeuge zur Nachrichtenproduktion.

Table of Contents

1 Introduction

Since the beginning 1980' we have discovered a radical technological revolution that allowed the digitalisation of existing audio-visual media. Due to swift developments in hardware technology (e.g. networkable and storage intensive computers, CD-ROMs, DVD, video and camcorders, IP-telephony, Webcams, synthesisers, MIDI, DAW, etc.) the digitisation of the media domain has been proceeding rapidly with respect to storage, reproduction, and transportation of information. At the same time the notion of the 'digital' as the capability to combine atomic information fragments became apparent. The idea of 'semantic and semiotic productivity' allowing an endless montage of signs inspired a great deal of research in computer environments that embody mechanisms to interpret, manipulate or generate visual media [Bloch 1986, Parkes 1989, Aguierre-Smith & Davenport 1992, Nagasaka & Tanaka 1992, Sack 1993, Chakravarthy 1994, Zhang & Smoliar 1994, Tonomura et. al 1994, Gordon & Domeshek 1995, Davis 1995, Yeung et. al 1995, Nack 1996, Brooks 1999, Lindley 2000]. Similar developments were acquired for audible information [Bloom 1985, Hirata 1995, Pfeifer et. al 1996, Wold et al. 1996, Robertson et. al. 1998, Borchers & Mühlhäuser 1998, TALC 1999].

The steady infiltration of the gained results into everyday production environments, such as the non-linear video editing systems FAST 601, Softimage Ds, or audio systems such as Cubase VST, have already deeply changed the social way of exchanging information. More and more people are acquainted with the creative process of producing and receiving audio-visual information and, as being exemplified by the popularity of the internet, make use of their skills. However, like in traditional written communication the means of montage are merely used during the production process while the final product is still understood in a context restricted way. This is an instance of Marshall McLuhan's observation that new media technology is used initially to solve old problems. The deeper impact of digital media is to redefine the forms of media, blurring the boundaries between traditional categories like pre-production, production, and post production, and radically altering the structure of information flow from producers to consumers. It is possible, for example, to generate customised news programs, using a high level model of overall program structure, but selecting content from news video databases according to the particular interests and needs of a viewer. The source databases, belonging to different agencies, may be distributed within an E-commerce infrastructure, and material may remain for reuse in contexts requiring background information or for historical exploration and analysis. New material may be gathered for a particular production, with the secondary aim of reuse for other productions, or may be gathered as raw source material for broad dissemination to a variety of end viewer presentations. Systems for selecting and presenting media content may use mechanisms to dynamically composite video images, including the incorporation of animated characters and objects. Raw video may also be processed to extract its data content in more useable forms, for example, mosaicing sequences to extract their backgrounds and separating their foreground object sequences for reuse in different contexts.

Within this overall media infrastructure, a given component of media exists independently of its use in any given production. Systems are therefore necessary to manage such media objects and representations of their semantics for use in many different productions with a potentially wide range of forms, such as search or filtering of information, media understanding (surveillance, intelligent vision, smart cameras, etc.), or media conversion (speech to text, picture to speech, visual transcoding, etc.). Systems are also required for the authoring of representations creating a particular production, such as a narrative film, a documentary, or an interactive game. The molecular processing of fixated hypermedia information becomes difficult because the provided information is not only buried within the established relations between the single media units but – and that is even harder to tackle – because the main semiotic information is hidden in the unified structure of the single image, video, audio or tactile unit that results from the composition of all its elements.

The emerging cyberspace described above is best understood a semantic network based on the relationship between the signs of the audio-visual information unit and the idea it represents, according to the creator's intention, as well as the differing connotations that can be attributed to the signs, depending on the circumstances and abductive presuppositions of the receiver at the time of perception, along with the various legitimated codes and sub-codes the receiver uses as interpretational channels [Arnheim 1956, Peirce, 1960, Eco, 1977; 1985,Greimas 1983, Bordwell 1989]. Such an information space requires new work environments, which support the technical and creative aspects of media information units. Specifically the creative aspects of the production process are of particular interest, since they form the bases of the complex, resource strong activity, which develops and composes a multi-dimensional hierarchical clustered network of relationships between different kinds of information units. Thus, what

we really need are tools which allow people to use their creativity in a way they are already used to but in addition use the human activity to extract the significant syntactic, semantic and semiotic aspects of its content [Brachman & H. J. Levesque 1985] which then can be transformed into a description based on a formal description language.

Research in industry and academia have opened inroads to characterise audio-visual information not only on a conceptual level by using keywords, but also on a perceptual level by using objective measurements based on image or sound processing, pattern recognition, etc. [Aigrain et al. 1995, Gupta & Jain 1997, Del Bimbo 1999, Mills et al. 2000, Johnson et al. 2000, Melucci & Orio 2000, Lemström & Tarhio 2000]. However, the problem with these approaches is that they merely use low-level perceptual descriptors with limited semantics for content representation (for an approach over several semantic levels see among others [Colombo et al. 1999, Hunter 1999, Pachet & Cazaly 2000]). Moreover, the general problem with such retrospective-exclusive approaches is, that they use the final media product which does not provide important cognitive, content and context based information, such as editing lists with decision descriptions. This type of semantic information, which is required to enrich the automatically extracted information, is usually provided through manual annotation – an expensive endeavour normally not covered by the production or archival budget.

In this paper we present ongoing research of the Mobile Group at GMD-IPSI, which is concerned with the representation of media content during its production, allowing the reuse of the gathered structures, meta information and media units (from now on called data). Our work is directed towards the application of IT support for all stages of the media production process within the A4SM framework (A4SM is pronounced APHORISM — **A**uthoring **S**ystem for **S**yntactic, **S**emantic and **S**emiotic **M**odelling). The aim of A4SM is to design a distributed digital media production environment supporting the creation, manipulation, and archiving/retrieval of media material. The project goal is to suggest a framework for semi-automated annotation of audio-visual objects to establish a growing information space and to demonstrate and asses the applicability and acceptability of this framework in a news production environment. The interesting aspect of the news domain for our research is that here dynamic structures and implicit connections are required to establish statements, context and discourse. For example, a news item showing Bill Clinton and Monica Levinsky only becomes of interest after their relationship changed into the context of 'scandal'. As a result we are forced to reconstruct the relations between existing material on two levels. Firstly, on a physical basis, by integrating an existing piece of video into a newly created piece of newscast, using it perhaps as a temporal reference, and secondly on a representational basis, by modifying an existing descriptional unit with an additional relation. Thus news as a domain is interesting because the underlying ontological representations require to describe the physical world and abstract mental and cultural concepts, though only a shallow level will be sufficient for the 'micro world' of news. At the same time, the content representations must be related to the intrinsic structures of the media unit to be described, so that the translation between one representation and the other does not result in the loss of salient features of either representation.

In the following article we outline our model of the news production process, describe the A4SM architecture as well as a number of tools using the A4SM framework, and finally address the representational structures, which are based on MPEG-7's formal description language [MPEG Requirements Group 2000c] for the description of audio-visual content.

