View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Crossref

Designing Annotation Before It's Needed

Frank Nack
CWI, Amsterdam
Kruislaan 413, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Tel: +31 20 5924223
E-mail: Frank.Nack@cwi.nl

Wolfgang Putz
GMD-IPSI, Darmstadt
Dolivostr 15
D-64293 Darmstadt, Germany
Tel.: +49 6151 869926
E-mail: Wolfgang.Putz@gmd.de

ABSTRACT

This paper 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 and outline the underlying XML-Schema based content description structures of A4SM. We then describe tools for a news and 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. Finally, we discuss the pros and cons of our approach of evolving semantic networks as the basis for audio-visual content description.

Keywords

News production, semantic networks, automated annotation, XML Schema, MPEG-7

1. INTRODUCTION

Since the beginning of the 1980's 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 [4, 36, 1, 34, 41, 52, 46, 11, 51, 29, 9, 24]. Similar developments were acquired for audible information [5, 18, 39, 47, 40, 6, 45].

However, the provided technology follows the strains of traditional written communication by supporting the linear representation of an argument resulting in a final multimedia product of context-restricted content. This is an instance of Marshall McLuhan's observation that new media technology is used initially to solve old problems. Conversely, 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.

A media infrastructure supporting such aims requires that a given component of media exists independently of its use in any given production. Making use of such independent media items necessitate the extraction of the relationship between the signs of the audio-visual information unit and the idea they represent [3, 38, 13, 16,7].

Systems are therefore necessary to manage independent 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 (intelligent vision, etc.), or media conversion (speech to text, visual transcoding, etc.). Systems are also required for the authoring of representations creating a particular production, such as news items, a documentary, or an interactive game. In other words we need 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 [8] which then can be transformed into a description based on a formal description language.

In this paper we present work that is directed towards the application of IT support for all stages of the media production process within the A4SM framework (pronounced APHORISM — **A**uthoring **S**ystem for **S**yntactic, **S**emantic and **S**emiotic **M**odelling), including 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 assess the applicability and acceptability of this framework in a news production environment.

In the following article we first outline our workflow model of the news production process and present requirements for semi-automated digital news production, which are based on a three month investigation at WDR (WDR: Westdeutscher Rundfunk – Germany largest public broadcasting station) and HR (Hessischer Rundfunk – the broadcasting station of the federal state of Hesse) where we talked with and observed the work of reporters, cameramen, and editors during actual productions. We then describe the A4SM architecture, its semantic network based approach for data storage and management, and address its XML Schema-based representational structures. Then we illustrate a number of tools we implemented for the news environment. Finally, we assess our achievements thus far, and provide an overview of further work.

2. NEWS PRODUCTION – ONE TYPE OF MEDIA PRODUCTION

Media production, such as for news, documentaries, interactive games, or virtual environments is a complex, resource demanding, and distributed process, 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, 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.

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 that today's researchers and archivists try to analyse and re-construct out of the final product - usually with limited success.

Moreover, current professional IT tools support the tendency of 'lost information'. These applications assist in the processes of transforming ideas into scripts (e.g. a text editor), sustain in digital/analogue editing (Media Composer, FAST 601, etc.), or support production management (Media-PPS, SAP R/2, 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.

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, as well as already processed and archived material, should immediately be retrievable by all members of the editorial staff.

To facilitate this 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.

Direct access on archived material requires additional information. Here the emphasis lies on the provision of dynamic structures and implicit connections 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 materials 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, 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 translation between one representation and the other does not result in the loss of salient features of either representation. The challenge is to provide an environment that integrates the instantiation and maintenance of these dynamic structures into the actual working process. However, no such environment does exist to the current day.

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 newsroom. The proposed requirements are:

- A device that automatically stores the acquired video stream together with an associated stream of relevant metadata like co-ordinates, camera work or lens movement
- An annotation tool for the reporter to provide real-time semantic annotation during acquisition (such as the contextual quality of images and the soundtrack), and to include this information time-coded into the stream of metadata
- An on-site editing tool for the reporter that provides means for stratified annotation on segment level and that incorporates the edit decision list and the annotations into the metadata stream
- An architecture for fast and secure 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

Since our work is directed towards the application of IT support for all stages of the media production process, we will cover in the following sections the first three and the sixth requirement. We will now introduce A4SM and show how it can support the news production model.

