Published in: Proceedings of the 1994 European Conference of the Society of Motion Picture and Television Engineers. Köln 1994, S. 98-106

Integrating Video with Information Technology - Prospects and Challenges

Ulrich Frank
Institut für Wirtschaftsinformatik, Universität Koblenz
Rheinau1, 56075 Koblenz
Germany

"The transformation we are concerned with is not a technical one, but a continuing evolution of how we understand our surrounding and ourselves ... "

Terry Winograd und Fernando Flores

Abstract

The paper will give an overview of how future multimedia information systems could look like and how they could be produced and maintained. Those systems will no longer make a difference between the handling of traditional data and video or audio. Instead they will focus on conveniently providing a requested information content together with the appropriate presentation - no matter whether it is text, graphics, video or audio. These features however do not come for free. Instead a number of challenges has to be faced. The paper will discuss those challenges and present an evolutionary approach with a number of measures and strategies to overcome them.

1. Introduction

We are at the dawning of a new information age. With increasingly powerful digital computers penetrating more and more private homes and emerging high bandwidth Wide Area Networks the grounds are laid for a new generation of computer applications that will integrate a wide range of media. While this development will certainly result in a vast amount of digitized information it will also provide us with more powerful ways to analyze, manipulate, and reuse information. Thereby we will gain the chance to benefit from (or being annoyed by) new ways of communication, entertainment, teaching, and learning. However, in order to fully exploit the potential of this new technology we will also have to develop new ways of preparing, organizing, and maintaining information.

We will first look at the current situation. Some of the existing multimedia applications are already rather impressive. Nevertheless they are certainly way apart from what future systems will look like - and how they will be produced. In order to give an idea of the added value that will result from integrating audio and video with digital information technology we will characterize some attractive services those systems may offer. Analyzing these features reveals that traditional audio and video material on its own is not well suited as input for computer applications. On an abstract level traditional information system design has been dealing with similar problems. Therefore we will apply well established design principles to the problem of handling and preparing video for serving as integrated parts of future information systems.

2. The Current Situation: Adapting new Technology to traditional Perspectives

Today's multimedia applications can be divided into two main streams. The first stream consists of traditional computer systems extended by capabilities to present photos or sequences of video or audio. Typically those applications are specialized information retrieval systems where certain chunks of information are associated with multimedia presentations. Among the most common examples are sales support systems that allow to retrieve objects a customer may be interested in and present them using photos or videos. The second stream of applications aims at supporting video

(post) production. Operating on digitized video images they provide an impressive range of powerful manipulations which can be used in a convenient way. From our point of view it is remarkable that those applications essentially increase productivity and flexibility of video post production. However, they do not attempt to change the professional approach of how to produce videos (in fact, if they did, they would probably be less successful).

Within the first stream there is one particular system that has been a tremendous success so far and that is still attracting an increasing number of sometimes enthusiastic users - specially within the scientific community: World Wide Web (WWW). WWW gives an impression of what can be accomplished with the computer infrastructure usually available at today's research sites - and in tomorrow's homes. It also gives an idea of the cultural changes that go along with the new information age. WWW is an architecture that lies on top of the Internet. It allows to create hypermedia documents. Such a document consists of nodes which may contain formatted text, images, audio, or video. The document nodes are stored within a flexible number of information servers. Any node may contain references to other nodes - no matter where they are physically located. In order to allow the user to conveniently browse through hypermedia documents spread around the world WWW includes specifications for client software. Meanwhile clients exist for all major platforms. The clients also provide means to add information and to organize it in order to support certain ways of interaction. Information retrieval is fostered by various dictionaries and a small set of full text retrieval capabilities. WWW thereby already provides the functionality that is required for video on demand, although the data exchange rates commonly available are not sufficient. Figure 1 illustrates the user interface of a WWW-client.

With thousands of motivated and skillful users at universities and research sites around the world it was no surprise that it did not take long until a vast amount of information was produced within numerous hypermedia documents. It is remarkable however that the features provided by WWW also fostered a new quality of information access and communication. For instance: research results can be quickly disseminated not only as text. Furthermore they can be annotated with pictures or videos - either of the research topic or people involved in the work. Thereby WWW not only affects the way people deal with information - both in providing and accessing it. Like the Internet in general it also allows for new ways to communicate - by providing guided access to members of certain world wide communities and by fostering less formal ways to interact.

What is the lesson we may learn from the current situation? While systems like WWW certainly give a first glance of future multimedia systems the integration of video and information technology is mainly restricted to a basic technical level: digital representation and compression, synchronization etc. At the same time many computer scientists dealing with multimedia systems concentrate on technical problems like building interfaces to analogous devices or widening performance bottlenecks. On the application level videos are treated as black boxes, or - to use a phrase from database technology - as "BLOBs" (Binary large Objects). Typically computer programmers abstract from the content of a video - while it should be the other way around: concentrating on the content and abstracting from technological constraints. On the other hand video professionals only use information technology to increase the efficiency of traditional production processes. They usually intend to produce neat, but stand alone videos instead of thinking about how to produce a video in order to make it well suited for computer applications.

Fig. 1: User interface of a WWW client session

3. Challenges and Strategies

Computer programs usually allow to read and/or write data. However, this can be done on very different levels. Usually it is desirable for an application to provide the user with concepts that fit his perception and conceptualization of the domain that is represented within the application. The more a designer of an application succeeds in accomplishing this goal the better are the chances to build user friendly programs. With familiar concepts being mapped to the application information can be retrieved or manipulated by directly applying the associations a user has in mind when he is thinking about the relevant subject.

Enhancing video with application domain concepts opens a wide range of features - on different levels of complexity. The following examples illustrate some of the services that could be provided:

- retrieve all videos that feature sport events
- retrieve all videos that feature ball games
- retrieve all movie dramas from 1984 starring Robert de Niro
- retrieve all tennis matches where a particular player lost the first set but finally won the match
- retrieve the movie that contains the line “Do you feel lucky?”
- retrieve the movie and the particular scene that contains the line: “Go ahead make my day”.
- retrieve the goal Germany scored in its loss against Bulgaria during the Championship in 1994
- retrieve the murder scene that happened after the protagonist was released from jail
- retrieve all movies that contain sequences of baseball games
- retrieve all TV shows where film previews were presented
- retrieve all movies where a male protagonist kills his lover
- retrieve all soccer games in 1994 where a defender scores a goal after he had received the ball from another defender
- in a particular movie replace the sequence showing a soccer game with a tennis match.
- within the movie “In the Line of Fire” substitute Clint Eastwood with John Wayne.

Apparently some of these services could already be provided by today’s applications. Other services seem to be harder to accomplish. There are two main challenges to accomplish applications that could handle requests like those exemplified above. The first challenge has already been mentioned: applications that operate on videos should have access to concepts describing the content of these videos. In order to be more precise we could also say: It is desirable to provide as much formalized semantics with a video as the user could have in mind for his requests. The semantic content of a representation depends on the formal interpretations it allows for: the more interpretations are excluded the higher the level of semantics. For instance: If you look at a video represented only as a sequence of byte arrays (each of which representing an image) together with a synchronized data stream for audio this video might contain anything - in other words: It does not contain much semantics. In order to overcome this problem you have to explicitly enrich a video with semantics. The strategies discussed below demonstrate how this approach can be pursued in more or less ambitious ways - according to principles well known from information system design.

3.1 Video as a Black Box: Attributes and Annotations

If an application does not know anything about the way information is represented within a video there are two ways to include video into information processing. The first approach treats videos as attributes of objects that are represented in data models. For instance: An actor who is listed in a TV station’s database may be assigned the attribute "example monologue" that is actually a video sequence featuring the actor. While this approach may be appropriate for enhancing given databases with additional presentations it does not directly relate to the content of a video. A well known approach to allow for operations regarding the content of something that itself is not directly represented in a data model is to add textual annotations. To enhance the above example with annotations we could use extra attributes to hold keywords that inform about the content of the video.

If you look at videos with a complex content the annotation could be something like a comprehensive natural language description which in turn could be operated on by a full text retrieval mech-

anism. However, if you provide such a description without further structuring for the whole video it only helps to find a particular video, not a sequence within it. For this reason one could divide a video in a number of sequences of an appropriate duration and assign each sequence a textual description. Most text retrieval or database systems do not include a notion of time. However, there are some prototypical systems that feature retrieval languages which allow for temporal comparisons (like "the sequence *before* sequence x") [like Dean/McDermott or Sripada].

3.2 Classification, Generalization and Specialization

One of the essential principles not only of designing information system but also of systematic real world descriptions in general is classification: You do not intend to solely describe certain instances of the real world. Instead you rather try to define features that are common to a set of instances. In other words: You build classes (or concepts) like the class "sport event". The features all instances have in common may be attributes (like "starting time", "duration", etc.) or services you expect the instances to provide (like "show the last n minutes"). If a number of classes has a common set of features it is a good idea to extract these features into a more general class. For instance: If different video types like "sport event", "drama", etc. all have common attributes like "duration" you could introduce a general class "video". This principle is called "generalization". Its advantage is obvious: It allows to avoid redundant specifications thereby fostering the maintenance of class descriptions. Furthermore it is the prerequisite for applying more sophisticated retrieval operations on a given description: A request can always be related to the most general concept that is appropriate. It will then implicitly be applied (by logical deduction) to the less general concepts. On the other hand you may realize that a given class specification (like "sport event") is not sufficient for a specific case (for instance "tennis match"). Specialization offers a measure to use a given specification and enhance it by additional features. For instance: "tennis match" would inherit all features defined for "sport event" and would additionally have one or more specific features.

Fig. 2 shows an example of how to apply these fundamental principals to video objects.

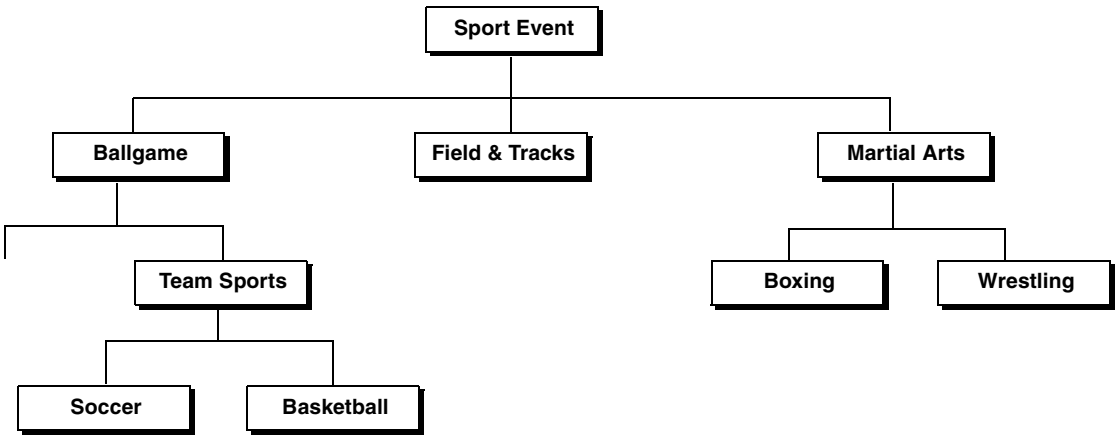

Fig. 2: Classes of sport videos

3.3 Logical Structure

While classification, generalization and specialization are an important step to treat videos as information objects they do not directly provide means to describe the logical structure of a video. Modelling the logical structure enhances the range of meaningful operations - both for read - and

write-access. For instance: You could focus on the second game of the second set of a tennis match. At the same time it is a prerequisite for more sophisticated ways of managing stored videos. For instance: If a logical part of a video uses a part of another video or audio document you would not have to copy this part. Instead there could be a reference to this part of the other source - thereby reducing redundancy and fostering maintenance.