2 News production – one type of media production

Media production, such as for news, documentaries, soaps, feature films, interactive games, edutainment applications or virtual environments is a complex, resource demanding, and distributed process with the aim to provide interesting and relevant information by composing a multi-dimensional network of relationships between different kinds of audio-visual information units. Though the produced item, e.g. a news clip or a film, itself is representing linearity, the process of making is rather of a circular and organic nature, e.g. improvisatory methods are important. Hence, it is more likely that a holistic approach is applied in media production than the unidirectional movement of the manufacturing line. Nevertheless, for convenience reasons media production is traditionally arranged in three parts, i.e. *preproduction*, *production*, and *postproduction*. The activities associated with these phases may vary according to the type of production. For example, the methodology involved in the creation of dramatic film, documentary, and news programs is very different. Commercial dramatic film production is typically a highly planned and linear process, while documentary is much more iterative with story structure often being very vague until well into the editing process. News production is structured for very rapid assembly of material from very diverse sources. Media information systems must be able to accommodate all of these styles of production, providing a common framework for the storage of media content and assembly of presentations.

Within the news production process the different phases represent the following aspects. The preproduction phase covers the identification of events and the schedule planning. The production part includes shooting and transmission of the news feed to the studio. The postproduction is directed towards editorial decisions based on reviewing the material, editing, sound mixing, broadcasting, and archiving.

It must be emphasised that the interrelationships between different stages within the process (e.g. the influence of personal attributes on decisions or the comparing of different solutions) is complex and extremely collaborative. The nature of these decisions is important because as feedback is collected from many people, it is the progression through the various types of reviewers that effects the nature of the work.

Hence, each of the different phases of news production provides important information on a technical, structural, and a descriptional level. However, today's news production is mainly a one time application for design and production. This means that primarily oral discussions on the choice of events or paper based descriptions of activities on the set, production schedules, editing lists with decision descriptions, and organisational information will be lost after the production is finished. Ironically, it is this kind of cognitive content and context based information which today's researchers and archivists try to analyse and re-construct out of the final product - usually with limited success.

Moreover, the tendency of 'lost information' is supported by current professional IT tools. These applications assist in the processes of transforming ideas into scripts (e.g. a text editor, Dramatica), sustain in digital/analogue editing (Media 1oo, Media Composer, FAST 601, etc.), or support production management (Media-PPS, SAP R/2, SESAM, etc.). Most of the available tools are often based on incompatible and closed proprietary architectures. Hence, it is not possible to establish an automatic information flow between different tools, nor is it possible to support the information flow between distinct production phases. The result is that compatibility between different systems is either restricted (the different applications are provided by one company) or compatibility is not applicable at all.

However, due to the most critical constraints within news production, i.e. time, it is in particular the flow of information that is of paramount interest. As a result of a knowledge elicitation among tv-reporters, cameramen, and editors for news we identified that the optimisation of the workflow within a news department requires that incoming material, such as news feeds or transmissions from field reporters, should immediately be retrievable by all members of the editorial staff. To facilitate this, the incoming feed has to incorporate a certain set of descriptive metadata, such as location, title, topic, content description, camera work or duration. Current news feeds provide this information (at some point in time) via file transfer. An automatic link of this information to the video material itself is in most cases not possible because the delivery time of this information is usually unspecified.

To improve this situation we propose a news environment that enriches the content from an early production state with relevant metadata and carries this enriched material to the news room. The proposed environment consists of the following components:

- A hard disk camcorder that automatically stores the acquired video stream together with an associated stream of relevant metadata like co-ordinates, camera work or lens movement

- A speech, keyboard or pen-based annotation tool for the reporter to provide real-time annotation during acquisition and includes this information time-coded into the stream of metadata
- An on-site editing tool for the reporter that allows editing the material, provides means for stratified annotation on segment level and that incorporates the edit decision list and the annotations into the metadata stream
- A transmission of the metadata stream to a news provider or TV broadcaster
- An import and logging tool for the recipient that provides automatic recording, extraction of metadata, key frame extraction and generation of frame-accurate low-res proxies for browsing and rough cut
- A content management system that accepts all this information and allows immediate access to incoming material with only a minimum delay time
- A news ticker application that notifies the news room editors when new material is coming in and that allows to query the metadata, to preview material, to rough cut on the low-res proxy and to download the results of the rough cut process into professional editing suites as well as directly into the air play system

Based on the above discussion we will now introduce A4SM and show how it can support the news production model.

3 News production and A4SM

The aim of A4SM, as described in Figure 1, is to provide a distributed digital media production environment supporting the creation, manipulation, and archiving/retrieval of audio-visual material during and after its production. The environment is based on a digital library of consistent data structures, around which associated tools are grouped to support the distinct phases of the media production process. Each of the available tools forms an independent object within the framework and can be designed to assist the particular needs of a specialised user. It is essential for our developments, that the tools should not put extra workload on the user – she should concentrate on the creative aspects of the work in exactly the same way as she is used to. Nevertheless, due to the tool's affiliation to the A4SM environment it supports the required interoperability.
For the news environment as described in section 2 we identified two important phases of the framework that can be supported by IT, i.e. production and post-production (note: the pre-production phase of scripting is, due to the time constraints and the unpredictability of events, not applicable for news (see [Nack & Steinmetz, 1998] and [Nack & Lindley, 2000] for a description of how such a script editor should look like).

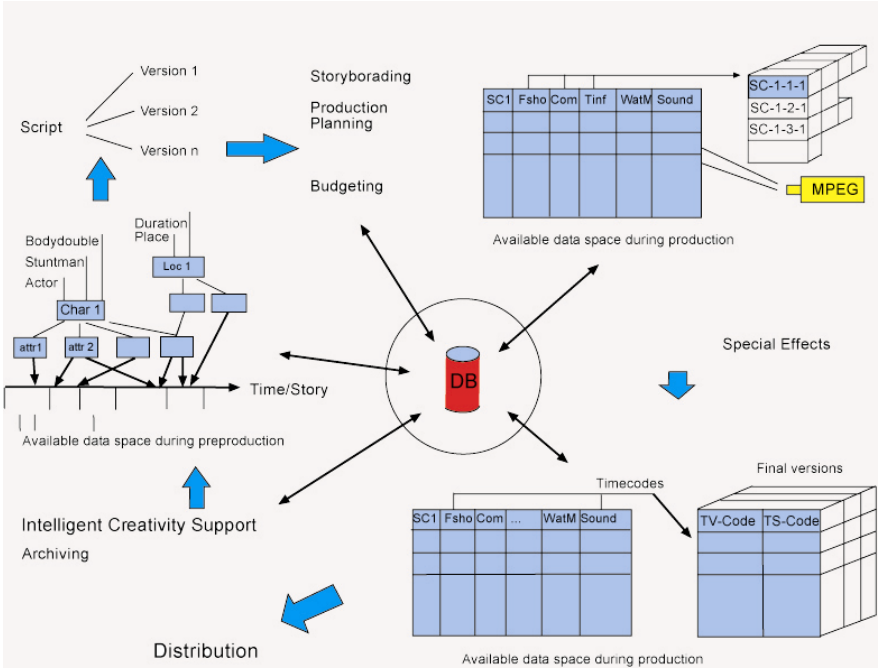


Figure 1: The A4SM architecture

Within the production phase it is the acquisition of material which can be improved by supporting the collaboration between reporter and cameraman. The common procedure for this process is that the general concept for the news-clip is designed on the way to the location of the event to be portrayed. Refinements of the concept might be performed at the location. Thus, there is a need for a set environment within which the reporter can annotate structural (e.g. scene id, etc.) and content information (e.g. importance of s shot with respect to audio or visual elements) while the cameraman is shooting. As a result we designed and developed a MPEG–7 hard disk camera (see Figure 2) that automatically stores the acquired video stream together with an associated MPEG-7 description structure of relevant information, such as co-ordinates, camera work or lens movement. Additionally we provide a mobile handheld annotation tool for the reporter (see Figure 2) to provide real-time annotation during acquisition on a basic semantic level, i.e. in and out points for sound and images.
The second important phase in news production to be supported by IT is the post-production, in which the recorded material is made ready for telecasting. Here it is the collaboration between reporter and editor which needs attention. During the knowledge elicitation it was mentioned by a great number of reporters, that it would be excellent to have a simple editing suite in form of a lab-top based application so that they do not have to rely on an editor and can increase the topicality of their work. Thus, we designed and prototyped an on-site editing suite for a reporter that allows editing of the material, provides means for