3. NEWS PRODUCTION AND A4SM

Research in industry and academia has 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. [2, 15, 12, 28, 23, 27, 25]. 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 [10, 19, 42, 35]). Moreover, the general difficulty 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.

A4SM tries to overcome these limitations by providing a distributed digital environment covering all sorts of media production. The aim of the framework is to support 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 for media items and related content descriptions, 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, each of the tools relies on the consistent data structures, which guarantee the composition of multi-dimensional network of relationships between different kinds of information units.

The flexibility of the A4SM framework becomes clearer by applying the news domain to it. In news, for example, we don't find physical scripting due to time constraints and the unpredictability of events. However, mental concepts are developed on the way to and on the set and most of them are reflected in the gathered material. Thus, complex scripting tools as suggested in [31] and [33] for commercial dramatic films or interactive stories, are not applicable here. However, the existing syntactic and semantic data structures still need to be annotated – if not during the pre-production, then during the production phase, which requires particular news tools. Though the tools change the underlying representational structures remain.

Based on the discussion with reporters, cameramen, and editors, we not only identified a set of required data structures and relations between them (see section 4) but also developed tools for efficient IT support during production and post-production (see section 5).

Before we introduce the particular tools and describe, how they interoperate, we would first like to describe the general concepts and assumptions of the repository and the representation schemata for news.

4. 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. Every degree of cognition might be illuminative for other interests and should remain accessible.

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, there will be temporal gaps in annotations, we will see 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 representation structures with the capability to change and grow. This also 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.

4.1 General representational concepts in A4SM

To facilitate the dynamic use of audio-visual material, A4SM's general attempt at content description applies the strata oriented approach [1] in combination with the setting concept [37]. 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, which comprise not only textual descriptions but also descriptions of image decomposition based on feature extraction.

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. In this way, complex structured human behaviour or spatial concepts can be represented and hence the audio-visual material retrieved on this basis. Figure 1 shows a layered description of a shot consisting of hundred frames, featuring the actions of a single character.

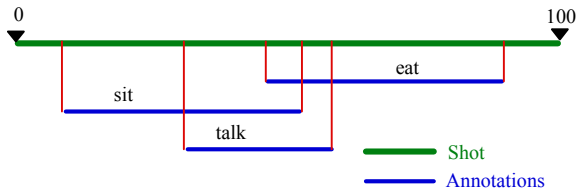

Figure 1: Actions annotated in layers in a 100 frame shot

The horizontal lines in Figure 1 denote actions, whereas the vertical lines delimit the various action-based content descriptions 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 2. 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.

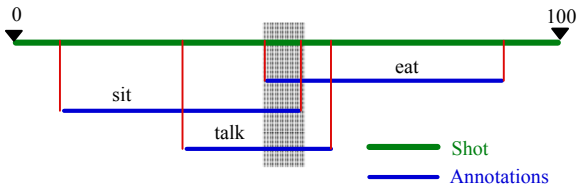

Figure 2: Relevant shot segment for a query for all three actions

The described flexible organisation of media content description and the related media data requires not an enclosed hierarchical structure, as we know it from document structures, but rather the adaptable construction in form of a semantic network.

This concept emerged from research in knowledge representation [8, 44, 17] and features three significant functions, which make it suitable as a platform for supporting the needs of media production:

- It provides semantic, episodic, and technical memory structures (i.e. information nodes) with the capability to change and grow, thus allowing an ongoing task specific process of inspection and interpretation of source material
- It facilitates the dynamic use of audio-visual material using links, enabling the connection from multi-layered information nodes to data on a temporal, spatial and spatio-temporal level. Moreover, since the description of media content holds constant for the associated time interval, we are now in the position to handle multiple content descriptions for the same media unit and also to handle gaps
- It enables the semantic connection between information nodes using typed relations, thus structuring the information space on a semantic as well as syntactic level.

The following sections explain in more detail how A4SM supports these different functions. We begin with nodes and outline then our handling of links and relations.