Since it would require a film director to develop a meaningful logical structure for a certain class of films we look at a similar domain: In the area of document management there are a number of logical models, some of them even subject of international standards [see for instance Appelt]. Figure 3 exemplifies how documents of a certain class could be structured.

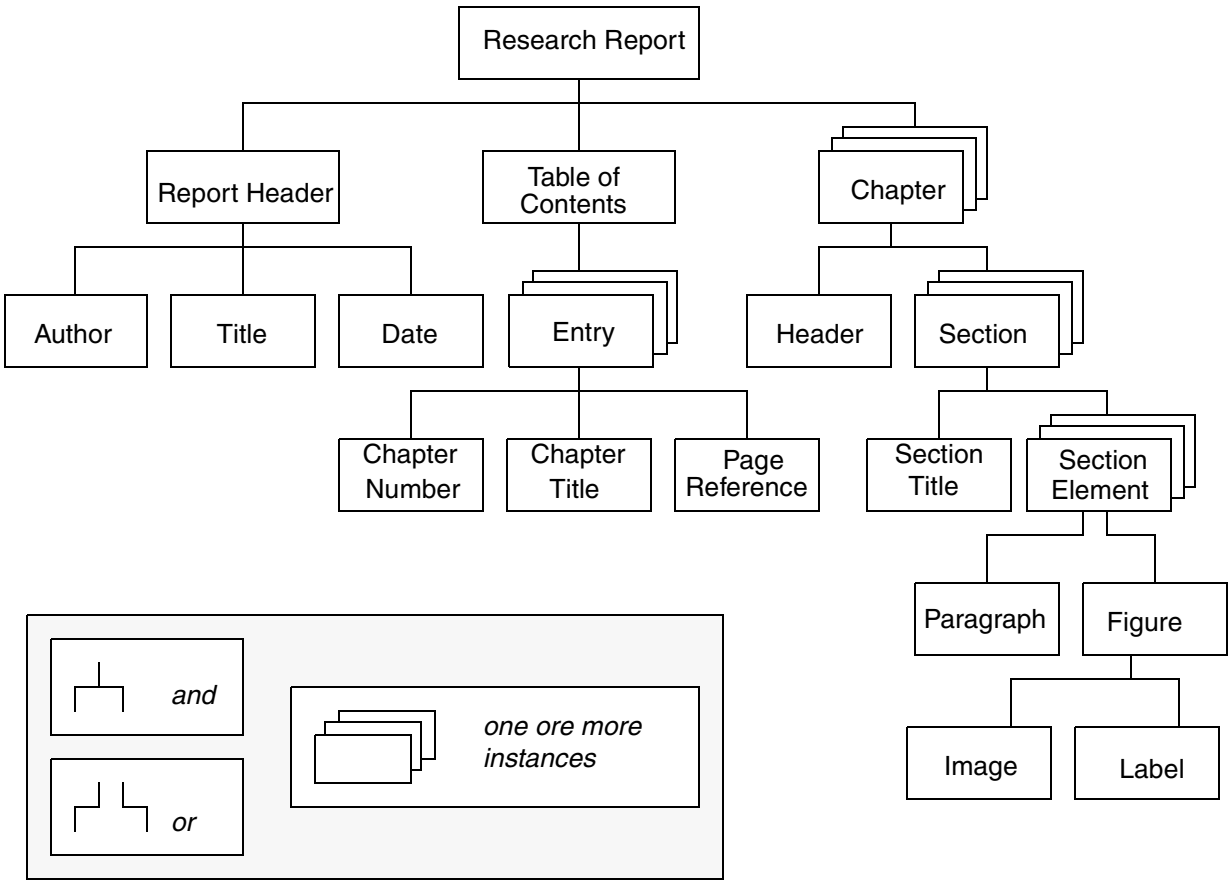

Fig. 3: Logical architecture of compound documents

3.4 Extended Semantic Modelling

The measures proposed so far would not be sufficient to satisfy all of the requests listed above. Those cases require a more detailed description of the relevant concepts. In other words: We need models of the video's subject that incorporate the semantics required to handle the requests. Such models are basically conceptual descriptions. There are a number of ways to structure these descriptions. Traditional data models like the Entity Relationship Model [see Chen] use objects (respectively entities) and relationships between objects. Classes of objects (entity types) are described by a set of attributes. Such approaches are not well suited for the purpose of semantically modelling the content of videos because of their limited expressive capabilities. Research in Artificial Intelligence produces modelling methodologies which are more appropriate. Schank and Abelson for instance designed a formal language to script real life scenes - like visiting a restaurant - in order to build programs that could answer questions regarding these scenes. They suggest two

basic conceptualizations to describe a scene, an active conceptualization ("Actor Action Object Direction") and a stative conceptualization ("Object is in State with value"). Winston introduced an approach to model Shakespearean tragedies in order to recognize situations which are analogies to a given one.

Special Artificial Intelligence approaches would certainly be well suited for modelling the content of videos. However, they are rather exotic. Usually they are not well documented, there are no robust tools to support them, and only few people are familiar with them. Another approach, which was also inspired by Artificial Intelligence research, has gained enormous attention during the last years: object-oriented modelling and implementation is going to be the leading paradigm for future software engineering. Not only that it provides means to model a video's content on a high level of semantic, there are also text books [for instance Booch, Rumbaugh et al.], tools, programming languages and special Database Management Systems available. The last aspect is of outstanding importance because it means tremendous help with implementing an actual application. Furthermore standards for object-oriented technologies are currently emerging. In principle object-oriented modelling suggests to describe an application domain in terms of objects and relationships between objects. Objects are grouped into classes. Generalization and specialization can be applied. Each class is characterized by a set of attributes and services.

Figure 4 gives an example of an object-oriented semantic model with one class described in more detail.

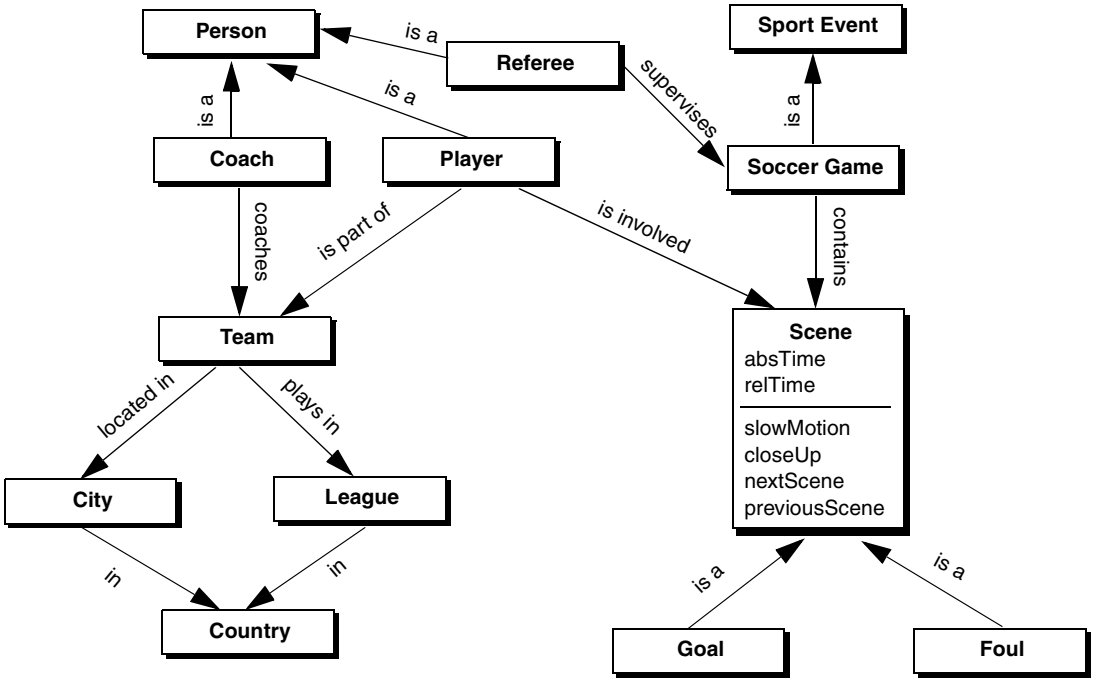

Fig. 4: Example of an object model for sport events with a specification of one class

Object-oriented models allow to describe application domains using concepts that correspond closely to the conceptualizations preferred by the relevant users. At the same time they are based on a solid software technological foundation which fosters implementation and maintenance. One problem however remains to be dealt with in every single modelling project: How to identify the relevant objects? While Meyer, one of the protagonists of object-oriented software development, states "The objects are there for the picking" it is a matter of fact that merely asking the domain

experts for objects or concepts usually does not result in a comprehensive model. Instead it is usually recommended to use a heuristic approach that helps with identifying the relevant objects. There are two promising heuristics. One is to concentrate on processes - which is always preferable when processes are a preferred way to describe a domain. For instance: If you want to design an application for an editor at a TV station you could ask him for the relevant tasks or processes he is involved in. Thereafter you would decompose the tasks or processes into smaller units and then ask for the information objects that are required there. The other heuristics is scripting: By interviewing people or using existing documents you get a script of the relevant domain. This script can then be preprocessed in order to produce a list of nouns (candidates for objects) and predicates (candidates for relationships), which would then be used as the input for designing an object model.

3.5 Instantiation as the Second Major Challenge

Even if the - sometimes tremendous - effort to design comprehensive semantic models can be accomplished there is still one major problem to be solved. A model is an abstraction of a specific case. This is for a good reason: We talk about concepts rather than about instances. However, when it comes to derive a specific application from a model we look at instances: Looking at a video on a particular soccer game we are not only interested in the concepts "player" or "goal" but also in the specific instances. While deriving instances from concepts is a well known procedure - called instantiation - for data processing it marks an outstanding challenge for applications including videos.

In principle there would be the chance to apply sophisticated pattern matching and natural language recognizing algorithms to analyze a video: The patterns would have to be identified as instances of concepts defined in a related semantic model. For instance: A foul in a soccer game. Furthermore it would be desirable to identify the state of this instance. For example: Where on the field did the foul occur? Experience gathered in Artificial Intelligence research on less complex tasks however indicates, that such an approach will only lead to poor (if any) results and is extremely expensive at the same time:

> "I recognized the depth of the difficulties in getting a machine to understand language in any but a superficial and misleading way, and am convinced that people will be much better served by machines that do well-defined and understandable things that those that appear to be like persons until something goes wrong (which won´t take long), at which point there is only confusion."
>
> *Winograd (in Bobrow/Hayes)*

Instead it seems to be much more promising to adapt the production of videos to the needs of multimedia applications. In other words: It is easier and safer to reduce ambiguity by appropriate measures than to apply automatic procedures to resolve it. How could such an approach look like? For all videos which base on events organized for being filmed (like movies or sport events) there is the chance for electronically marking at least some of the relevant instances. Marking means to allow for an affordable and save automatic detection of the instances. That does not have to affect the way a human viewer perceives the video. For instance: Actors in a play could be marked as instances of certain classes (like detective, hero, lover, father, mother, etc.). In order to characterize the relationships between the figures the script that has to be prepared anyway could be enhanced to express important relationships in a way that they would allow for automatic interpretation.[1] A similar approach could be applied to sport events. For instance: The various players in a soccer game could be marked as well as segments of the field.

1. It is no doubt that this would certainly require a major cultural change - and that many people are probably not fond of this idea at all. It is remarkable however that some companies have already adapted the process of writing technical documentation to the requirements of language parsing procedures.

4. Concluding Remarks

For a new generation of multimedia applications to become reality a number of technological challenges have to be overcome. However, the crucial challenges are of another kind: Only if artists, video experts, and computer scientists succeed in cooperating on a conceptual level, the synergy necessary for redesigning current production processes will emerge - without the risk to reinvent the wheel every other day. Cooperation requires communication which in turn stresses the importance of models that foster a common understanding of the essential problems. Therefore the emphasis has to be on conceptual modelling, not on implementation details which are not stable anyway.