stratified annotation on segment level, and incorporates the edit decision list as well as the annotations into the MPEG–7 description structure.
Before we introduce the MPEG-7 based content description structures and the general concepts and assumptions of the repository we would like to describe the different tools in a bit more detail.

4 The Tools

4.1 The camera

The aim of the Digital Camera is to provide, besides recorded MPEG-1 or MPEG-2 video, metadata associated with the video. The metadata describes image capture parameters, such as lens movement, lens state, camera distance, camera position, camera angle, shot colour, etc, as it is this technology which manifests itself in the medium's unique expressiveness [Bordwell 1989, Eco 1977, 1985]. Figure 3 shows the currently collected information, i.e. data about camera movement (pan and tilt), lens action (zoom and focus), shutter, gain, and iris position. Moreover, we do collect information about the spatial position of the camera. For the latter we use a magnetic tracker. The global position of the camera is part of future developments.

The picked information units were chosen to demonstrate the general access and archival mechanisms. A complete set of camera descriptors might be closer to the parameters suggested by [SMPTE 1999] and by the GMD virtual studio [Fehlis 1999].

Figure 2: Design study of the MPEG-7 camera and MPEG-7 Handheld device

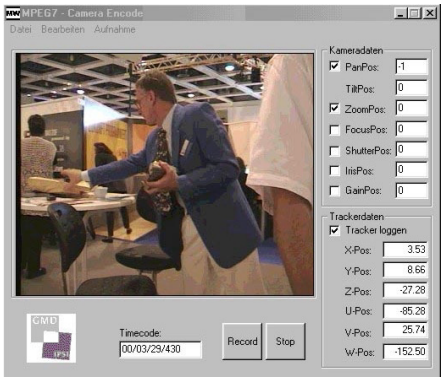


Figure 3: Interface for the camera handling

The actual hardware components used in our demonstration environment are:

- Polihemus FastScan Tracker,
- Sony EVID-30 Camera,
- a Videodisc Photon MPEG-1 Encoder.

The synchronisation (i.e. linking) between data and annotation is resolved via timecodes in SMPTE notation (hh:mm:ss:msms) combined with a scene identifier. Before shooting, camera and handheld (see section 4.2) exchange the scene id. While shooting the annotation process polls every 20 ms for the required camera changes and if modifications are identified the relevant Description Schemata (i.e. XML-Schema documents), for storing the collected information, are created. In such a way we establish a document network, where each change will be represented in a single document (temporal actions longer than 20ms, such as zooms, are collected in one document only). The detailed structures, their generation and use is described later in section 5.

The only extra work for the cameraman is to establish the required set of description schemata. Due to the particular requirements of productions it is important to provide a larger (e.g. for feature films) or smaller (e.g. for news) set of automated annotations. This sort of 'user profile' is stored on a programmable code-card, which is installed into the camera before shooting starts.

4.2 The handheld

Figure 4 demonstrates the interface simulation within our demonstration environment. The device provides a monitor for screening the recorded material of the camera in real-time (the transition rate is 13 images per second, since the reporter is only interested in gaining an idea of the framing and content). Furthermore, the handheld supplies a set of buttons

- to mark the importance of sound and images with in- and out points for sound and image and

- conceptual shot dependency via a scene id, such as 1-1,1-2,2-1, etc., where the first digit represents the scene and the second digit stands for the shot number.

Figure 4: Simulation of a handheld device for the annotation of simple semantic information

The scene id can be adjusted by the reporter at any time if the camera is not recording. Once the camera is active, it first identifies the current id setting of the handheld, which will provide the id for all description schemata created for the current recording. The synchronisation of the in and out points, which can be set by the reporter at any time during recording, with the audio or video is achieved via timecodes and the scene id. The generation of the DSs is the same as outlined for the camera info.
For the reporter this means, that she has to do exactly the same information gathering in our production environment as she is used to in current environments. The difference is, that now the synchronisation between camera and important conceptual information, such as what is relevant information and at what time did it happen, is not anymore based on the adjustment of camera time code and reporter watch but rather automatically. Furthermore, the reporter can now concentrate on the action, since only a few buttons need to be pressed, instead of scribbling timecodes and related nodes on paper. The DSs produced by the handheld are described in more detail in section 5.

4.3 The editing suite

This tools uses the annotations of the audio-visual media to automatically group the material based on conceptual dependencies within news. Rules for grouping material cover such aspects as

- Groups of video material are based on scene ids and their versions
- short clips are more important than long ones
- annotations of in and out points increase the value of a shot, thus it should be presented more prominently, etc.

As a result the reporter gets an instant overview of the available material and its intrinsic relations represented through the spatial order of its presentation (see Figure 5a as an example). The reporter is now able to mark the relevant video clips for the newscast by pointing at them. The order of pointing indicates the sequential appearance of the clips (see Figure 5b). Finally the reporter provides the overall time of the clip (see Figure 5c). Based on a simple planner the editing suite is then performing an automated composition of the news clip, which the reporter can tune interactively, e.g. to adjust it according to his voice over (see the time line at the top of the interface, Figure 5d).

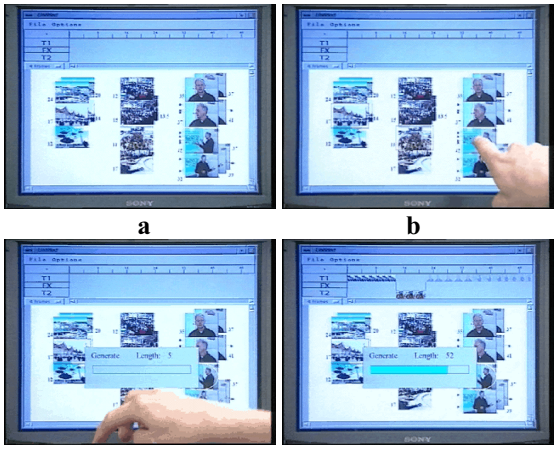


Figure 5 (a – d): The handling of a semi-automated MPEG-7 editing suite

Rules for the automated clipping of a shot might look like this (all rules are concerned with the rhythmical shape of a sequence) :
Strategy 31 and Strategy 34 reflect the fact that a viewer perceives the image in a close-up shot in a relatively short time (≈ 2-3 seconds), whereas the full perception of a long shot requires more time.[1] Moreover, the composition of shots may vary in number of subjects, number and speed of actions, and so on, which also influences the time taken to perceive the image in its entirety. Finally, the stage of a sequence in which a shot features also influences the time taken to perceive the entire image. For example, a long shot used in the motivation phase takes longer to appreciate, since the location and subjects need to be recognised, whereas in the resolution phase the same shot type may be shorter in duration, since in this case the viewer can orient himself or herself much more quickly.