4.2 Nodes in A4SM

A node in A4SM is an instantiated schema providing either denotative or semantically loaded technical characteristics of the data. The available number of node schemata is restricted, thus indexing and classification can be performed in a controlled way, whereas the number of provided nodes in the descriptional information space is not.

The obvious choice for representing these structures would have been using the DDL of MPEG-7 [21] or suggested schemata by MPEG-7 [22].

The objective of the MPEG-7 group [31] 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 schemata (DS) and descriptors (D). 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, camera motion in a video, the actors in a movie, etc. Descriptors and description schemata are represented in the MPEG-7 Description Definition Language (DDL). The current version of the DDL is XML Schema based [48, 49, 50], 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 [32].

At the end we decided not to use any of the MPEG-7 description schemata. There are mainly three reasons for that decision. First, MPEG-7 is hierarchy centred, which is not astonishing since efficient access and retrieval were and still are the driving development forces. Even though Part 5 of the Standard, which is about Multimedia Description Schemes, provides a Graph DS [22, chap.7.6], this Graph DS is merely aimed to establish semantic relations between parts of a description. Reading chapter 5 of the same document, which talks about the forming of MPEG-7 schema valid instance documents and description units or fragments, it becomes obvious that the description of data in MPEG-7 is enclosed in one document that itself is structured in form of a tree. The schemata for this document type are fixed and cannot be altered. This approach is far to restrictive for the described production and use of meta data for audio-visual data in news.

The second reason, why we chose not to use provided MPEG-7 schemata is based on the distinction in MPEG-7 between a complete description (using MPEG-7Main as root element) and partial description units (using MPEG-7Unit as root element). As illustrated above, annotating is dynamic and iterative work that maps the not strictly pre-structured process in the perception of a concept. The distinction of a complete and fragmental description is sort of academic and adds an unnecessary extra level of descriptional complexity.

Finally, we opted against the use of suggested MPEG-7 schemata due to their great number and interlocked nature, which makes the use of few constructive schemata in isolation merely impossible.

However, as we intended to be MPEG-7 compliant we decided to use the MPEG-7 DDL, which allows the creation of new description schemata. The only problem we have been facing is that up to the time of writing the DDL has not been put into an agreed format and also still lacks parser support. Thus, we chose to build our implementation using XML Schema and perform the necessary changes, if any, ones an MPEG-7 parser is available.

The authors are aware of the fact that RDF Schema (RDFS) would actually be an even better approach for describing relation-based semantics than XML Schema. However, at the time of writing this alternative is merely under discussion in MPEG-7 [20]. The authors would highly welcome the inclusion of RDFS but will wait with further implementation, due to conformity reasons, until decisions are made in the MPEG-7 standard.

For our news environment we developed a set of 18 schemata, which provide information on a denotative and technical level. The design of the technical oriented schemata fulfils two aspects: these schemata can be automatically instantiated and they do describe technical information that in itself offers semantic value, e.g. a zoom-in usually focuses the viewers attention on something important. The description schemata we developed and use in the 'A4SM for News' implementation are as follows:

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)
TSR	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
Mediadevice	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

With these 18 structures it is possible to access material on the content level (e.g. search for a person, situation, place in a 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).

4.3 Links and relations in A4SM for News

As pointed out in section 4.1, we organize all meta information about the actual audio and video stream in form of a semantic network. Figure 3 describes a possible network of A4SM descriptions for news clips, which are represented by the rectangular boxes. Figure 3 also shows the two ways of annotating the data, either as part of a complete newscast (the upper media stream represented in form of a rectangle), or as single clips as portrayed for the media stream 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').

Hence, we require two sorts of connections attached to nodes:

- Connection between data and descriptional node (link)
- Connection between the nodes themselves (relation)

A **Link** enables the connection from a description to data on a temporal, spatial and spatio-temporal level.