Furthermore it is evident that we do not only face an intellectual/cultural challenge: Not everything that could be done, will be affordable. This is the case for semantic modelling as well as for instantiation. Similar to software we already have today, the effort spent for a particular system very much depends on the economy of scale: Since copying is cheap, the price mainly depends on the size of the market. So we will probably encounter both rather simple systems and systems that caused a high amount of production costs but are sold on mass markets.

References

Appelt, W.: Document Architecture in Open Systems. The ODA Standard. Berlin, Heidelberg, New York etc. 1991

Bobrow, D.G.; Hayes, P.J.: Artificial Intelligence - Where Are We? In: Artificial Intelligence 25, 1985, pp. 375-415

Booch, G.: Object-Oriented Analysis and Design with Applications. Redwood City 1994

Chen, P.P.: The Entity-Relationship Model: Towards a Unified View of Data. In: ACM TODS, Vol. 1, No. 1, 1976, pp. 9-36

Dean, T.; McDermott, D.V.: Temporal Data Base Management. In: Artificial Intelligence, Vol. 32, 1987, pp. 1-55

Loomis, M.E.S.: OODBMS - The Basics. In: Journal of Object-Oriented Programming. Vol. 3, No. 1, 1990, pp. 77-88

Meyer, B.: Object-Oriented Software Construction. Englewood Cliffs, N.J. 1988

Rumbaugh, J. et al.: Object-oriented modeling and design. Englewood Cliffs, N.J. 1991

Schank, R.; Abelson, R.: Scripts, Plans, Goals and Understanding. Hillsdale 1975

Sripada, S.M.: A logical framework for temporal deductive databases. In: Banchilhon, F.; De Witt, D.J. (Hg.): Proceedings of the 14th International Conference on Very Large Database Systems. Los Altos 1988, pp. 171-182

Winograd, T.; Flores, F.: Understanding Computers and Cognition - a new Foundation for Design. Norwood, N.J. 1986

Winston, P.H.: Learning and Reasoning by Analogy. In: Communications of the ACM. Vol. 23, 1980, pp. 689-703

Published in: Proceedings of the 1994 European Conference of the Society of Motion Picture and Television Engineers. Köln 1994, S. 98-106

Integrating Video with Information Technology - Prospects and Challenges

Ulrich Frank
Institut für Wirtschaftsinformatik, Universität Koblenz
Rheinau1, 56075 Koblenz
Germany

"The transformation we are concerned with is not a technical one, but a continuing evolution of how we understand our surrounding and ourselves ... "

Terry Winograd und Fernando Flores

Abstract

The paper will give an overview of how future multimedia information systems could look like and how they could be produced and maintained. Those systems will no longer make a difference between the handling of traditional data and video or audio. Instead they will focus on conveniently providing a requested information content together with the appropriate presentation - no matter whether it is text, graphics, video or audio. These features however do not come for free. Instead a number of challenges has to be faced. The paper will discuss those challenges and present an evolutionary approach with a number of measures and strategies to overcome them.

1. Introduction

We are at the dawning of a new information age. With increasingly powerful digital computers penetrating more and more private homes and emerging high bandwidth Wide Area Networks the grounds are laid for a new generation of computer applications that will integrate a wide range of media. While this development will certainly result in a vast amount of digitized information it will also provide us with more powerful ways to analyze, manipulate, and reuse information. Thereby we will gain the chance to benefit from (or being annoyed by) new ways of communication, entertainment, teaching, and learning. However, in order to fully exploit the potential of this new technology we will also have to develop new ways of preparing, organizing, and maintaining information.

We will first look at the current situation. Some of the existing multimedia applications are already rather impressive. Nevertheless they are certainly way apart from what future systems will look like - and how they will be produced. In order to give an idea of the added value that will result from integrating audio and video with digital information technology we will characterize some attractive services those systems may offer. Analyzing these features reveals that traditional audio and video material on its own is not well suited as input for computer applications. On an abstract level traditional information system design has been dealing with similar problems. Therefore we will apply well established design principles to the problem of handling and preparing video for serving as integrated parts of future information systems.

2. The Current Situation: Adapting new Technology to traditional Perspectives

Today's multimedia applications can be divided into two main streams. The first stream consists of traditional computer systems extended by capabilities to present photos or sequences of video or audio. Typically those applications are specialized information retrieval systems where certain chunks of information are associated with multimedia presentations. Among the most common examples are sales support systems that allow to retrieve objects a customer may be interested in and present them using photos or videos. The second stream of applications aims at supporting video

(post) production. Operating on digitized video images they provide an impressive range of powerful manipulations which can be used in a convenient way. From our point of view it is remarkable that those applications essentially increase productivity and flexibility of video post production. However, they do not attempt to change the professional approach of how to produce videos (in fact, if they did, they would probably be less successful).

Within the first stream there is one particular system that has been a tremendous success so far and that is still attracting an increasing number of sometimes enthusiastic users - specially within the scientific community: World Wide Web (WWW). WWW gives an impression of what can be accomplished with the computer infrastructure usually available at today's research sites - and in tomorrow's homes. It also gives an idea of the cultural changes that go along with the new information age. WWW is an architecture that lies on top of the Internet. It allows to create hypermedia documents. Such a document consists of nodes which may contain formatted text, images, audio, or video. The document nodes are stored within a flexible number of information servers. Any node may contain references to other nodes - no matter where they are physically located. In order to allow the user to conveniently browse through hypermedia documents spread around the world WWW includes specifications for client software. Meanwhile clients exist for all major platforms. The clients also provide means to add information and to organize it in order to support certain ways of interaction. Information retrieval is fostered by various dictionaries and a small set of full text retrieval capabilities. WWW thereby already provides the functionality that is required for video on demand, although the data exchange rates commonly available are not sufficient. Figure 1 illustrates the user interface of a WWW-client.

With thousands of motivated and skillful users at universities and research sites around the world it was no surprise that it did not take long until a vast amount of information was produced within numerous hypermedia documents. It is remarkable however that the features provided by WWW also fostered a new quality of information access and communication. For instance: research results can be quickly disseminated not only as text. Furthermore they can be annotated with pictures or videos - either of the research topic or people involved in the work. Thereby WWW not only affects the way people deal with information - both in providing and accessing it. Like the Internet in general it also allows for new ways to communicate - by providing guided access to members of certain world wide communities and by fostering less formal ways to interact.

What is the lesson we may learn from the current situation? While systems like WWW certainly give a first glance of future multimedia systems the integration of video and information technology is mainly restricted to a basic technical level: digital representation and compression, synchronization etc. At the same time many computer scientists dealing with multimedia systems concentrate on technical problems like building interfaces to analogous devices or widening performance bottlenecks. On the application level videos are treated as black boxes, or - to use a phrase from database technology - as "BLOBs" (Binary large Objects). Typically computer programmers abstract from the content of a video - while it should be the other way around: concentrating on the content and abstracting from technological constraints. On the other hand video professionals only use information technology to increase the efficiency of traditional production processes. They usually intend to produce neat, but stand alone videos instead of thinking about how to produce a video in order to make it well suited for computer applications.

Fig. 1: User interface of a WWW client session

3. Challenges and Strategies

Computer programs usually allow to read and/or write data. However, this can be done on very different levels. Usually it is desirable for an application to provide the user with concepts that fit his perception and conceptualization of the domain that is represented within the application. The more a designer of an application succeeds in accomplishing this goal the better are the chances to build user friendly programs. With familiar concepts being mapped to the application information can be retrieved or manipulated by directly applying the associations a user has in mind when he is thinking about the relevant subject.

Enhancing video with application domain concepts opens a wide range of features - on different levels of complexity. The following examples illustrate some of the services that could be provided:

- retrieve all videos that feature sport events
- retrieve all videos that feature ball games
- retrieve all movie dramas from 1984 starring Robert de Niro
- retrieve all tennis matches where a particular player lost the first set but finally won the match
- retrieve the movie that contains the line “Do you feel lucky?”
- retrieve the movie and the particular scene that contains the line: “Go ahead make my day”.
- retrieve the goal Germany scored in its loss against Bulgaria during the Championship in 1994
- retrieve the murder scene that happened after the protagonist was released from jail
- retrieve all movies that contain sequences of baseball games
- retrieve all TV shows where film previews were presented
- retrieve all movies where a male protagonist kills his lover
- retrieve all soccer games in 1994 where a defender scores a goal after he had received the ball from another defender
- in a particular movie replace the sequence showing a soccer game with a tennis match.
- within the movie “In the Line of Fire” substitute Clint Eastwood with John Wayne.

Apparently some of these services could already be provided by today’s applications. Other services seem to be harder to accomplish. There are two main challenges to accomplish applications that could handle requests like those exemplified above. The first challenge has already been mentioned: applications that operate on videos should have access to concepts describing the content of these videos. In order to be more precise we could also say: It is desirable to provide as much formalized semantics with a video as the user could have in mind for his requests. The semantic content of a representation depends on the formal interpretations it allows for: the more interpretations are excluded the higher the level of semantics. For instance: If you look at a video represented only as a sequence of byte arrays (each of which representing an image) together with a synchronized data stream for audio this video might contain anything - in other words: It does not contain much semantics. In order to overcome this problem you have to explicitly enrich a video with semantics. The strategies discussed below demonstrate how this approach can be pursued in more or less ambitious ways - according to principles well known from information system design.

3.1 Video as a Black Box: Attributes and Annotations

If an application does not know anything about the way information is represented within a video there are two ways to include video into information processing. The first approach treats videos as attributes of objects that are represented in data models. For instance: An actor who is listed in a TV station’s database may be assigned the attribute "example monologue" that is actually a video sequence featuring the actor. While this approach may be appropriate for enhancing given databases with additional presentations it does not directly relate to the content of a video. A well known approach to allow for operations regarding the content of something that itself is not directly represented in a data model is to add textual annotations. To enhance the above example with annotations we could use extra attributes to hold keywords that inform about the content of the video.

If you look at videos with a complex content the annotation could be something like a comprehensive natural language description which in turn could be operated on by a full text retrieval mech-

anism. However, if you provide such a description without further structuring for the whole video it only helps to find a particular video, not a sequence within it. For this reason one could divide a video in a number of sequences of an appropriate duration and assign each sequence a textual description. Most text retrieval or database systems do not include a notion of time. However, there are some prototypical systems that feature retrieval languages which allow for temporal comparisons (like "the sequence *before* sequence x") [like Dean/McDermott or Sripada].

3.2 Classification, Generalization and Specialization

One of the essential principles not only of designing information system but also of systematic real world descriptions in general is classification: You do not intend to solely describe certain instances of the real world. Instead you rather try to define features that are common to a set of instances. In other words: You build classes (or concepts) like the class "sport event". The features all instances have in common may be attributes (like "starting time", "duration", etc.) or services you expect the instances to provide (like "show the last n minutes"). If a number of classes has a common set of features it is a good idea to extract these features into a more general class. For instance: If different video types like "sport event", "drama", etc. all have common attributes like "duration" you could introduce a general class "video". This principle is called "generalization". Its advantage is obvious: It allows to avoid redundant specifications thereby fostering the maintenance of class descriptions. Furthermore it is the prerequisite for applying more sophisticated retrieval operations on a given description: A request can always be related to the most general concept that is appropriate. It will then implicitly be applied (by logical deduction) to the less general concepts. On the other hand you may realize that a given class specification (like "sport event") is not sufficient for a specific case (for instance "tennis match"). Specialization offers a measure to use a given specification and enhance it by additional features. For instance: "tennis match" would inherit all features defined for "sport event" and would additionally have one or more specific features.

Fig. 2 shows an example of how to apply these fundamental principals to video objects.