E-Strategy 31 *If camera distance of a shot = close-up*
then
clip it to a length = 60 Frames.

E-Strategy 34 *If camera distance of a shot > medium-long sequence. kind = realisation or resolution*
then
clip it to a length = 136 Frames.

Strategies such as E-strategies 31 - 34 indicate the need to trim a shot. However, the above strategies may cause problems, as the actions portrayed may simply require more time than suggested by the rules. The necessary steps to be performed to achieve the keeping a necessary chain of actions while still allowing to trim are to identify the startframe of the first relevant action, and to detect any overlap with the second relevant action. If such an overlap exists, then it is possible to cut away the section of the shot in which the first action is performed in isolation. The same mechanism applies to the end of the shot. It is, of course, important that the established spatial and temporal continuity between the shot and its predecessor and successor are still valid. Figure 6 describes the application of edge trimming for a shot of 140 frames. The shot should portray an actor walking and then sitting, but should not be longer than 108 frames.

[1] The time values used in all the following examples of editing strategies are based on estimates provided by the editors at the WDR.

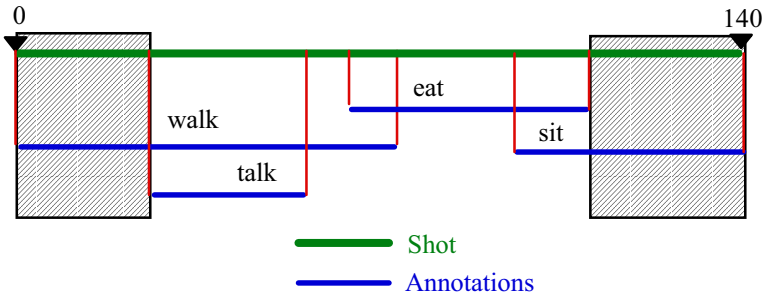


Figure 6: Trimming of a shot from 140 to 108 frames

E-Strategies 36 represent an example of temporal clipping as discussed immediately above, focussing on frame elimination from a single shot that sequentially portrays a number of actions.

E-Strategy 36 *If close-up < camera distance < long and*
sequence.action.tempform = equivalence and
number of frames > as calculated in E-Strategy 32 or 33 and
number of performed actions in the shot > 3
then
verify the frame overlap of first and second action as X
verify the overlap of last and last - 1 action as Y
cut away the pure frame numbers for the first action if X > 36
cut away the pure frame numbers for the last action if Y> 36

For a more detailed theoretical discussion of automated film editing see [Nack and Parkes 1997].
At the current stage of development our prototypical editing environment uses the acquired meta-information from the annotation process (e.g. handheld in or out points, camera position, lens position, etc.), to support the editing process with advanced information structure and presentation functionality (e.g. search, order, and approximation to the relation between text and video material). Figure 7 displays parts of the search engine of our test environment, which allows to search for annotation types and directly locate them in the video.

Figure 7: Search engine of the A4SM test environment

The option of producing a rough cut versions automatically is part of our ongoing development.
Having introduced the tools in more detail, we are now in the position to describe the general concepts and assumptions of the repository and some of the MPEG-7 based content description structures underlying the just described tools.

5 A4SM repository

Looking at the phases of news production it becomes apparent that the audio-visual material undergoes constant changes, e.g. from the shooting to editing, where parts of the material usually will become reshaped on a temporal and spatial basis. This dynamic use of the material has a strong influence on the descriptions and annotations of the media data created during the production process. That is, the annotations will have gaps, overlaps, double- or triple annotations, etc., or in other words, the annotations will be incomplete and change over time.

As a result it is important to provide semantic, episodic, and technical memory structures with the capability to change and grow. This requires relations between the different type of structures with a flexible and dynamic ability for combination. To achieve this, media annotations cannot form a monolithic document but must rather be organised as a network of specialised content description documents.

The representation of our description schemata is based on the MPEG-7 standardisation effort [MPEG Requirements Group 1999a, 1999b, 2000c]. The objective of the MPEG-7 group is to standardise ways how to describe different types of multimedia information. The emphasis is audio-visual content with the goal to extend the limited capabilities of proprietary solutions in identifying content by providing a set of description schemes (DS) and descriptors (D) for the purpose to make various types of multimedia content accessible. In this context a DS specifies the structure and semantics of the relationships between its components, which may be both descriptors and description schemes. A descriptor defines the syntax and the semantics of a distinctive characteristic of the media unit to be described, e.g. the colour of an image, pitch of a speech segment, rhythm of an audio segment, camera motion in a video, style of a video, the actors in a movie, etc. Descriptors and description schemes are represented in the MPEG-7 Description Definition Language (DDL). The current version of the DDL is XML Schema [XML Schema Part 0, Part 1, Part 2] based providing means to describe temporal and spatial features of audio-visual media as well as to connect these descriptions on a temporal spatial basis within the media. For more details on MPEG-7 see [Nack & Lindsay 1999].

To facilitate the dynamic use of audio-visual material, A4SM's general attempt at content description applies the strata oriented approach [Aguierre-Smith & Davenport 1992] in combination with the setting concept [Parkes 1989]. The usefulness of combining these two approaches results in gaining the temporality of the multi-layered approach without the disadvantage of using keywords, as keywords have been replaced by a structured content representation.

The general structure of our content representation understands each description as a layer. The connection between the different layers and the data to be described (e.g. the actual audio, video or audio-visual stream) is realised by applying a triple identifier, which indicates the *media identifier*, the *start time* and the *end time*. For example, an actor may perform a number of actions in the same time span. The temporal relation between them can be identified using the start and end point with which those actions are associated. For example, they may all share the same start and end point, and may be performed simultaneously. In this way, complex structured human behaviour or spatial concepts can be represented and hence the audio-visual material retrieved on this basis. Figure 8 shows a layered description of a shot consisting of hundred frames, featuring the actions of a single character.

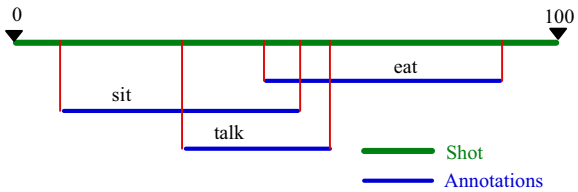


Figure 8: Actions annotated in layers in a 100 frame shot

The horizontal lines in Figure 8 denote actions, whereas the vertical lines delimit the various content based layers that can be extracted from this shot. Applying this schema to all descriptive units enables the retrieval of particular material with no restrictions on the complexity level of a query. Take the simple example described in Figure 8. If there is a need for a character, who eats, sits and talks simultaneously, we are now in the position to isolate the essential part of a shot, as shown in Figure 9.