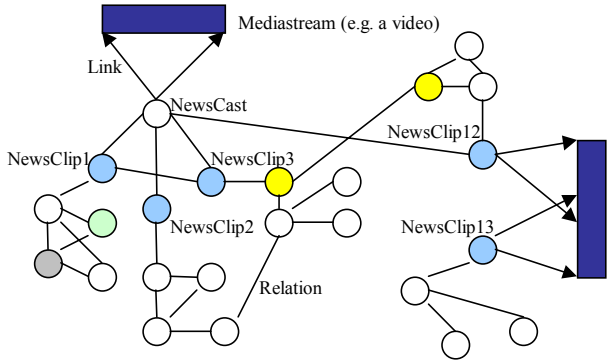

Figure 3: A4SM Description Network

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 that 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 3), 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). As we will see later, the instantiation of links in A4SM for News is always performed automatically (see the section 5 on tools below). The structure of a link in A4SM is described in Figure 4.

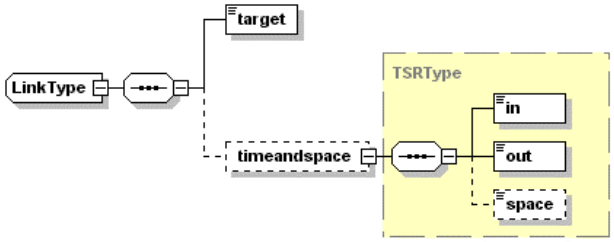

Figure 4: Structure of an A4SM link

A **Relation** enables the connection of descriptors within descriptions as well as connections between distinct descriptions. There can be up to n relations between two nodes. For providing a coherent semantics, it is necessary to define the domain relevant relation types. As a result of our knowledge elicitation we defined the following for the news environment:

- **Events:** follows, precedes, must include, supports, opposes, introduces, , motivation, conflict-resolution, evidence, justification, interpretation, summary, opposition, emotional

- **Character, Setting, Object:** part-of, synonym, association, before, equal, meets, overlaps, during, starts, finishes.

Relations will be instantiated in a semi-automated way during the production process (see section 5: Tools).

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) as well as the validation of documents ourselves.

The different XML schemata, which represent the structures of potential nodes in the network, are defined using XMLSpy Version 3.5. This tool supports the XML Schema version from October 2000 [48, 49, 50] and also converts older versions of XML Schema into the current version. Moreover, this parser allows through a scripting interface its adaptation to user specific applications.

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.

Based on the above discussion on the general structures of our semantic network based repository, we describe now how the structures are instantiated in A4SM. For that we will introduce the tools we have been developing over the last couple of years for the news domain.

5. THE TOOLS

As mentioned above the most critical constraint within news production is time and thus the aim was to support the actual production process in such a way that no extra burden is put on the involved specialists by adding the annotations during production. As outlined in section 3, we had to come up with solutions for merely two production phases, i.e. the production itself and the postproduction, which had to cover important pre-production information as well.

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 shot with respect to audio or visual elements) while the cameraman is shooting. As a result we designed and developed a hard disk camera that automatically stores the acquired video stream together with relevant information, such as co-ordinates, camera work or lens movement (see section 5.1). Additionally we provided a mobile handheld annotation tool for the reporter to allow for real-time annotation during acquisition on a basic semantic level, i.e. capturing in and out points for sound and images (see section 5.2).

Post-production calls for the support of the collaboration between reporter and editor. During the interviews 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. However, the tool should provide additional information so that an editor at the broadcasting station should be able to adjust the material in an appropriate way, if required. 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 typed in annotations into the XML-schema description structure (see section 5.3).

We now describe the different tools we developed in our lab in more detail. We start with the camera and the handheld device and then show how the generated annotations of these tools are used during the post-production phase for finalising the news unit. The section will conclude with an evaluation of the gained insights.

5.1 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 [7,13]. Figure 5 demonstrates the interface in our lab settings for determining, which camera settings are traced, i.e. data about camera movement (pan and tilt), lens action (zoom and focus), shutter, gain, and iris position. Moreover, we collect information about the spatial position of the camera. For the latter we use a magnetic tracker. The image in portrays the current video being shot and thus represents what the cameraman would see in the viewfinder.

