About the influence of computer semiotics on communal Intelligence[1]

Frank Nack
CWI, Amsterdam
Frank.Nack@cwi.nl

Introduction

Since the beginning 1980' we have discovered a radical technological revolution that allowed the digitisation 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.

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, the established culture of the information space has not yet taken a specific shape, tools are in their infancy and hence it is still time to develop new visions for techniques, sign systems, organisational structures and rule mechanisms, which will allow us to establish an 'communal' way of thinking, to bundle intelligent and creative powers, and for solving complex problems. For example, we will be able 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. 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, such as, mosaicing sequences to extract their backgrounds and separating their foreground object sequences for reuse in different contexts.

One of the most elementary conditions for a communal intelligence is that communication tools must make knowledge publicly available and provide means to make others aware of available sources. This requires a language that integrates in a natural way into daily activities by supporting at the same time the semi-automated generation of 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. The language should also support 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.

In other words, communal intelligence is about the management of meaning and meaning is always inextricably linked to the medium and its content, which form and carry the message. The goal of analysing the medium is to reveal the textual or formal aspects of it, and how these support the creation of meaning.

The following descriptions of projects provide examples of how the concept of communal intelligence can be implemented. Over the last six years the author has been involved mainly in two projects, which focused on the definition of methods to perform them-based video manipulation based on its semantic and syntactic aspects (AUTEUR), and the representation of media content during its production, allowing the reuse of the gathered structures, meta information and media units (A4SM).

AUTEUR (Nack 96, Nack & Parkes 97)

The aim of AUTEUR (**A**rtificial Intelligence **U**tilities for **T**hematic Film **E**diting using Context **U**nderstanding and **E**diting **R**ules) was to establish a system that synchronises automatic story generation for visual media with the stylistic requirements of narrative and medium related presentation. The AUTEUR system consists of:

- *An ontological representation of narrative elements such as actions, events, and emotional and visual codes,* based on a semantic net of conceptual structures related via six types of semantic links (e.g. synonym, subaction, opposition, ambiguity, association, conceptual). A coherent action-reaction dynamic is provided by the introduction of three event phases, i.e.

[1] The term 'communal intelligence' was coined by Pierre Lévy in his book 'L'intelligence collective. Pour une anthropologie du cyberspace'. Édition la Découverte, Paris, 1994.

motivation, realisation and resolution. The essential categories for the structures are action, character, object, relative position, screen position, geographical space, functional space and time. *The textual representation of this ontology describes semantic, temporal and relational features of video in hierarchically organised structures,* which overcomes the limitations of key word-based approaches.

- *A set of 26 humour strategies*, which combine the logic of narrative structures with functional operators of comic primitives. The identified primitives within humour are grouped into two classes: supportive primitives (exaggeration, timing and readiness) and constructive primitives (incongruity and derision).
- *A simplified model of the editing process,* which covers the juxtaposition of takes, shots and scenes, for the rough cut stage.
- *A set of 37 editing strategies*, which introduce schemes for the appropriate visual and cinematographic presentation of a narrative event. The strategies support continuity editing, and focus on the essential narrative aspects, context and form. With these rules the system is able to relate the intention of a narrative to its presentation in shot form. Furthermore, the rules provides means to visually guide the viewer of a video sequence, so that he or she can identify information as relevant or purely descriptive. Moreover, the system is in the position to establish and maintain spatial and temporal continuity over several shots, based partly on the visual decomposition of a character and/or actions, and partly on the decomposition of spatial relationships between characters and between characters and their screen positions. Finally, the system can shape the temporal rhythm of a sequence by means of physical clipping.

The AUTEUR system and its theoretical foundation are best regarded as a platform that demonstrates the feasibility of automated thematic film editing in restricted, yet complex, domains. The system demonstrates how automated film editing can support automated help, and, despite the implementational complexity of the system, that the extracted principles will also provide a source of help to designers and workers in other related fields.