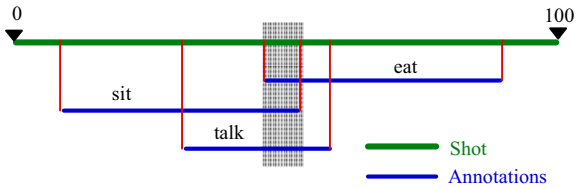


Figure 9: Relevant shot segment for a query for all three actions

The media content representation formalism pays specific attention to the maintenance of objectivity in the description of content. In other words, the description of media content holds constant for the associated time interval. This not only allows multiple content descriptions for the same media unit but also to handle gaps.
For our news environment we developed a set of 18 episodic and technical description schemes. Each DS is represented in MPEG-7 DDL. Once a DS is instantiated it forms a node within the descriptional network that holds the annotations of a news clip or a complete newscast. The description schemes we use are the following:

Newscast	high level organisation scheme of a new cast, containing references to all related news clips and moderations
Newsclip	high level organisation scheme of a new clip, containing all references such as links to relevant annotations and relations to other clips
Link	link structure describing the connection between description scheme and the av-material to be described (data)
Relative time and space	relative (to a given link) temporal or spatial reference to the data
Relation	structure describing the relation between descriptions
Formaldes	formal information about the news clip, such as broadcaster, origin, language, etc.
Bpinfo	production and broadcasting information: when was the clip broadcasted (produced), on which channel, etc.
Subjective_c	subjective description of an event, such as comments of the audience
Media_device	media specific technical information of the data, e.g. lens state, camera movement, etc.
Person	persons participating in the production of the clip, such as reporter, cameraman, technicians, producer
Event	the event covered by the description
Object	object, existing or acting in the event
Character	the relevant character
Action	action of an object or character
Dialogue	spoken dialogues and comments of the event
Setting	setting information of an event, such as country, city, place etc.
Archive	archiving value of the news clip according its content and compositing
Access	access right info, IPR, rights management of the clip

Examples of description schemata for newsclip, media_device and event are described in Figure 10 to 12(the format for all the description schemata is XML Schema):

```
<schema targetNamespace='http://www.darmstadt.gmd.de/mobile/MPEG7/newsclip'
        version='0.1'  xmlns='http://www.mpeg7.org/mpeg7.ddl'
        xmlns:rel='http://www.darmstadt.gmd.de/mobile/MPEG7/relation'
        xmlns:lnk='http://www.darmstadt.gmd.de/mobile/MPEG7/link' >

        <element name="Newsclip" type="NewsclipType"/>

        <complexType name='NewsclipType'>
        <attribute name='clipname' type='string'/>
        <all>
        <element name='Relation' type='RelationType' />
        <element name='MediaFormat' type='MediaFormatType'/>
        <element name='MediaType' type base='MediaTypeType'/>
        <element name='MediaRatio' type='RatioType'/>
        <element name='VideoData' type='VideoDataType' />
        <element name='AudioTrans' type='AudioTransType' />
```

```
            <element name='AudioData' type='AudioDataType' />
            </all>
            </complexType>

<!—Data types used in the previously defined type -->

            <complexType name='RatioType'>
             <sequence>
            <element name='h_ratio' base='positive-integer'/>
            <element name='v_ratio' base='positive-integer'/>
            </sequence>
            </complexType>

<complexType base='rel:relation' name='RelationType' minOccurs='1' maxOccurs='*'/>
<simpleType name='MediaFormatType' base='string'/>
            <simpleType name='MediaTypeType' base='string'/>
            <complexType name='MediaRatioType' base='Ratio'/>
            <complexType name='VideoDataType' base='lnk:link' />
            <complexType name='AudioTransType' base='lnk:link' minOccurs='0' maxOccurs='*'/>
            <complexType name='AudioDataType' base='lnk:link' minOccurs='0' maxOccurs='*'/>

</schema>
```

Figure 10: A4SM newsclip Description Schemata

```
<schema targetNamespace='http://www.darmstadt.gmd.de/mobile/MPEG7/media_device'
             version='0.1'  xmlns='http://www.mpeg7.org/mpeg7.ddl'
             xmlns:cam='http://www.darmstadt.gmd.de/mobile/MPEG7/camera'
             xmlns:aud='http://www.darmstadt.gmd.de/mobile/MPEG7/audio'
             xmlns:tim='http://www.darmstadt.gmd.de/mobile/MPEG7/TimeDS' >

            <element name=“media_device“ type=“media_deviceType“/>

            <complexType name='media_deviceType'>
            <attribute name='name' type='string'/>
            <all>
            <complexType name='effects_type' minOccurs='0' maxOccures='1'/>
            <!--time: relative to data-->
            <element name='start_point' type='tim:TimeDS' />
            <!--time: relative to data-->
            <element name='end_point' type='tim:TimeDS' />
            <element name='camera_data' type= 'cam:camera' minOccurs='0' maxOccurs='1'/>
            <element name='audio_data' type='aud:audio' minOccurs='0' maxOccurs='1'/>
            </all>
            </complexType>

            <simpleType name='effects_type' base='string'/>

</schema>
```

Figure 11: A4SM media_device Description Schemata

```
<schema targetNamespace='http://www.darmstadt.gmd.de/mobile/MPEG7/event'
             version='0.1'  xmlns='http://www.mpeg7.org/mpeg7.ddl'
             xmlns:rts='http://www.darmstadt.gmd.de/mobile/MPEG7/rt_ts'
             xmlns:rel='http://www.darmstadt.gmd.de/mobile/MPEG7/relation' >

            <element name=“event“ type=“eventType“/>

            <complexType name='eventType'>
            <attribute name='name' type='string'/>
            <all>
            <element name='event_type' type='string'/>
            <element name='abstract' type='string'/>
            <element name='HH_video' type='rel:relation' minOccurs='0' maxOccurs='*'/>
            <element name='HH_audio' type='rel:relation' minOccurs='0' maxOccurs='*'/>
            <element  name='content_relations' type=rel:relation' minOccurs='1' maxOccures='*' />
            <element name='timespace' type='rts:rt_ts' />
            </all>
            </complexType>
</schema>
```

Figure 12: A4SM event Description Schemata

With these 18 structures it is possible to access material on the content level (e.g. search for a person, situation, place in the newsclip) as well as on an organisational level (show clips from a particular reporter or show clips related to the clip based on a relation type).
The syntax of the network is based on links and relations.

A **Link** enables the connection from a description to data on a temporal, spatial and spatio-temporal level. A DS providing links we call a 'hub'. The hub is actually the best potential entry point into the network. Moreover, it is the hub where 'absolute addresses' and 'absolute time' are determined. For our news environment we provide two types of description schemata which can hold links, i.e. the newscast-DS and the newsclip-DS. A newscast-DS always behaves as a hub. This means that all its temporal and spatial references are absolute, whereas the references in the associated clips, organised in newsclip-DS, are referential. If a newsClip-DS behaves as a hub (see the right clips within Figure 13), then its temporal and spatial references toward the media are absolute (note, that the annotation algorithm is aware when to use which temporal or spatial representation).
The instantiation of links is ideally performed automatically, though in most cases they will be established semi-automatically. Within our implementation, for example, links are created by the camera based on the scene id set via the handheld device. Please note, that while the …

A **Relation** enables the connection of descriptors within descriptions as well as connections between distinct descriptions. Relations are actually the icing of the cake within the Description Network. It is necessary to define their types, which we did for our news environment as follows:

- **Events:** follows, precedes, must include, supports, opposes, conflict-resolution, evidence, motivation, justification, interpretation, summary, opposition, emotional
- **Character, Setting, Object:** synonym, association, before, equal, meets, overlaps, during, starts, finishes.