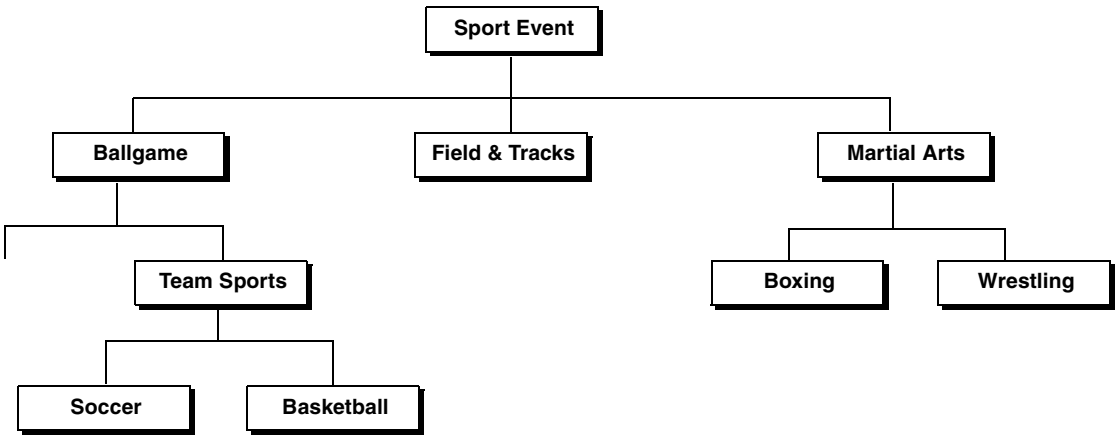

Fig. 2: Classes of sport videos

3.3 Logical Structure

While classification, generalization and specialization are an important step to treat videos as information objects they do not directly provide means to describe the logical structure of a video. Modelling the logical structure enhances the range of meaningful operations - both for read - and

write-access. For instance: You could focus on the second game of the second set of a tennis match. At the same time it is a prerequisite for more sophisticated ways of managing stored videos. For instance: If a logical part of a video uses a part of another video or audio document you would not have to copy this part. Instead there could be a reference to this part of the other source - thereby reducing redundancy and fostering maintenance.

Since it would require a film director to develop a meaningful logical structure for a certain class of films we look at a similar domain: In the area of document management there are a number of logical models, some of them even subject of international standards [see for instance Appelt]. Figure 3 exemplifies how documents of a certain class could be structured.

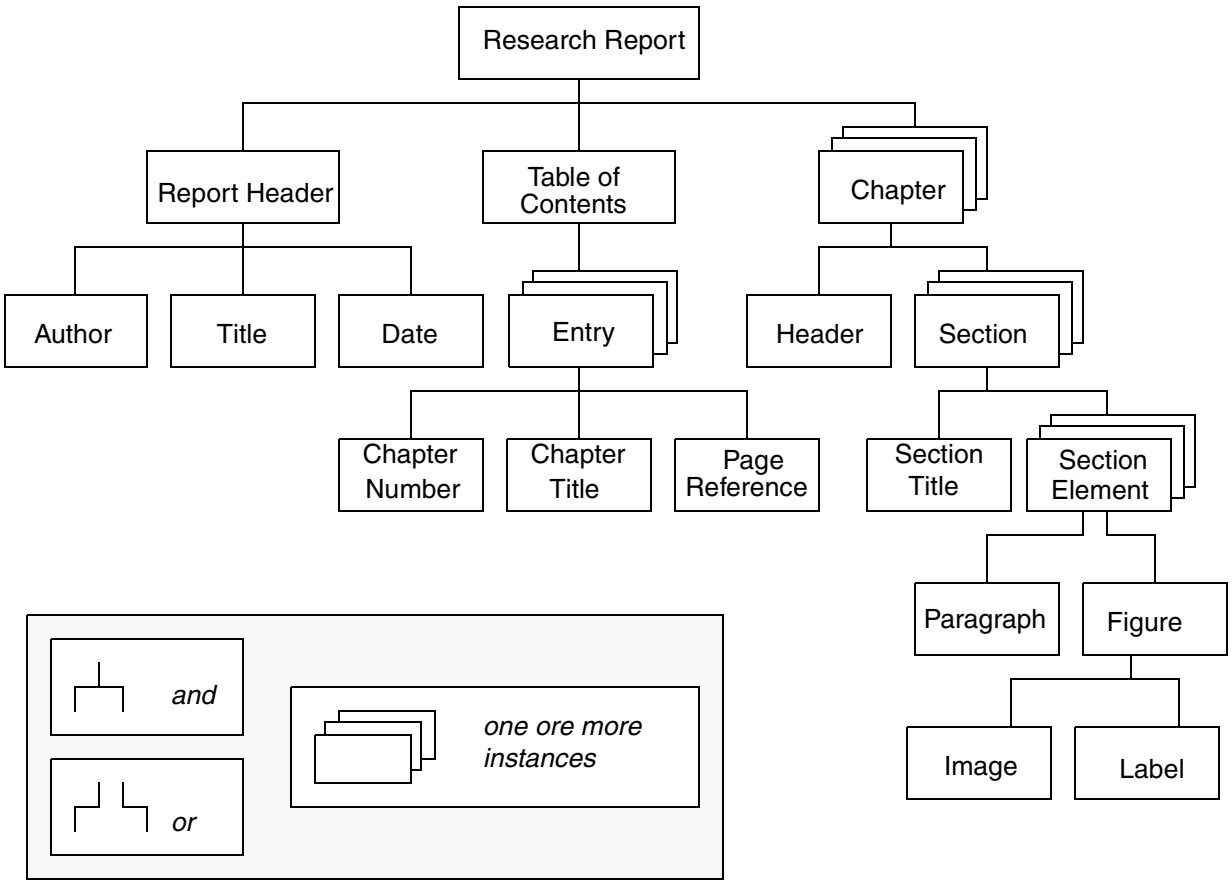

Fig. 3: Logical architecture of compound documents

3.4 Extended Semantic Modelling

The measures proposed so far would not be sufficient to satisfy all of the requests listed above. Those cases require a more detailed description of the relevant concepts. In other words: We need models of the video's subject that incorporate the semantics required to handle the requests. Such models are basically conceptual descriptions. There are a number of ways to structure these descriptions. Traditional data models like the Entity Relationship Model [see Chen] use objects (respectively entities) and relationships between objects. Classes of objects (entity types) are described by a set of attributes. Such approaches are not well suited for the purpose of semantically modelling the content of videos because of their limited expressive capabilities. Research in Artificial Intelligence produces modelling methodologies which are more appropriate. Schank and Abelson for instance designed a formal language to script real life scenes - like visiting a restaurant - in order to build programs that could answer questions regarding these scenes. They suggest two

basic conceptualizations to describe a scene, an active conceptualization ("Actor Action Object Direction") and a stative conceptualization ("Object is in State with value"). Winston introduced an approach to model Shakespearean tragedies in order to recognize situations which are analogies to a given one.

Special Artificial Intelligence approaches would certainly be well suited for modelling the content of videos. However, they are rather exotic. Usually they are not well documented, there are no robust tools to support them, and only few people are familiar with them. Another approach, which was also inspired by Artificial Intelligence research, has gained enormous attention during the last years: object-oriented modelling and implementation is going to be the leading paradigm for future software engineering. Not only that it provides means to model a video's content on a high level of semantic, there are also text books [for instance Booch, Rumbaugh et al.], tools, programming languages and special Database Management Systems available. The last aspect is of outstanding importance because it means tremendous help with implementing an actual application. Furthermore standards for object-oriented technologies are currently emerging. In principle object-oriented modelling suggests to describe an application domain in terms of objects and relationships between objects. Objects are grouped into classes. Generalization and specialization can be applied. Each class is characterized by a set of attributes and services.

Figure 4 gives an example of an object-oriented semantic model with one class described in more detail.

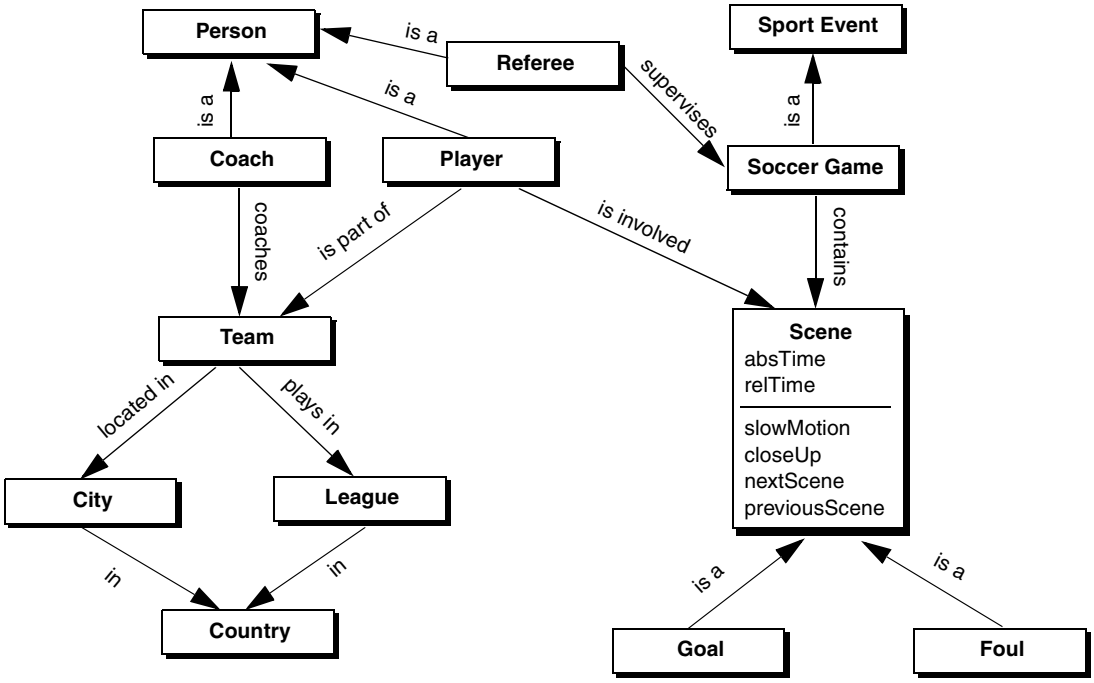

Fig. 4: Example of an object model for sport events with a specification of one class

Object-oriented models allow to describe application domains using concepts that correspond closely to the conceptualizations preferred by the relevant users. At the same time they are based on a solid software technological foundation which fosters implementation and maintenance. One problem however remains to be dealt with in every single modelling project: How to identify the relevant objects? While Meyer, one of the protagonists of object-oriented software development, states "The objects are there for the picking" it is a matter of fact that merely asking the domain

experts for objects or concepts usually does not result in a comprehensive model. Instead it is usually recommended to use a heuristic approach that helps with identifying the relevant objects. There are two promising heuristics. One is to concentrate on processes - which is always preferable when processes are a preferred way to describe a domain. For instance: If you want to design an application for an editor at a TV station you could ask him for the relevant tasks or processes he is involved in. Thereafter you would decompose the tasks or processes into smaller units and then ask for the information objects that are required there. The other heuristics is scripting: By interviewing people or using existing documents you get a script of the relevant domain. This script can then be preprocessed in order to produce a list of nouns (candidates for objects) and predicates (candidates for relationships), which would then be used as the input for designing an object model.

3.5 Instantiation as the Second Major Challenge

Even if the - sometimes tremendous - effort to design comprehensive semantic models can be accomplished there is still one major problem to be solved. A model is an abstraction of a specific case. This is for a good reason: We talk about concepts rather than about instances. However, when it comes to derive a specific application from a model we look at instances: Looking at a video on a particular soccer game we are not only interested in the concepts "player" or "goal" but also in the specific instances. While deriving instances from concepts is a well known procedure - called instantiation - for data processing it marks an outstanding challenge for applications including videos.