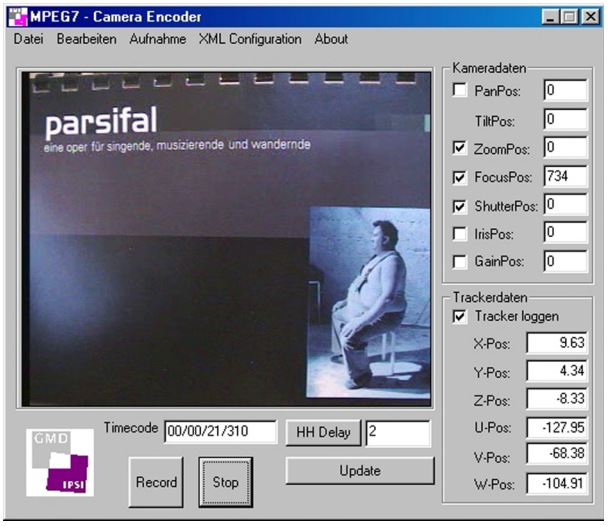

Figure 5: **Interface for the camera handling**

The limited 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 [43] and by the GMD virtual studio [14].

The actual hardware components used in our demonstration environment are:

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

As mentioned in section 4, it is the aim of the A4SM framework to provide as much automatic annotation as possible. The most potential areas are those of technical descriptions. We use the camera to show how the instantiation processes of those annotations or the spatio-temporal identification marks for audio-visual data are performed, based on a linking mechanism. The advantage of the generic approach is that the expert, here the cameraman, can concentrate on his tasks without being concerned about storage organisation or general presentation.

Before the cameraman starts shooting, he establishes the set of features that should be annotated during filming. This can either been done by using an interface as suggested in Figure 5, or via a programmable code-card, which is installed into the camera before shooting starts. 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.

The synchronisation (i.e. linking) between data and annotation is resolved via time codes 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. When the camera is then recording the digital video in MPEG format, the annotation algorithm polls every 20 ms for changes on the relevant image capture parameters. In case a change is detected a mediadevice-DS structure will be instantiated with the start and end time of the event, the parameter type, e.g. zoom, and its descriptional value. Furthermore, the node will be immediately associated to the relevant hub by providing connections to the relevant documents (using the scene ID). In the case of a mediadevice-DS this might be a connection to the relevant newsclip-DS or newscast-DS.

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).

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 only extra work for the cameraman is to establish the required set of description schemata.

The approach taken in this implementation of the camera goes beyond what current cameras offer. First of all, there are only a small number of professional digital cameras that are able to capture additional audio/visual meta-data, and not surprisingly all of them are developed for digital news gathering (DNG). As suggested by our approach these professional cameras uses an external memory source (e.g. a 4Gbyte harddisc recorder with AVID technology), where markers are set while the video is recorded (Ikegami). However, the particular markers are provided in a proprietary format, instead of the open XML based format offered by A4SM.

5.2 The handheld

Figure 6 demonstrates the interface simulation of the handheld device within our demonstration environment. The handheld device is instantiating event description schemata in a semi-automated way.

The device provides a monitor for screening the recorded material of the camera in real-time (the frame rate is 13 fps, enough for 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.

The reporter can adjust the scene id at any time if the camera is not recording. Once the camera is active, it first identifies the current id setting of the handheld (a handshake protocol on infrared basis is used), 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 time codes (provided by the camera) and the scene id.

Figure 6: **Simulation of a handheld device for the annotation of simple semantic information**

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 time codes and related nodes on paper.

Having introduced the production tools, we can now discuss the pre-production environment.

5.3 The editing suite

This tool uses the annotations of the audio-visual media to automatically group the material based on conceptual dependencies within news. Based on the discussion with editors on how they group incoming material we distilled the following rules for grouping material:

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.

At the current stage of development our prototypical editing environment uses the acquired meta-information from the annotation process 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).

The semi-automated editing suite provides the reporter with an instant overview of the available material and its intrinsic relations represented through the spatial order of its presentation (see Figure 7a as an example). The reporter is now able to mark the relevant video clips for the newscast by pointing. The order of pointing indicates the sequential appearance of the clips and the length of pointing their importance. The different shapes around the selected clips resemble the role of the clip within the news unit, i.e. the square stands for the introduction, whereas the circle represents the important part. Finally the reporter provides the overall time of the clip. 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 7b).