Relations will be mainly instantiated in a manual way during the production process (see section 3: Tools).
The instantiation of a DS might be completely automatic, i.e. in the case of the media_device-DS provided by the camera, or semi-automated, as for the event-DS, which is partly instantiated (i.e. the in and out points) using the handheld device.
We use the media_device-DS to explain how description schemata are automatically instantiated and integrated into the content annotation network.
While the camera is recording the digital video in MPEG format, the annotation algorithm polls every 20 ms for changes on image capture parameters. In case a change is detected a media_device-DS structure will be instantiated with the start and end time of the event, the parameter type, e.g. zoom, and its descriptional value. If the camera capture event performs on a longer time span than 20ms, the end time will be entered after the first unsuccessful poll (the algorithm corrects the temporal delay automatically). Once the DS is fully instantiated it is hooked into the document network by providing connections to the relevant documents. In the case of a media_device-DS this might be a connection to the relevant newsclip-DS or newscast-DS.
Figure 13 describes a possible network of A4SM descriptions for news clips, which are represented by the rectangular boxes. Figure 13 also shows the two ways of annotating clips, either as part of a complete newscast (the upper clip), or as a single clip as portrayed for the clip on the right side. It is important to mention that different annotation networks can be related towards each other (e.g. a newsclip about 'Clinton at a press conference' refers to an another clip from an older newscast showing 'Mr. Clinton and Ms. Levinsky').
The described mechanism of generating description schemata semi-automatically in real time supports the idea that the description of audio-visual material is an ongoing process. A description in form of a semantic network (i.e. instantiated description schemata connected via relations) allows easily to create new annotations and relate them to existing material. If new information structures are required, new templates can be designed using the DDL. Moreover, the decomposition of the annotation in small temporal or spatial units supports the streaming aspect of the media units. In case we wish to provide meta-information with the streamed data we can now just use those annotation units which are relevant for the temporal period and which are of interesting for the application using the stream.

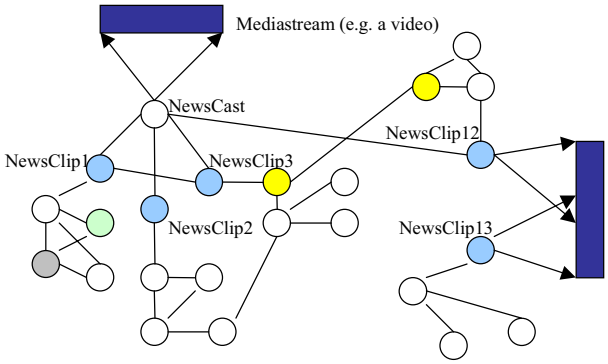


Figure 13: A4SM Description Network and partial parsing approach

However, the approach of relating small unit description schemes, each forming a document and thus a node of the network, also generates problems, mainly with respect to search and content validation.

Search:
The complex structure of the semantic net does not allow easy detection of required information units. It is not difficult to detect the right entry point for the search (usually a hub, but other nodes might also be applicable) but the traversal of the network is, compared to a simple hierarchical tree structure, more complex. Due to its flexibility it is rather problematic to generate orientation structures such as a table of content for a newscast. A potential solution to this problem might be the introduction of a schema header (containing general information about the DS type, the links and relations, and other organisational info) and the schema body with the particular descriptive information.

Validation:
For cases in which new nodes are added, established nodes are changed, or new relations between existing nodes are introduced we have to validate that these operations and the created documents are valid. In our opinion this can only be achieved via partial parsing (see also Figure 14). This means the parser validates only a particular part of the network (e.g. a number of hubs). In such a way we avoid that a complete network needs to be parsed if only a tiny section is effected.

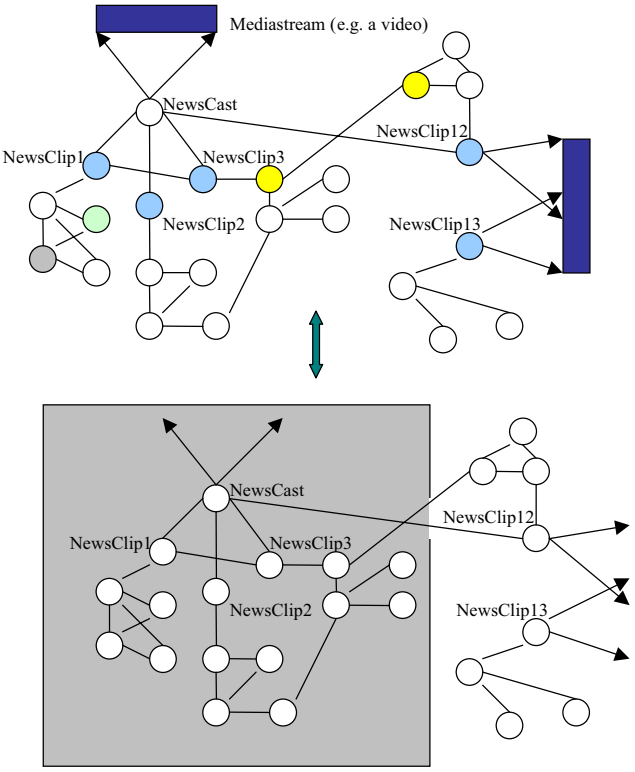


Figure 14: A4SM Description Network and partial parsing approach

We understand that the flexibility of our approach is extending the complexity of maintaining the descriptional structure. Further research has to prove if this is acceptable.
In the appendix of this article we include instantiated description schemata to show how real results look like. Our current implementation organises an annotation network in form of a file system. This means that we have to do the data management (i.e. storage and retrieval of metadata as well as connection between data and annotations) ourselves. However, we would prefer having the structures stored in an XML based object oriented database (see, among others, Tamino from Software AG and dbXML form The dbXML Group, and the XML enabled databases from IBM – XML Extender, Informix Internet Foundation.2000, Oracle - XML Developer's Kit, etc.) At the moment, however, none of them performs storage and retrieval in an acceptable access time.

6 Conclusion

In this paper we presented A4SM, a framework for the creation, manipulation, and archiving/retrieval of media documents, applied for the domain of news. We demonstrated with a basic set of syntactic and semantic descriptors how video material can be annotated in real time and how this information can not only be used for retrieval but also how it can be used during the different phases of the production process itself.

The emphasis in this work is on the provision of tools and technologies for the manual authorship of linear and interactive media production, due to the fact that the description of audio-visual material is an ongoing task specific process. In fact, we believe that a great deal of useful annotation can just be provided by manual labour but also that there is not such a thing as a single and all inclusive content description. We see the need for collective sets of descriptions growing over time (i.e. no annotation will be overwritten but extensions or new descriptions will appear in the form of new documents). Thus, there is not only the requirement for flexible formal annotation mechanisms and structures but also for tools which firstly support human creativity for creating the best material for the required task but secondly also use the creative act to extract the significant syntactic, semantic and semiotic aspects of the content description.

We are aware that the approach described in this article is but a small step towards the intelligent use and reuse of media production material. Nevertheless, we believe that the work undertaken will inform research into the generation of interactive media documents, in particular, and research into media representation, in general. At present, we are engaged in research on tools and techniques for semi-automated, interactive narrative generation, mainly for the domain of documentary.