In principle there would be the chance to apply sophisticated pattern matching and natural language recognizing algorithms to analyze a video: The patterns would have to be identified as instances of concepts defined in a related semantic model. For instance: A foul in a soccer game. Furthermore it would be desirable to identify the state of this instance. For example: Where on the field did the foul occur? Experience gathered in Artificial Intelligence research on less complex tasks however indicates, that such an approach will only lead to poor (if any) results and is extremely expensive at the same time:

> "I recognized the depth of the difficulties in getting a machine to understand language in any but a superficial and misleading way, and am convinced that people will be much better served by machines that do well-defined and understandable things that those that appear to be like persons until something goes wrong (which won´t take long), at which point there is only confusion."
>
> *Winograd (in Bobrow/Hayes)*

Instead it seems to be much more promising to adapt the production of videos to the needs of multimedia applications. In other words: It is easier and safer to reduce ambiguity by appropriate measures than to apply automatic procedures to resolve it. How could such an approach look like? For all videos which base on events organized for being filmed (like movies or sport events) there is the chance for electronically marking at least some of the relevant instances. Marking means to allow for an affordable and save automatic detection of the instances. That does not have to affect the way a human viewer perceives the video. For instance: Actors in a play could be marked as instances of certain classes (like detective, hero, lover, father, mother, etc.). In order to characterize the relationships between the figures the script that has to be prepared anyway could be enhanced to express important relationships in a way that they would allow for automatic interpretation.[1] A similar approach could be applied to sport events. For instance: The various players in a soccer game could be marked as well as segments of the field.

1. It is no doubt that this would certainly require a major cultural change - and that many people are probably not fond of this idea at all. It is remarkable however that some companies have already adapted the process of writing technical documentation to the requirements of language parsing procedures.

4. Concluding Remarks

For a new generation of multimedia applications to become reality a number of technological challenges have to be overcome. However, the crucial challenges are of another kind: Only if artists, video experts, and computer scientists succeed in cooperating on a conceptual level, the synergy necessary for redesigning current production processes will emerge - without the risk to reinvent the wheel every other day. Cooperation requires communication which in turn stresses the importance of models that foster a common understanding of the essential problems. Therefore the emphasis has to be on conceptual modelling, not on implementation details which are not stable anyway.

Furthermore it is evident that we do not only face an intellectual/cultural challenge: Not everything that could be done, will be affordable. This is the case for semantic modelling as well as for instantiation. Similar to software we already have today, the effort spent for a particular system very much depends on the economy of scale: Since copying is cheap, the price mainly depends on the size of the market. So we will probably encounter both rather simple systems and systems that caused a high amount of production costs but are sold on mass markets.

References

Appelt, W.: Document Architecture in Open Systems. The ODA Standard. Berlin, Heidelberg, New York etc. 1991

Bobrow, D.G.; Hayes, P.J.: Artificial Intelligence - Where Are We? In: Artificial Intelligence 25, 1985, pp. 375-415

Booch, G.: Object-Oriented Analysis and Design with Applications. Redwood City 1994

Chen, P.P.: The Entity-Relationship Model: Towards a Unified View of Data. In: ACM TODS, Vol. 1, No. 1, 1976, pp. 9-36

Dean, T.; McDermott, D.V.: Temporal Data Base Management. In: Artificial Intelligence, Vol. 32, 1987, pp. 1-55

Loomis, M.E.S.: OODBMS - The Basics. In: Journal of Object-Oriented Programming. Vol. 3, No. 1, 1990, pp. 77-88

Meyer, B.: Object-Oriented Software Construction. Englewood Cliffs, N.J. 1988

Rumbaugh, J. et al.: Object-oriented modeling and design. Englewood Cliffs, N.J. 1991

Schank, R.; Abelson, R.: Scripts, Plans, Goals and Understanding. Hillsdale 1975

Sripada, S.M.: A logical framework for temporal deductive databases. In: Banchilhon, F.; De Witt, D.J. (Hg.): Proceedings of the 14th International Conference on Very Large Database Systems. Los Altos 1988, pp. 171-182

Winograd, T.; Flores, F.: Understanding Computers and Cognition - a new Foundation for Design. Norwood, N.J. 1986

Winston, P.H.: Learning and Reasoning by Analogy. In: Communications of the ACM. Vol. 23, 1980, pp. 689-703

Published in: Proceedings of the 1994 European Conference of the Society of Motion Picture and Television Engineers. Köln 1994, S. 98-106

Integrating Video with Information Technology - Prospects and Challenges

Ulrich Frank
Institut für Wirtschaftsinformatik, Universität Koblenz
Rheinau1, 56075 Koblenz
Germany

"The transformation we are concerned with is not a technical one, but a continuing evolution of how we understand our surrounding and ourselves ... "

Terry Winograd und Fernando Flores

Abstract

The paper will give an overview of how future multimedia information systems could look like and how they could be produced and maintained. Those systems will no longer make a difference between the handling of traditional data and video or audio. Instead they will focus on conveniently providing a requested information content together with the appropriate presentation - no matter whether it is text, graphics, video or audio. These features however do not come for free. Instead a number of challenges has to be faced. The paper will discuss those challenges and present an evolutionary approach with a number of measures and strategies to overcome them.

1. Introduction

We are at the dawning of a new information age. With increasingly powerful digital computers penetrating more and more private homes and emerging high bandwidth Wide Area Networks the grounds are laid for a new generation of computer applications that will integrate a wide range of media. While this development will certainly result in a vast amount of digitized information it will also provide us with more powerful ways to analyze, manipulate, and reuse information. Thereby we will gain the chance to benefit from (or being annoyed by) new ways of communication, entertainment, teaching, and learning. However, in order to fully exploit the potential of this new technology we will also have to develop new ways of preparing, organizing, and maintaining information.

We will first look at the current situation. Some of the existing multimedia applications are already rather impressive. Nevertheless they are certainly way apart from what future systems will look like - and how they will be produced. In order to give an idea of the added value that will result from integrating audio and video with digital information technology we will characterize some attractive services those systems may offer. Analyzing these features reveals that traditional audio and video material on its own is not well suited as input for computer applications. On an abstract level traditional information system design has been dealing with similar problems. Therefore we will apply well established design principles to the problem of handling and preparing video for serving as integrated parts of future information systems.

2. The Current Situation: Adapting new Technology to traditional Perspectives

Today's multimedia applications can be divided into two main streams. The first stream consists of traditional computer systems extended by capabilities to present photos or sequences of video or audio. Typically those applications are specialized information retrieval systems where certain chunks of information are associated with multimedia presentations. Among the most common examples are sales support systems that allow to retrieve objects a customer may be interested in and present them using photos or videos. The second stream of applications aims at supporting video

(post) production. Operating on digitized video images they provide an impressive range of powerful manipulations which can be used in a convenient way. From our point of view it is remarkable that those applications essentially increase productivity and flexibility of video post production. However, they do not attempt to change the professional approach of how to produce videos (in fact, if they did, they would probably be less successful).

Within the first stream there is one particular system that has been a tremendous success so far and that is still attracting an increasing number of sometimes enthusiastic users - specially within the scientific community: World Wide Web (WWW). WWW gives an impression of what can be accomplished with the computer infrastructure usually available at today's research sites - and in tomorrow's homes. It also gives an idea of the cultural changes that go along with the new information age. WWW is an architecture that lies on top of the Internet. It allows to create hypermedia documents. Such a document consists of nodes which may contain formatted text, images, audio, or video. The document nodes are stored within a flexible number of information servers. Any node may contain references to other nodes - no matter where they are physically located. In order to allow the user to conveniently browse through hypermedia documents spread around the world WWW includes specifications for client software. Meanwhile clients exist for all major platforms. The clients also provide means to add information and to organize it in order to support certain ways of interaction. Information retrieval is fostered by various dictionaries and a small set of full text retrieval capabilities. WWW thereby already provides the functionality that is required for video on demand, although the data exchange rates commonly available are not sufficient. Figure 1 illustrates the user interface of a WWW-client.

With thousands of motivated and skillful users at universities and research sites around the world it was no surprise that it did not take long until a vast amount of information was produced within numerous hypermedia documents. It is remarkable however that the features provided by WWW also fostered a new quality of information access and communication. For instance: research results can be quickly disseminated not only as text. Furthermore they can be annotated with pictures or videos - either of the research topic or people involved in the work. Thereby WWW not only affects the way people deal with information - both in providing and accessing it. Like the Internet in general it also allows for new ways to communicate - by providing guided access to members of certain world wide communities and by fostering less formal ways to interact.

What is the lesson we may learn from the current situation? While systems like WWW certainly give a first glance of future multimedia systems the integration of video and information technology is mainly restricted to a basic technical level: digital representation and compression, synchronization etc. At the same time many computer scientists dealing with multimedia systems concentrate on technical problems like building interfaces to analogous devices or widening performance bottlenecks. On the application level videos are treated as black boxes, or - to use a phrase from database technology - as "BLOBs" (Binary large Objects). Typically computer programmers abstract from the content of a video - while it should be the other way around: concentrating on the content and abstracting from technological constraints. On the other hand video professionals only use information technology to increase the efficiency of traditional production processes. They usually intend to produce neat, but stand alone videos instead of thinking about how to produce a video in order to make it well suited for computer applications.

Fig. 1: User interface of a WWW client session

3. Challenges and Strategies

Computer programs usually allow to read and/or write data. However, this can be done on very different levels. Usually it is desirable for an application to provide the user with concepts that fit his perception and conceptualization of the domain that is represented within the application. The more a designer of an application succeeds in accomplishing this goal the better are the chances to build user friendly programs. With familiar concepts being mapped to the application information can be retrieved or manipulated by directly applying the associations a user has in mind when he is thinking about the relevant subject.

Enhancing video with application domain concepts opens a wide range of features - on different levels of complexity. The following examples illustrate some of the services that could be provided:

- retrieve all videos that feature sport events
- retrieve all videos that feature ball games
- retrieve all movie dramas from 1984 starring Robert de Niro
- retrieve all tennis matches where a particular player lost the first set but finally won the match
- retrieve the movie that contains the line “Do you feel lucky?”
- retrieve the movie and the particular scene that contains the line: “Go ahead make my day”.
- retrieve the goal Germany scored in its loss against Bulgaria during the Championship in 1994
- retrieve the murder scene that happened after the protagonist was released from jail
- retrieve all movies that contain sequences of baseball games
- retrieve all TV shows where film previews were presented
- retrieve all movies where a male protagonist kills his lover
- retrieve all soccer games in 1994 where a defender scores a goal after he had received the ball from another defender
- in a particular movie replace the sequence showing a soccer game with a tennis match.
- within the movie “In the Line of Fire” substitute Clint Eastwood with John Wayne.

Apparently some of these services could already be provided by today’s applications. Other services seem to be harder to accomplish. There are two main challenges to accomplish applications that could handle requests like those exemplified above. The first challenge has already been mentioned: applications that operate on videos should have access to concepts describing the content of these videos. In order to be more precise we could also say: It is desirable to provide as much formalized semantics with a video as the user could have in mind for his requests. The semantic content of a representation depends on the formal interpretations it allows for: the more interpretations are excluded the higher the level of semantics. For instance: If you look at a video represented only as a sequence of byte arrays (each of which representing an image) together with a synchronized data stream for audio this video might contain anything - in other words: It does not contain much semantics. In order to overcome this problem you have to explicitly enrich a video with semantics. The strategies discussed below demonstrate how this approach can be pursued in more or less ambitious ways - according to principles well known from information system design.

3.1 Video as a Black Box: Attributes and Annotations

If an application does not know anything about the way information is represented within a video there are two ways to include video into information processing. The first approach treats videos as attributes of objects that are represented in data models. For instance: An actor who is listed in a TV station’s database may be assigned the attribute "example monologue" that is actually a video sequence featuring the actor. While this approach may be appropriate for enhancing given databases with additional presentations it does not directly relate to the content of a video. A well known approach to allow for operations regarding the content of something that itself is not directly represented in a data model is to add textual annotations. To enhance the above example with annotations we could use extra attributes to hold keywords that inform about the content of the video.