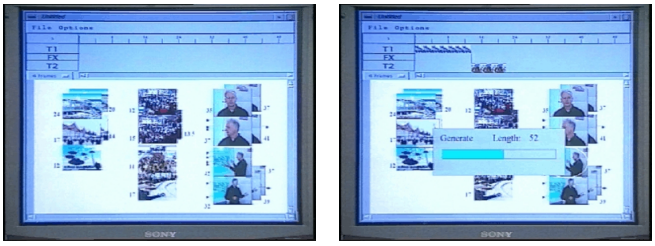

Figure 7 (a – b): The handling of a semi-automated editing suite

The simple planner for news composition is based on 22 rules for the automated clipping of a shot and juxtaposition of shots. Below we describe those rules that are concerned with the rhythmical shape of a sequence.

Strategy A and Strategy B 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 (values are based on estimates provided by editors). 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.

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

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

For a more detailed theoretical discussion of automated film editing see [30].

5.4 Evaluation of repository and tools

The responses cameramen, editors, and reporters regarding the tools either presented in our lab or at the TV fare in Berlin and on the occasion of 10th Anniversary of ERCIM (European Research Consortium for Informatics and Mathematics) in Amsterdam were encouraging.

In general it turned out that most people at broadcasting stations feel quite reserved towards new ways of production, at least among editorial staff. However, reporters, above all the younger ones , said that they would use tools as beeing shown.

A different experience we had with cameramen. Nearly all of them understood the extra value of annotating material during production and all of them immediately saw the advantage for them to document their work – might that be in annotating a new camera effect as their invention or by simply identifying, which parts of a film were done by them. They liked in particular the idea that no extra burden was placed on them. However, they detested the idea of the visual screen on the handheld – which every one of them described as a control mechanism for the reporter and thus as a way of degrading their artistic freedom.

On the representational level we showed that 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 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 XML schema, or MPEG-7's 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.

However, the approach of relating small unit description schemata, 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 allo w 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. 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 affected.

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.

6. CONCLUSION AND FUTURE WORK

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 18 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 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. In fact, there is lots of evidence 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. We also include the missing four sections of the requirements for digital news production.

7. ACKNOWLEDGMENTS

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 WDR (Westdeutscher Rundfunk - Köln) and 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

[1] 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

[2] 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

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

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

[5] 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)

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

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

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

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

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

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

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

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

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

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

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

[17] Halasz, F.G. (1988). Reflection On Notecards: Seven Issues For The Next Generation Of Hypermedia Systems. Communications of the ACM, July 1988, 31 (7).

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

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

[20] Hunter, J. & Lagoze, C. (2001). Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles. In: The Tenth International World Wide Web Conference, Hong Kong pp. 457—466, May 1-5, 2001

[21] ISO MPEG-7(2001). Text of ISO/IEC FCD 15938-2 Information Technology - Multimedia Content Description Interface – Part 2 Description Definition Language, ISO/IEC JTC 1/SC 29/WG 11 N4002, March 2001

[22] ISO MPEG-7(2001). Text of ISO/IEC 15938-5/FCD Information Technology - Multimedia Content Description Interface - Part 5 Multimedia Description Schemes, ISO/IEC JTC 1/SC 29/WG 11 N3966, March 2001

[23] 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

[24] 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.

[25] 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

[26] MPEG Requirements Group (2000). Overview of the MPEG-7 Standard (version4.0), Doc. ISO/MPEG N3752, MPEG La Baule Meeting, October 2000.

[27] 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

[28] 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

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

[30] 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.

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

[32] Nack, F. and 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

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

[34] Nagasaka, A. and 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

[35] Pachet, F. and 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

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] Santini, S. and Ramesh J. (2000). Integrated Browsing and Querying for Image Databases. IEEE MultiMedia ,July - September 2000,pp. 26 - 39, IEEE Computer Society

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

[44] Sowa, J. F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley Publishing Company.

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

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

[47] 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

[48] XML Schema Part 0 (2000). Primer, W3C Candidate Recommendation, 24 October 2000, http://www.w3.org/TR/xmlschema-0/

[49] XML Schema Part 1 (2000). Structures W3C Candidate Recommendation, 24 October 2000, http://www.w3.org/TR/xmlschema-1/

[50] XML Schema Part 2 (2000). Datatypes W3C Candidate Recommendation, 24 October 2000, http://www.w3.org/TR/xmlschema-2/

[51] 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.

[52] 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.