7 Acknowledgements

We thank Michael Weigand and Angelo Barreiros for programming the MPEG-7 camera. We'd also like to acknowledge the MPEG-7 experts, in particular Jane Hunter, Ernest Wan, Olivier Avaro and Arnd Steinmetz, who provided support in the development of the Description Schemata and for useful discussion during the development of this work. We also thank the HR (Hessischer Rundfunk - Frankfurt) for supporting this work by offering access to their practical sessions.

This work is funded by GMD-IPSI.

8 References

Aguierre Smith, T. G., & Davenport, G. (1992). The Stratification System. A Design Environment for Random Access Video. In *ACM workshop on Networking and Operating System Support for Digital Audio and Video*. San Diego, California

Aigrain, P., Joly, P., & Longueville, V. (1995). Medium Knowledge-Based Macro-Segmentation of Video into Sequences. In M. Maybury (Ed.) (pp. 5-16), *IJCAI 95 - Workshop on Intelligent Multimedia Information Retrieval*. Montréal: August 19, 1995

Arnheim, R. (1956). *Art and Visual Perception: A Psychology of the creative eye*. London: Faber & Faber.

Bloch, G. R. (1986) *Elements d'une Machine de Montage Pour l'Audio-Visuel*. Ph.D., Ecole Nationale Supérieure Des Télécommunications.

Bloom, P.J. (1985). High-quality digital audio in the entertainment industry: an overview of achievements and challenges, IEEE Acoust. Speech Signal Process. Mag., 2, 2-25 (1985)

Borchers, J. & Mühlhäuser, M. . (1998) design Patterns for Interactive Musical Systems. IEEE Multimedia Magazine, Vol.5, No. 3, pp. 36 – 46, July-September 1998

Bordwell, D. (1989). *Making Meaning - Inference and Rhetoric in the Interpretation of Cinema.* Cambridge, Massachusetts: Harvard University Press.

Brachman, R.J. & Levesque, H.J. (1983), *Readings in Knowledge Representation*. San Mateo, California: Morgan Kaufmann Publishers.

Brooks, KM (1999). "Metalinear Cinematic Narrative: Theory, Process, and Tool. " MIT Ph.D. Thesis.

C. Colombo, A. Del Bimbo, and P. Pala (1999), "Semantics in visual information retrieval". IEEE Multimedia, 6(3):38-53, IEEE 1999.

Chakravarthy, A. S. (1994). Toward Semantic Retrieval of Pictures and Video. In C. Baudin, M. Davis, S. Kedar, & D. M. Russell (Ed.), *AAAI-94 Workshop Program on Indexing and Reuse in Multimedia Systems*, (pp. 12 - 18). Seattle, Washington: AAAI Press.

Davis, M. (1995) *Media Streams: Representing Video for Retrieval and Repurposing.* Ph.D., MIT.

Del Bimbo, A. (1999). "Visual Information Retrieval", Morgan Kaufmann Ed, San Francisco, USA

Eco, U. (1985). *Einführung in die Semiotik*. München: Wilhelm Fink Verlag.

Eco, U. (1977). *A Theory of Semiotics*. London: The Macmillan Press.

Fehlis, H. (1999) Hybrides Trackingsystem für virtuelle Studios. Fernseh- + Kinotechnik; Bd. 53, Nr. 5

Gupta A. & Jain R. (1997) Visual information retrieval. Communications of the ACM, 40:71-79.

Gordon, A. S., & Domeshek, E. A. (1995). Conceptual Indexing for Video Retrieval. In *IJCAI 95 - Workshop on Intelligent Multimedia Information Retrieval*. Montréal, Canada: 19th of August 1995

Greimas, J. (1983). *Structural Semantics: An Attempt at a Method.* Lincoln: University of Nebraska Press.

Hirata, K. (1995). "Towards Formalizing Jazz Piano Knowledge witha Deductive Object-Oriented Approach". Proceedings of Artificial intelligence and Music, IJCAI, pp. 77 – 80, Montreal.

Hunter, J& Armstrong,L. (1999)."A Comparison of Schemas for Video Metadata Representation", Proceedings of the WWW8, Toronto, May 10-14.

Johnson, S.E., Jourlin, P., Spärk jones, K. 7 Woodland P.C. (2000). Audio Indexing and retrieval of Complete Broadcast News Shows. RIAO' 2000 Conference proceedings, Vol 2, pp. 1163 – 1177, Collége de France, Paris, France, April 12-14 2000

Lindley, C. (2000). *A Video Annotation Methodology for Interactive Video Sequence Generation,* BCS Computer Graphics & Displays Group Conference on Digital Content Creation, Bradford, UK, 12-13 April 2000.

Lemström, K. & Tarhio, J. (2000). Searching Monophonic Patterns within Polyphonic Sources. RIAO' 2000 Conference proceedings, Vol 2, pp. 1163 – 1177, Collége de France, Paris, France, April 12-14 2000

MPEG Requirements Group (1999 a): *"MPEG-7: Context, Objectives and Technical Roadmap", Doc. ISO/MPEG N, MPEG Melbourne Meeting, October 1999*

MPEG Requirements Group (1999b) *"MPEG-7Requirements Document V.11", Doc. ISO/MPEG N, MPEG Melbourne Meeting, October 1999*

MPEG Requirements Group (2000c) *"MPEG-DDL Working Draft V 2.0", Doc. ISO/MPEG W3293, MPEG Noordwijkerhout Meeting, April 2000.*

Melucci, M. & Orio, N.. (2000). SMILE: a System for Content-based Musical Information Retrieval Environments. RIAO' 2000 Conference proceedings, Vol 2, pp. 1261 - 1279, Collége de France, Paris, France, April 12-14 2000

Mills, T.J., Pye, D., hollinghurst, N.J. & Wood, K.R. (2000). At&TV: Broadcast Television and Radio Retrieval. RIAO' 2000 Conference proceedings, Vol 2, pp. 1135 – 1144, Collége de France, Paris, France, April 12-14 2000

Nack, F. (1996) "AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing," Ph.D., Lancaster University, 1996.

Nack, F. and Parkes, A. (1997). Towards the Automated Editing of Theme-Oriented Video Sequences. In Applied Artificial Intelligence (AAI) [Ed: Hiroaki Kitano], Vol. 11, No. 4, pp. 331-366.

Nack, F. & A. Steinmetz (1998). *Approaches on Intelligent Video Production*. Proceedings of ECAI-98 Workshop on AI/Alife and Entertainment, August 24, 1998, Brighton.

Nack, F. & Lindsay, A. (1999) Everything you wanted to know about MPEG-7: Part I & II IEEE MultiMedia ‚July - September 1999,pp. 65 - 77, October - December 1999,pp 64 – 73, IEEE Computer Society

Nack, F. & C. Lindley (2000) "Environments for the production and maintenance of interactive stories", Workshop on Digital Storytelling, Darmstadt, Germany, 15-16/6/2000.

Nagasaka, A., & Tanaka, Y. (1992). Automatic video indexing and full-search for video appearance. In E. Knuth & I. M. Wegener (Eds.), *Visual Database Systems* (pp. 113 - 127). Amsterdam: Elsevier Science Publishers

Pachet, F. & Cazzly, D. (2000). A Taxonomy of Musical Genres. RIAO' 2000 Conference proceedings, Vol 2, pp. 1238 – 1245, Collége de France, Paris, France, April 12-14 2000

Parkes, A. P. (1989) *An Artificial Intelligence Approach to the Conceptual Description of Videodisc Images*. Ph.D. Thesis, Lancaster University.