If you look at videos with a complex content the annotation could be something like a comprehensive natural language description which in turn could be operated on by a full text retrieval mech-

anism. However, if you provide such a description without further structuring for the whole video it only helps to find a particular video, not a sequence within it. For this reason one could divide a video in a number of sequences of an appropriate duration and assign each sequence a textual description. Most text retrieval or database systems do not include a notion of time. However, there are some prototypical systems that feature retrieval languages which allow for temporal comparisons (like "the sequence *before* sequence x") [like Dean/McDermott or Sripada].

3.2 Classification, Generalization and Specialization

One of the essential principles not only of designing information system but also of systematic real world descriptions in general is classification: You do not intend to solely describe certain instances of the real world. Instead you rather try to define features that are common to a set of instances. In other words: You build classes (or concepts) like the class "sport event". The features all instances have in common may be attributes (like "starting time", "duration", etc.) or services you expect the instances to provide (like "show the last n minutes"). If a number of classes has a common set of features it is a good idea to extract these features into a more general class. For instance: If different video types like "sport event", "drama", etc. all have common attributes like "duration" you could introduce a general class "video". This principle is called "generalization". Its advantage is obvious: It allows to avoid redundant specifications thereby fostering the maintenance of class descriptions. Furthermore it is the prerequisite for applying more sophisticated retrieval operations on a given description: A request can always be related to the most general concept that is appropriate. It will then implicitly be applied (by logical deduction) to the less general concepts. On the other hand you may realize that a given class specification (like "sport event") is not sufficient for a specific case (for instance "tennis match"). Specialization offers a measure to use a given specification and enhance it by additional features. For instance: "tennis match" would inherit all features defined for "sport event" and would additionally have one or more specific features.

Fig. 2 shows an example of how to apply these fundamental principals to video objects.

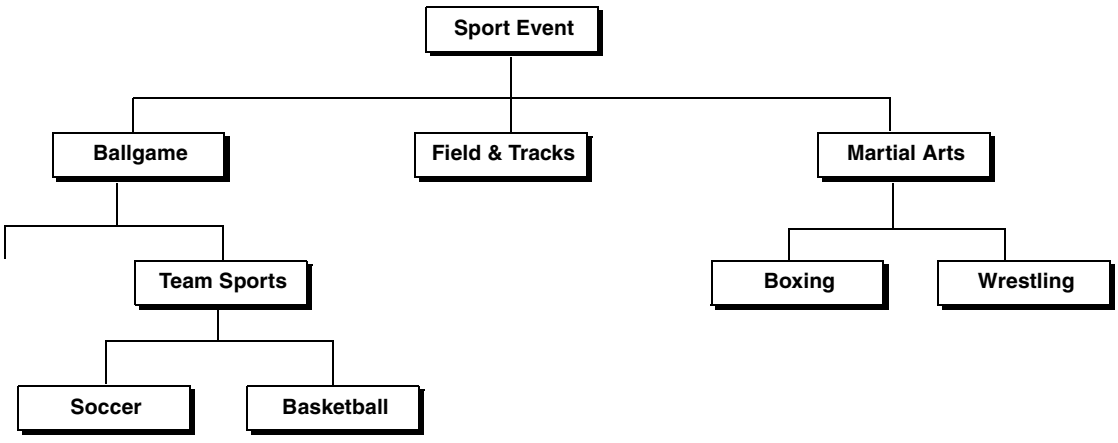

Fig. 2: Classes of sport videos

3.3 Logical Structure

While classification, generalization and specialization are an important step to treat videos as information objects they do not directly provide means to describe the logical structure of a video. Modelling the logical structure enhances the range of meaningful operations - both for read - and

write-access. For instance: You could focus on the second game of the second set of a tennis match. At the same time it is a prerequisite for more sophisticated ways of managing stored videos. For instance: If a logical part of a video uses a part of another video or audio document you would not have to copy this part. Instead there could be a reference to this part of the other source - thereby reducing redundancy and fostering maintenance.

Since it would require a film director to develop a meaningful logical structure for a certain class of films we look at a similar domain: In the area of document management there are a number of logical models, some of them even subject of international standards [see for instance Appelt]. Figure 3 exemplifies how documents of a certain class could be structured.

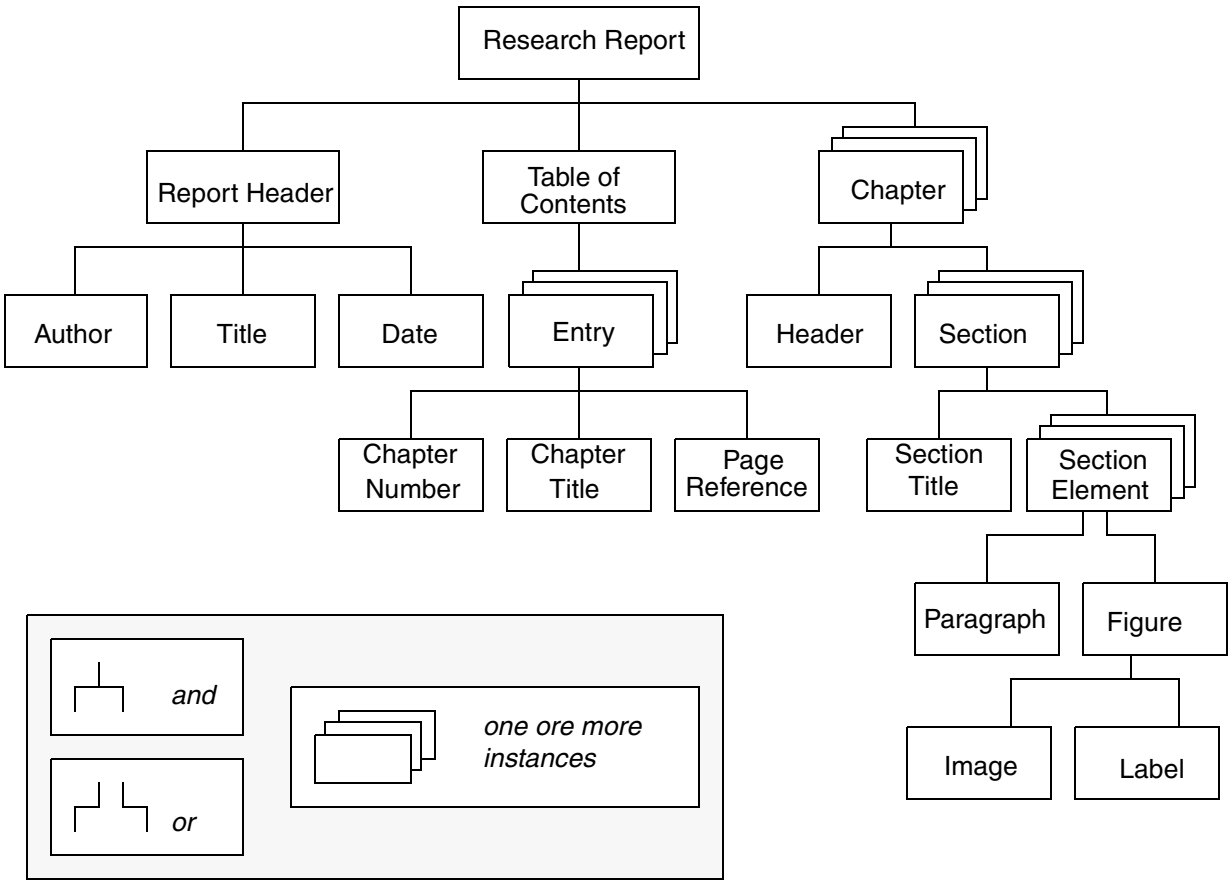

Fig. 3: Logical architecture of compound documents

3.4 Extended Semantic Modelling

The measures proposed so far would not be sufficient to satisfy all of the requests listed above. Those cases require a more detailed description of the relevant concepts. In other words: We need models of the video's subject that incorporate the semantics required to handle the requests. Such models are basically conceptual descriptions. There are a number of ways to structure these descriptions. Traditional data models like the Entity Relationship Model [see Chen] use objects (respectively entities) and relationships between objects. Classes of objects (entity types) are described by a set of attributes. Such approaches are not well suited for the purpose of semantically modelling the content of videos because of their limited expressive capabilities. Research in Artificial Intelligence produces modelling methodologies which are more appropriate. Schank and Abelson for instance designed a formal language to script real life scenes - like visiting a restaurant - in order to build programs that could answer questions regarding these scenes. They suggest two

basic conceptualizations to describe a scene, an active conceptualization ("Actor Action Object Direction") and a stative conceptualization ("Object is in State with value"). Winston introduced an approach to model Shakespearean tragedies in order to recognize situations which are analogies to a given one.

Special Artificial Intelligence approaches would certainly be well suited for modelling the content of videos. However, they are rather exotic. Usually they are not well documented, there are no robust tools to support them, and only few people are familiar with them. Another approach, which was also inspired by Artificial Intelligence research, has gained enormous attention during the last years: object-oriented modelling and implementation is going to be the leading paradigm for future software engineering. Not only that it provides means to model a video's content on a high level of semantic, there are also text books [for instance Booch, Rumbaugh et al.], tools, programming languages and special Database Management Systems available. The last aspect is of outstanding importance because it means tremendous help with implementing an actual application. Furthermore standards for object-oriented technologies are currently emerging. In principle object-oriented modelling suggests to describe an application domain in terms of objects and relationships between objects. Objects are grouped into classes. Generalization and specialization can be applied. Each class is characterized by a set of attributes and services.

Figure 4 gives an example of an object-oriented semantic model with one class described in more detail.

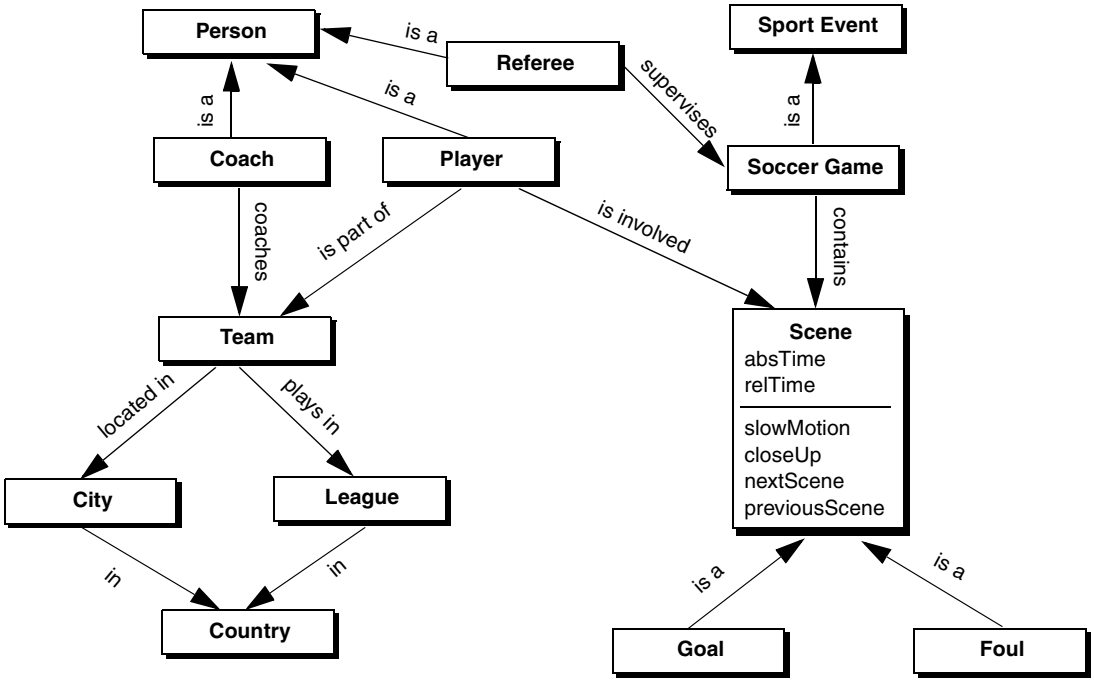

Fig. 4: Example of an object model for sport events with a specification of one class

Object-oriented models allow to describe application domains using concepts that correspond closely to the conceptualizations preferred by the relevant users. At the same time they are based on a solid software technological foundation which fosters implementation and maintenance. One problem however remains to be dealt with in every single modelling project: How to identify the relevant objects? While Meyer, one of the protagonists of object-oriented software development, states "The objects are there for the picking" it is a matter of fact that merely asking the domain

experts for objects or concepts usually does not result in a comprehensive model. Instead it is usually recommended to use a heuristic approach that helps with identifying the relevant objects. There are two promising heuristics. One is to concentrate on processes - which is always preferable when processes are a preferred way to describe a domain. For instance: If you want to design an application for an editor at a TV station you could ask him for the relevant tasks or processes he is involved in. Thereafter you would decompose the tasks or processes into smaller units and then ask for the information objects that are required there. The other heuristics is scripting: By interviewing people or using existing documents you get a script of the relevant domain. This script can then be preprocessed in order to produce a list of nouns (candidates for objects) and predicates (candidates for relationships), which would then be used as the input for designing an object model.