The structured textual approach to the representation of video content provides an objective representation and thus does not restrict possible connotative meanings of the material. At this moment in time, the structures provided appear to be sufficiently rich to describe complex film specific features, and the denotative aspects of film, though some of the structures are rudimentary, such as the representation of gestures and the representation of groups of subjects. In theory, the current representation enables the annotation of video material of arbitrary length and content. It should be noted that the notion of arbitrary re-use of video material is illusory, since in general, the computational effort involved in comparing large numbers of highly varied content based descriptions to establish continuity would result in an unacceptable degradation of system performance. However, if we focus on *domain dependent applications*, a suitable selection of material would reduce this complexity. For example, we may ensure that our database contains only shots in which:

- a small set of characters are found in similar locations,
- the locations are simple,
- the actions are available from different angles, point of views, etc.

Due to the swift development in computer animation, the best approach, however, would be to generate the required sequences with narrative strategies, as described in the AUTEUR system and then generate the required animation sequences as required.

AUTEUR is a research platform, and as such achieves limited success in automatically generating video material to suggest a given theme. The current version of AUTEUR produces only a restricted range of humorous scenes, predominantly of the so-called slapstick style, e.g. "slipping on a banana skin". However, AUTEUR achieves this in ways that take account of knowledge of filmic acceptability.

The analysis of the cinematic image shows that a very large number of codes are involved in the meaning of images. The current version of AUTEUR makes use of but a few of these, such as:

- *emotional codes*, which are represented as conceptual structures (e.g. for *rage* or *pleasure*), and their visualisation in the form of gestures of body, face, hands and limbs (e.g. smiling, nodding, pointing), or actions (e.g. whistling as an indication of pleasure);
- *cinematic codes*, as exemplified by the relationship between camera distance and hierarchical representations of subjects, in combination with conceptual relationships between filmic devices and narrative functionality. Additionally, there are the spatial relationships between shot distances;
- *cultural codes*, in that our representation of jokes, editing and story structures, reflect the humour, film and narrative schemes specific to European culture.

Each of the above code systems could be dealt with more extensively. For example, *emotional codes* could be improved by extending our action-gesture-centred approach, used in the representation of video content, with a more complete representational scheme for hand and body gestures. Useful augmentation to the *cinematic codes* would be colour codes (e.g. bright colours support the impression of a good mood, and there is a relationship between colours and certain abstract concepts). Further refinements of the *cinematic codes* could be achieved by including additional cinematographic aspects of video content (e.g. camera angle, or shot contrast) or denotative aspects of video content (e.g. season, structures of objects - form, colour, size, etc.), and linking these to conceptual structures of abstract concepts. It must be stressed that these amendments would not necessitate in great change to the existing representation structures, but would greatly improve AUTEUR's inferencing capabilities.

Spatial and temporal continuity editing is relatively well provided for by the editing scheme. There is a need, however, to improve the temporal aspects of physical editing, and, in particular, to adequately address the influence of the speed of action on the rhythmical structure of a sequence, which is important in slowing and accelerating the pace of a sequence.

The problems regarding the description of media semantics led soon into the question on how to support the instantiation and maintenance of relatively large and complex descriptive structures. As a result the A4SM framework was developed.

A4SM (Nack & Steinmetz 98, Nack and Lindley 2000)

In this following sections we present results of the Mobile Group at GMD-IPSI on the representation of media content during its production, still allowing a later reuse of the gathered meta-information. The 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 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. 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.

The aim of A4SM, 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 the 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 we identified two important phases of the framework that can be supported by IT, i.e. production and post-production

Within the production phase it is the acquisition of material which can be improved by supporting the collaboration between a reporter and a cameraperson. 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 cameraperson is shooting. As a result, we designed and developed an MPEG–7 hard disk camera that automatically stores the acquired video stream together with an associated MPEG-7 description structure of relevant information, such as co-ordinates, camera work and lens movement. Additionally we provide a mobile handheld annotation tool for the reporter 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 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 knowledge elicitation it was mentioned by a great number of reporters that it would be excellent to have a simple editing suite in the form of a lap-top 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.