Parkes, A. P. (1989). Settings and the Settings Structure: The Description and Automated Propagation of Networks for Perusing Videodisk Image States. In N. J. Belkin & C. J. van Rijsbergen (Ed.), *SIGIR '89*, (pp. 229 - 238). Cambridge, MA.

Peirce, C. S. (1960). *The Collected Papers of Charles Sanders Peirce - 1 Principles of Philosophy and 2 Elements of Logic, Edited by Charles Hartshone and Paul Weiss.* Cambridge, MA: The Belknap Press of Harvard University Press.

Pfeiffer, S.; Fischer, S. & Effelsberg (1996). "Automatic Audio Content Analysis". Proceedings of the ACM Multimedia 96, pp. 21-30, New York.

Robertson, J., De Quincey, A., Stapleford T. & Wiggins, G. (1998) Real-Time Music Generation for a Virtual Environment. Proceedings of ECAI-98 Workshop on AI/Alife and Entertainment, August 24, 1998, Brighton.

Sack, W. (1993). Coding News And Popular Culture. In The International Joint Conference on Artificial Intelligence (IJCA93) Workshop on Models of Teaching and Models of Learning. Chambery, Savoie, France.

SMPTE (1999). Dynamic Data Dictonary Structure, 6. Draft, September 1999.

Zhang, H., Gong, Y., & Smoliar, S. W. (1994). Automated parsing of news video. In *IEEE International Conference on Multimedia Computing and Systems*, (pp. 45 - 54). Boston: IEEE Computer Society Press.

TALC (1999). http://www.de.ibm.com/ide/solutions/dmsc/

Tonomura, Y., Akutsu, A., Taniguchi, Y., & Suzuki, G. (1994). Structured Video Computing. *IEEE MultiMedia*, 1(3), 34 - 43.

Wold, E., Blum, T., Keislar, D. & Wheaton, J.(1996). Content-Based Classification, Search, and Retrieval of Audio**.** IEEE Multimedia Magazine, Vol.3, No. 3, pp. 27 – 36, Fall 1996

XML Schema Part 0 (2000) Primer, W3C Working Draft, 7 April, http://www.w3.org/TR/xmlschema-0/

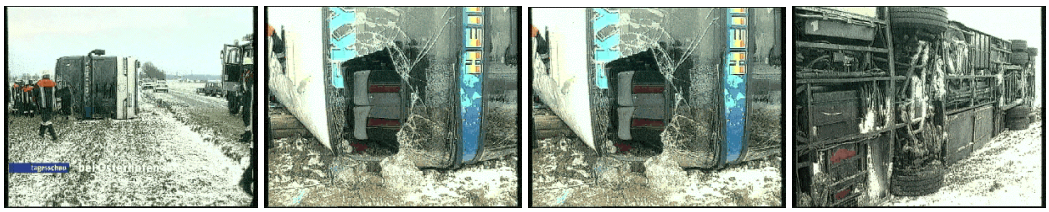
XML Schema Part 1 (2000) Structures W3C Working Draft, 7 April, http://www.w3.org/TR/xmlschema-1/

XML Schema Part 2 (2000) Datatypes W3C Working Draft, 7 April, http://www.w3.org/TR/xmlschema-2/

Yeung, M. M., Yeo, B., Wolf, W. & Liu, B. (1995). Video Browsing using Clustering and Scene Transitions on Compressed Sequences. In *Proceedings IS&T/SPIE '95 Multimedia Computing and Networking, San Jose*. SPIE (2417), 399 - 413.

Appendix: Example of a news annotation

The following example of the description of a newsclip is based on a our 18 different description schemata. The description language used is the MPEG-7 DDL from December 1999 [ref.]. What is shown below is not the complete description, since its volume would not fit the scope of this article. However, the example is detailed enough to to show the approach and the possibilities of the underlying model.
The clip we describe is a real newsclip, broadcasted by ARD-Tagesschau at 20th January 2000. The clip consists of 7 events, of which we show below the first four scenes representing the corresponding key-frames. Note, that the newsclip DS functions in this example as a hub.



event 1 event 2 event 3 event 4

Instantiated newsclip DS

```
<Newsclip clipname='Tagesschau200001202000'>
        <Link name='video_data'>
                <target>http:/www.darmstadt.gmd.de/mpeg7/ts200001202000_v</target>
                <time>00/13/22/21-00/13/49/14</time>
        </Link>
        <Link name='audio_data'>
                <target>http:/www.darmstadt.gmd.de/mpeg7/ts200001202000_a</target>
                <time>00/13/22/21-00/13/48/29</time>
        </Link>
        <Relation name='formaldescription'>
                <target>http://www.darmstadt.gmd.de/MPEG7/formaldes_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev1'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event01_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev2'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event02_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev3'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event03_inst</target>
                <uniorbi>uni </tuniorbi>
                <reltype>contains </reltype>
        </Relation>
        <Relation name='toev4'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event04_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev5'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event05_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev6'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event06_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='toev7'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event07_inst</target>
```

```
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>

        <media_format>MPEG-1</media_format>
        <media_type>video</media_type>
        <media_ratio>
                <h_ratio>4</h_ratio>
                <v_ratio>3</v_ratio>
        </media_ratio>
</Newsclip>
```

Instatntiated Formaldes DS

```
<formaldes name='fts200001202000'>
        <title>children's bus accident in Bavaria</title>
        <Relation name=''>
                <target>http://www.darmstadt.gmd.de/MPEG7/broadcast01</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name=''>
                <target>http://www.darmstadt.gmd.de/MPEG7/archive01</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name=''>
                <target>http://www.darmstadt.gmd.de/MPEG7/person01</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>

        <broadcaster>ARD</broadcaster>
        <keyword>accident</keyword>
        <keyword>children</keyword>
        <keyword>Bavaria</keyword>
        <category>news</category>
        <origin>own</origin>
        <language>german</language>
</formaldes>
```

Instatntiated event DS for event 2

```
<event name='e02_ts200001202000'>
        <event_type></event_type>
        <abstract>the broken front window of the schoolbus</abstract>
        <Relation name='eventchain'>
                <target>http://www.darmstadt.gmd.de/MPEG7/event03_inst</target>
                <uniorbi>bi</tuniorbi>
                <reltype>connects</reltype>
        </Relation>
        <Relation name='bus'>
                <target>http://www.darmstadt.gmd.de/MPEG7/bus02_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='setting'>
                <target>http://www.darmstadt.gmd.de/MPEG7/setting02_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='dialogue'>
                <target>http://www.darmstadt.gmd.de/MPEG7/dial02_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <Relation name='media1'>
                <target>http://www.darmstadt.gmd.de/MPEG7/media02_inst</target>
                <uniorbi>uni</tuniorbi>
                <reltype>contains</reltype>
        </Relation>
        <timespace>
                <in>00/00/04/09</in>
                <out>00/00/07/05</out>
```

```
        </timespace>
</event>
```

Instatntiated media_device DS used in the instantiation for event 2

```
<media_device name='m02_ts200001202000'>
        <comment> is empty in this example, because we dont have
                camera infos </comment>
        <effects_type>closeup</effects_type>
        <start_point>00/00/04/09 /start_point>
        <end_point>00/00/07/05</end_point>
</media_device>
```