3.5 Instantiation as the Second Major Challenge

Even if the - sometimes tremendous - effort to design comprehensive semantic models can be accomplished there is still one major problem to be solved. A model is an abstraction of a specific case. This is for a good reason: We talk about concepts rather than about instances. However, when it comes to derive a specific application from a model we look at instances: Looking at a video on a particular soccer game we are not only interested in the concepts "player" or "goal" but also in the specific instances. While deriving instances from concepts is a well known procedure - called instantiation - for data processing it marks an outstanding challenge for applications including videos.

In principle there would be the chance to apply sophisticated pattern matching and natural language recognizing algorithms to analyze a video: The patterns would have to be identified as instances of concepts defined in a related semantic model. For instance: A foul in a soccer game. Furthermore it would be desirable to identify the state of this instance. For example: Where on the field did the foul occur? Experience gathered in Artificial Intelligence research on less complex tasks however indicates, that such an approach will only lead to poor (if any) results and is extremely expensive at the same time:

> "I recognized the depth of the difficulties in getting a machine to understand language in any but a superficial and misleading way, and am convinced that people will be much better served by machines that do well-defined and understandable things that those that appear to be like persons until something goes wrong (which won´t take long), at which point there is only confusion."
>
> *Winograd (in Bobrow/Hayes)*

Instead it seems to be much more promising to adapt the production of videos to the needs of multimedia applications. In other words: It is easier and safer to reduce ambiguity by appropriate measures than to apply automatic procedures to resolve it. How could such an approach look like? For all videos which base on events organized for being filmed (like movies or sport events) there is the chance for electronically marking at least some of the relevant instances. Marking means to allow for an affordable and save automatic detection of the instances. That does not have to affect the way a human viewer perceives the video. For instance: Actors in a play could be marked as instances of certain classes (like detective, hero, lover, father, mother, etc.). In order to characterize the relationships between the figures the script that has to be prepared anyway could be enhanced to express important relationships in a way that they would allow for automatic interpretation.[1] A similar approach could be applied to sport events. For instance: The various players in a soccer game could be marked as well as segments of the field.

1. It is no doubt that this would certainly require a major cultural change - and that many people are probably not fond of this idea at all. It is remarkable however that some companies have already adapted the process of writing technical documentation to the requirements of language parsing procedures.

4. Concluding Remarks

For a new generation of multimedia applications to become reality a number of technological challenges have to be overcome. However, the crucial challenges are of another kind: Only if artists, video experts, and computer scientists succeed in cooperating on a conceptual level, the synergy necessary for redesigning current production processes will emerge - without the risk to reinvent the wheel every other day. Cooperation requires communication which in turn stresses the importance of models that foster a common understanding of the essential problems. Therefore the emphasis has to be on conceptual modelling, not on implementation details which are not stable anyway.

Furthermore it is evident that we do not only face an intellectual/cultural challenge: Not everything that could be done, will be affordable. This is the case for semantic modelling as well as for instantiation. Similar to software we already have today, the effort spent for a particular system very much depends on the economy of scale: Since copying is cheap, the price mainly depends on the size of the market. So we will probably encounter both rather simple systems and systems that caused a high amount of production costs but are sold on mass markets.

References

Appelt, W.: Document Architecture in Open Systems. The ODA Standard. Berlin, Heidelberg, New York etc. 1991

Bobrow, D.G.; Hayes, P.J.: Artificial Intelligence - Where Are We? In: Artificial Intelligence 25, 1985, pp. 375-415

Booch, G.: Object-Oriented Analysis and Design with Applications. Redwood City 1994

Chen, P.P.: The Entity-Relationship Model: Towards a Unified View of Data. In: ACM TODS, Vol. 1, No. 1, 1976, pp. 9-36

Dean, T.; McDermott, D.V.: Temporal Data Base Management. In: Artificial Intelligence, Vol. 32, 1987, pp. 1-55

Loomis, M.E.S.: OODBMS - The Basics. In: Journal of Object-Oriented Programming. Vol. 3, No. 1, 1990, pp. 77-88

Meyer, B.: Object-Oriented Software Construction. Englewood Cliffs, N.J. 1988

Rumbaugh, J. et al.: Object-oriented modeling and design. Englewood Cliffs, N.J. 1991

Schank, R.; Abelson, R.: Scripts, Plans, Goals and Understanding. Hillsdale 1975

Sripada, S.M.: A logical framework for temporal deductive databases. In: Banchilhon, F.; De Witt, D.J. (Hg.): Proceedings of the 14th International Conference on Very Large Database Systems. Los Altos 1988, pp. 171-182

Winograd, T.; Flores, F.: Understanding Computers and Cognition - a new Foundation for Design. Norwood, N.J. 1986

Winston, P.H.: Learning and Reasoning by Analogy. In: Communications of the ACM. Vol. 23, 1980, pp. 689-703

Published in: Proceedings of the 1994 European Conference of the Society of Motion Picture and Television Engineers. Köln 1994, S. 98-106

Integrating Video with Information Technology - Prospects and Challenges

Ulrich Frank
Institut für Wirtschaftsinformatik, Universität Koblenz
Rheinau1, 56075 Koblenz
Germany

"The transformation we are concerned with is not a technical one, but a continuing evolution of how we understand our surrounding and ourselves ... "

Terry Winograd und Fernando Flores

Abstract

The paper will give an overview of how future multimedia information systems could look like and how they could be produced and maintained. Those systems will no longer make a difference between the handling of traditional data and video or audio. Instead they will focus on conveniently providing a requested information content together with the appropriate presentation - no matter whether it is text, graphics, video or audio. These features however do not come for free. Instead a number of challenges has to be faced. The paper will discuss those challenges and present an evolutionary approach with a number of measures and strategies to overcome them.

1. Introduction

We are at the dawning of a new information age. With increasingly powerful digital computers penetrating more and more private homes and emerging high bandwidth Wide Area Networks the grounds are laid for a new generation of computer applications that will integrate a wide range of media. While this development will certainly result in a vast amount of digitized information it will also provide us with more powerful ways to analyze, manipulate, and reuse information. Thereby we will gain the chance to benefit from (or being annoyed by) new ways of communication, entertainment, teaching, and learning. However, in order to fully exploit the potential of this new technology we will also have to develop new ways of preparing, organizing, and maintaining information.

We will first look at the current situation. Some of the existing multimedia applications are already rather impressive. Nevertheless they are certainly way apart from what future systems will look like - and how they will be produced. In order to give an idea of the added value that will result from integrating audio and video with digital information technology we will characterize some attractive services those systems may offer. Analyzing these features reveals that traditional audio and video material on its own is not well suited as input for computer applications. On an abstract level traditional information system design has been dealing with similar problems. Therefore we will apply well established design principles to the problem of handling and preparing video for serving as integrated parts of future information systems.

2. The Current Situation: Adapting new Technology to traditional Perspectives

Today's multimedia applications can be divided into two main streams. The first stream consists of traditional computer systems extended by capabilities to present photos or sequences of video or audio. Typically those applications are specialized information retrieval systems where certain chunks of information are associated with multimedia presentations. Among the most common examples are sales support systems that allow to retrieve objects a customer may be interested in and present them using photos or videos. The second stream of applications aims at supporting video

(post) production. Operating on digitized video images they provide an impressive range of powerful manipulations which can be used in a convenient way. From our point of view it is remarkable that those applications essentially increase productivity and flexibility of video post production. However, they do not attempt to change the professional approach of how to produce videos (in fact, if they did, they would probably be less successful).

Within the first stream there is one particular system that has been a tremendous success so far and that is still attracting an increasing number of sometimes enthusiastic users - specially within the scientific community: World Wide Web (WWW). WWW gives an impression of what can be accomplished with the computer infrastructure usually available at today's research sites - and in tomorrow's homes. It also gives an idea of the cultural changes that go along with the new information age. WWW is an architecture that lies on top of the Internet. It allows to create hypermedia documents. Such a document consists of nodes which may contain formatted text, images, audio, or video. The document nodes are stored within a flexible number of information servers. Any node may contain references to other nodes - no matter where they are physically located. In order to allow the user to conveniently browse through hypermedia documents spread around the world WWW includes specifications for client software. Meanwhile clients exist for all major platforms. The clients also provide means to add information and to organize it in order to support certain ways of interaction. Information retrieval is fostered by various dictionaries and a small set of full text retrieval capabilities. WWW thereby already provides the functionality that is required for video on demand, although the data exchange rates commonly available are not sufficient. Figure 1 illustrates the user interface of a WWW-client.

With thousands of motivated and skillful users at universities and research sites around the world it was no surprise that it did not take long until a vast amount of information was produced within numerous hypermedia documents. It is remarkable however that the features provided by WWW also fostered a new quality of information access and communication. For instance: research results can be quickly disseminated not only as text. Furthermore they can be annotated with pictures or videos - either of the research topic or people involved in the work. Thereby WWW not only affects the way people deal with information - both in providing and accessing it. Like the Internet in general it also allows for new ways to communicate - by providing guided access to members of certain world wide communities and by fostering less formal ways to interact.

What is the lesson we may learn from the current situation? While systems like WWW certainly give a first glance of future multimedia systems the integration of video and information technology is mainly restricted to a basic technical level: digital representation and compression, synchronization etc. At the same time many computer scientists dealing with multimedia systems concentrate on technical problems like building interfaces to analogous devices or widening performance bottlenecks. On the application level videos are treated as black boxes, or - to use a phrase from database technology - as "BLOBs" (Binary large Objects). Typically computer programmers abstract from the content of a video - while it should be the other way around: concentrating on the content and abstracting from technological constraints. On the other hand video professionals only use information technology to increase the efficiency of traditional production processes. They usually intend to produce neat, but stand alone videos instead of thinking about how to produce a video in order to make it well suited for computer applications.

Fig. 1: User interface of a WWW client session

3. Challenges and Strategies

Computer programs usually allow to read and/or write data. However, this can be done on very different levels. Usually it is desirable for an application to provide the user with concepts that fit his perception and conceptualization of the domain that is represented within the application. The more a designer of an application succeeds in accomplishing this goal the better are the chances to build user friendly programs. With familiar concepts being mapped to the application information can be retrieved or manipulated by directly applying the associations a user has in mind when he is thinking about the relevant subject.