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 2000a, 2000b, 2000c]. The objective of the MPEG-7 group is to standardise ways of describing different types of multimedia information. The emphasis of the standard is audio-visual content description with the goal of extending the limited capabilities of proprietary solutions in identifying content by providing a set of description schemes (DS) and descriptors (D) for the purpose of making various types of multimedia content accessible. 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 the 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.

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], i.e. similar to the one used in AUTEUR. The usefulness of combining these two approaches results in gaining the temporality of the multi-layered approach without the disadvantage of using keywords, since keywords have been replaced by a structured content representation.

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 handles gaps.

For our news environment we developed a set of 18 episodic and technical description schemes (we came up with our own DSs because at the time of development none of the MPEG-7 DSs or Ds were actually usable).. Each DS is represented in XML-Schema (we decided to use this instead of the DDL, because the DDL was still under development – however, XML Schema is very close, so our schemata would be MPEG-7 compliant).. Once a DS is instantiated it forms a node within the description 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 broadcast (produced), on which channel, etc.

Subjective_c
subjective description of an event, such as comments of the audience

Media_device

media specific technical information about the data, e.g. lens state, camera movement, etc.

Person

persons participating in the production of the clip, such as reporter, cameraperson, 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 on the event

Setting

setting information for an event, such as country, city, place etc.

Archive

archiving value of the news clip according its content and compositing

Access

access right infuse, IPR, rights management of the clip

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 of its temporal and spatial references are absolute, whereas the references in the associated clips, organised in a newsclip-DS, are referential. If a newsClip-DS behaves as a hub (see the right clip within Figure 1), then its temporal and spatial references toward the media are absolute (note that the annotation algorithm is aware of 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 a handheld annotation device.

A **Relation** enables the connection of descriptors within descriptions as well as connections between distinct descriptions. Relations are actually the icing on 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.

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 in image capture parameters. When 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 description value. If the camera capture event executes over 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 1 describes a possible network of A4SM descriptions for news clips, which are represented by the rectangular boxes. Figure 1 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 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 the form of a semantic network (i.e. and instantiated description schema connected via relations) allows the easy creation of new annotations and the relation of 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 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.

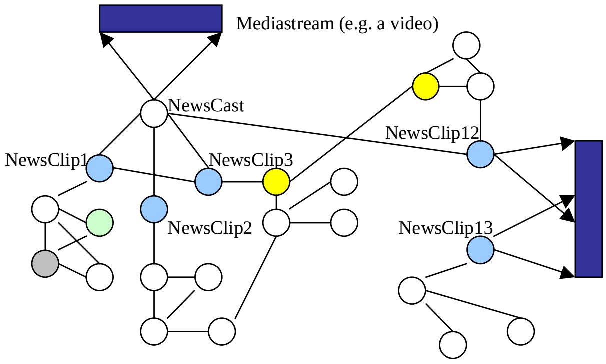

Figure 1: A4SM Description Network and partial parsing approach

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.

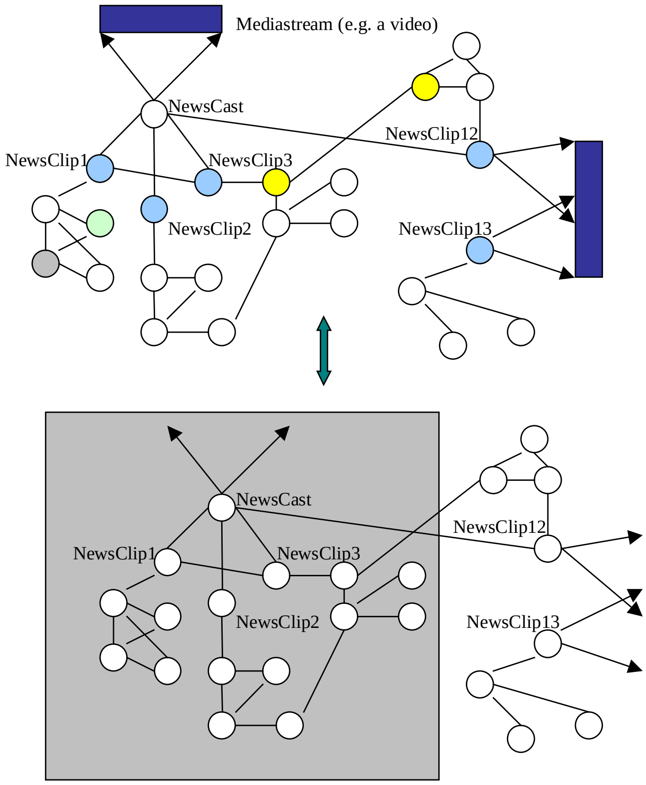

Figure 2: A4SM Description Network and partial parsing approach

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 contents 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 these operations and the created documents. In our opinion this can only be achieved via partial parsing (see also Figure 2). This means the parser validates only a particular part of the network (e.g. a number of hubs). In this way we avoid parsing a complete network if only a tiny section is effected.
We understand that the flexibility of our approach is extending the complexity of maintaining the description structure. Further research has to prove if this is acceptable.

Conclusion

Both projects described above emphasise the requirement for flexible formal annotation mechanisms and structures 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 manually, but also that there is no 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 and secondly also use the creative act to extract the significant syntactic, semantic and semiotic aspects of the content description.
Providing such environments and facilities we are then actually in the position to allow the support of perspective making and perspective taking. ‘Perspective making’ is the process whereby a community develops and strengthens its own knowledge and practice [Boland & Tenkasi 1995]. It is this process that underpins the building of a community’s identity: their basic assumptions, goals, terminology, and modes of discourse. ‘Perspective taking’, on the other hand, refers to the process of trying to engage with another community’s perspective. This can be difficult when their respective way of knowing assumes different agendas or does not match at all. The access and resulting presentations of relevant information in an appropriate stylistic way, i.e. shaped rhythmically and thematically into rich information textures, requires a perspective management in the form of a dynamic and adaptive generation of information presentation. Thus, users can investigate an unknown space provided with the most relevant material and its annotations for the actual moment, allowing a progressing experience of completing the understanding of complex concepts in procedural, and participatory means (i.e. interactive and investigative in a navigable encyclopaedic space, providing access to the full temporal and spatial means of the media items). Such an experience yields an understanding of a concept more primal and powerful than any appeal through normal text in a linear logical form.
At present, I am engaged in research on tools and techniques for semi-automated, interactive narrative generation and style oriented generation of multimedia presentations for web environments.

Literature

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

Boland, R.J.J. & Tenkasi, R.V. (1995). *Perspective Making and Perspektive Taking in Communeties of Knowing*. Organization Science, 6 (4), 350–372.

Lévy, P. (1994). L’intelligence collective. Pour une anthropologie du cyberspace. Édition la Découverte, Paris.

MPEG Requirements Group (2000a): *”MPEG-7: Overview V.14”, Doc. ISO/MPEG N3444, MPEG Geneva Meeting, May 2000*

MPEG Requirements Group (2000b) *”MPEG-7Requirements Document V.11”, Doc. ISO/MPEG N3446, MPEG Geneva Meeting, May 2000*

MPEG Requirements Group (2000c) *”MPEG-DDL Working Draft V 2.0”, Doc. ISO/MPEG W3293, MPEG Noordwijkerhout Meeting, April 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. & C. Lindley (2000) “Arbeitsumgebungen für die Entwicklung interaktiver Geschichten”, Workshop on Digital Storytelling, Darmstadt, Germany, 15-16/6/2000.

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.

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/