Enhancing video with application domain concepts opens a wide range of features - on different levels of complexity. The following examples illustrate some of the services that could be provided:

- retrieve all videos that feature sport events
- retrieve all videos that feature ball games
- retrieve all movie dramas from 1984 starring Robert de Niro
- retrieve all tennis matches where a particular player lost the first set but finally won the match
- retrieve the movie that contains the line “Do you feel lucky?”
- retrieve the movie and the particular scene that contains the line: “Go ahead make my day”.
- retrieve the goal Germany scored in its loss against Bulgaria during the Championship in 1994
- retrieve the murder scene that happened after the protagonist was released from jail
- retrieve all movies that contain sequences of baseball games
- retrieve all TV shows where film previews were presented
- retrieve all movies where a male protagonist kills his lover
- retrieve all soccer games in 1994 where a defender scores a goal after he had received the ball from another defender
- in a particular movie replace the sequence showing a soccer game with a tennis match.
- within the movie “In the Line of Fire” substitute Clint Eastwood with John Wayne.

Apparently some of these services could already be provided by today’s applications. Other services seem to be harder to accomplish. There are two main challenges to accomplish applications that could handle requests like those exemplified above. The first challenge has already been mentioned: applications that operate on videos should have access to concepts describing the content of these videos. In order to be more precise we could also say: It is desirable to provide as much formalized semantics with a video as the user could have in mind for his requests. The semantic content of a representation depends on the formal interpretations it allows for: the more interpretations are excluded the higher the level of semantics. For instance: If you look at a video represented only as a sequence of byte arrays (each of which representing an image) together with a synchronized data stream for audio this video might contain anything - in other words: It does not contain much semantics. In order to overcome this problem you have to explicitly enrich a video with semantics. The strategies discussed below demonstrate how this approach can be pursued in more or less ambitious ways - according to principles well known from information system design.

3.1 Video as a Black Box: Attributes and Annotations

If an application does not know anything about the way information is represented within a video there are two ways to include video into information processing. The first approach treats videos as attributes of objects that are represented in data models. For instance: An actor who is listed in a TV station’s database may be assigned the attribute "example monologue" that is actually a video sequence featuring the actor. While this approach may be appropriate for enhancing given databases with additional presentations it does not directly relate to the content of a video. A well known approach to allow for operations regarding the content of something that itself is not directly represented in a data model is to add textual annotations. To enhance the above example with annotations we could use extra attributes to hold keywords that inform about the content of the video.

If you look at videos with a complex content the annotation could be something like a comprehensive natural language description which in turn could be operated on by a full text retrieval mech-

anism. However, if you provide such a description without further structuring for the whole video it only helps to find a particular video, not a sequence within it. For this reason one could divide a video in a number of sequences of an appropriate duration and assign each sequence a textual description. Most text retrieval or database systems do not include a notion of time. However, there are some prototypical systems that feature retrieval languages which allow for temporal comparisons (like "the sequence *before* sequence x") [like Dean/McDermott or Sripada].

3.2 Classification, Generalization and Specialization

One of the essential principles not only of designing information system but also of systematic real world descriptions in general is classification: You do not intend to solely describe certain instances of the real world. Instead you rather try to define features that are common to a set of instances. In other words: You build classes (or concepts) like the class "sport event". The features all instances have in common may be attributes (like "starting time", "duration", etc.) or services you expect the instances to provide (like "show the last n minutes"). If a number of classes has a common set of features it is a good idea to extract these features into a more general class. For instance: If different video types like "sport event", "drama", etc. all have common attributes like "duration" you could introduce a general class "video". This principle is called "generalization". Its advantage is obvious: It allows to avoid redundant specifications thereby fostering the maintenance of class descriptions. Furthermore it is the prerequisite for applying more sophisticated retrieval operations on a given description: A request can always be related to the most general concept that is appropriate. It will then implicitly be applied (by logical deduction) to the less general concepts. On the other hand you may realize that a given class specification (like "sport event") is not sufficient for a specific case (for instance "tennis match"). Specialization offers a measure to use a given specification and enhance it by additional features. For instance: "tennis match" would inherit all features defined for "sport event" and would additionally have one or more specific features.

Fig. 2 shows an example of how to apply these fundamental principals to video objects.

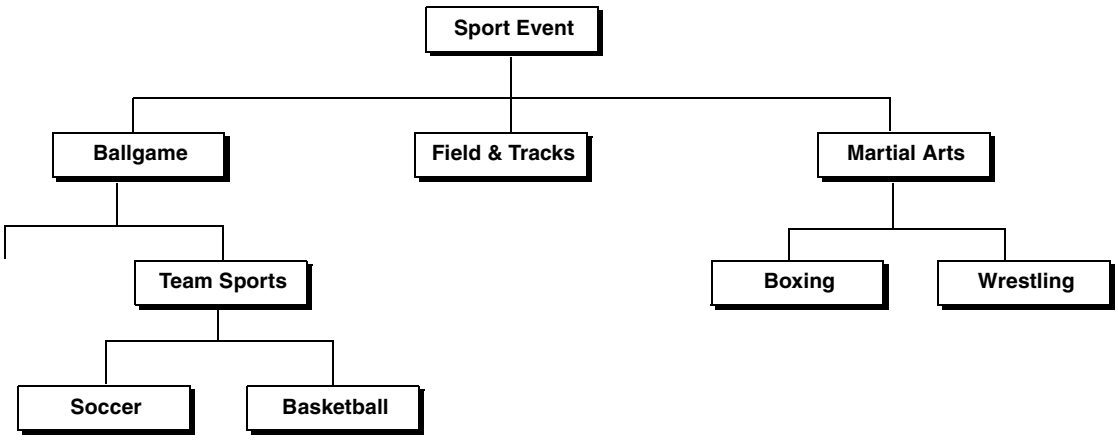

Fig. 2: Classes of sport videos

3.3 Logical Structure

While classification, generalization and specialization are an important step to treat videos as information objects they do not directly provide means to describe the logical structure of a video. Modelling the logical structure enhances the range of meaningful operations - both for read - and

write-access. For instance: You could focus on the second game of the second set of a tennis match. At the same time it is a prerequisite for more sophisticated ways of managing stored videos. For instance: If a logical part of a video uses a part of another video or audio document you would not have to copy this part. Instead there could be a reference to this part of the other source - thereby reducing redundancy and fostering maintenance.

Since it would require a film director to develop a meaningful logical structure for a certain class of films we look at a similar domain: In the area of document management there are a number of logical models, some of them even subject of international standards [see for instance Appelt]. Figure 3 exemplifies how documents of a certain class could be structured.

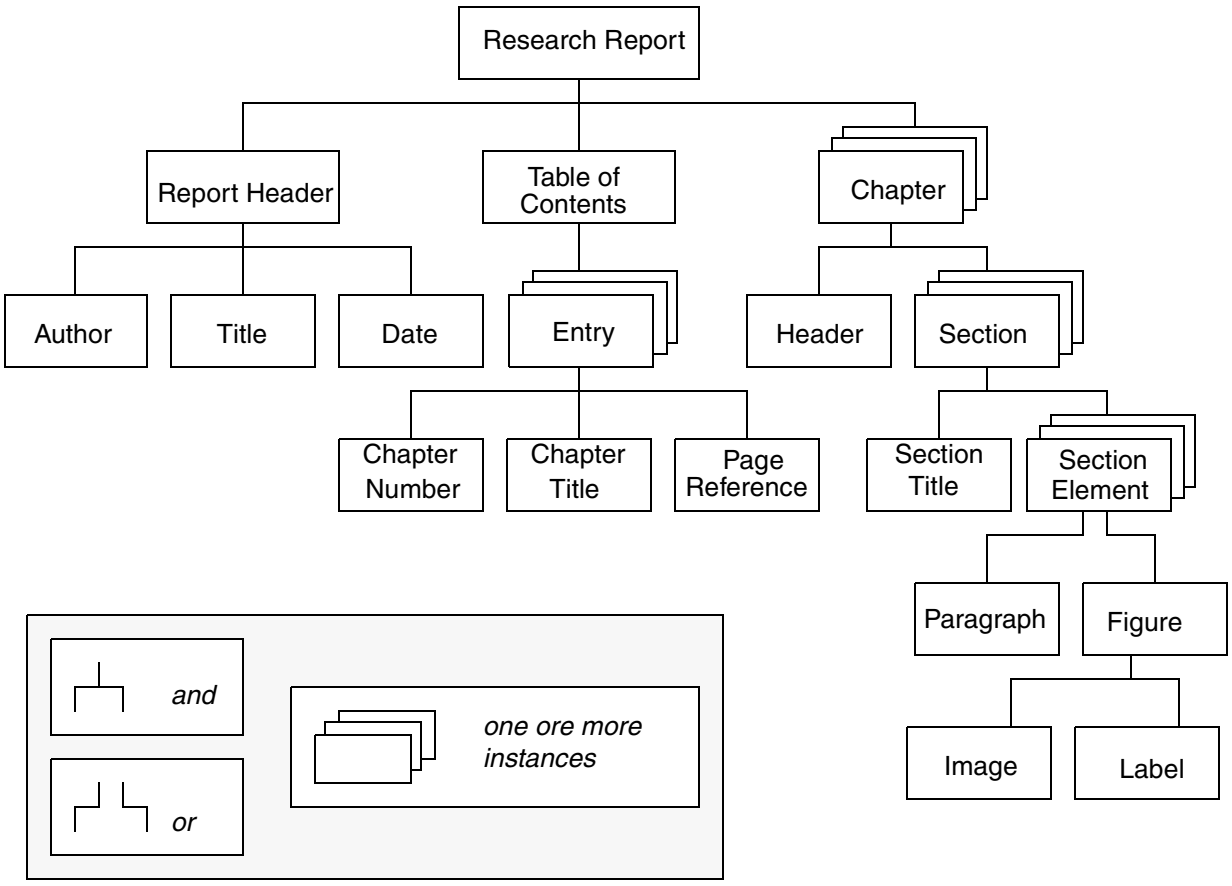

Fig. 3: Logical architecture of compound documents

3.4 Extended Semantic Modelling

The measures proposed so far would not be sufficient to satisfy all of the requests listed above. Those cases require a more detailed description of the relevant concepts. In other words: We need models of the video's subject that incorporate the semantics required to handle the requests. Such models are basically conceptual descriptions. There are a number of ways to structure these descriptions. Traditional data models like the Entity Relationship Model [see Chen] use objects (respectively entities) and relationships between objects. Classes of objects (entity types) are described by a set of attributes. Such approaches are not well suited for the purpose of semantically modelling the content of videos because of their limited expressive capabilities. Research in Artificial Intelligence produces modelling methodologies which are more appropriate. Schank and Abelson for instance designed a formal language to script real life scenes - like visiting a restaurant - in order to build programs that could answer questions regarding these scenes. They suggest two

basic conceptualizations to describe a scene, an active conceptualization ("Actor Action Object Direction") and a stative conceptualization ("Object is in State with value"). Winston introduced an approach to model Shakespearean tragedies in order to recognize situations which are analogies to a given one.

Special Artificial Intelligence approaches would certainly be well suited for modelling the content of videos. However, they are rather exotic. Usually they are not well documented, there are no robust tools to support them, and only few people are familiar with them. Another approach, which was also inspired by Artificial Intelligence research, has gained enormous attention during the last years: object-oriented modelling and implementation is going to be the leading paradigm for future software engineering. Not only that it provides means to model a video's content on a high level of semantic, there are also text books [for instance Booch, Rumbaugh et al.], tools, programming languages and special Database Management Systems available. The last aspect is of outstanding importance because it means tremendous help with implementing an actual application. Furthermore standards for object-oriented technologies are currently emerging. In principle object-oriented modelling suggests to describe an application domain in terms of objects and relationships between objects. Objects are grouped into classes. Generalization and specialization can be applied. Each class is characterized by a set of attributes and services.

Figure 4 gives an example of an object-oriented semantic model with one class described in more detail.

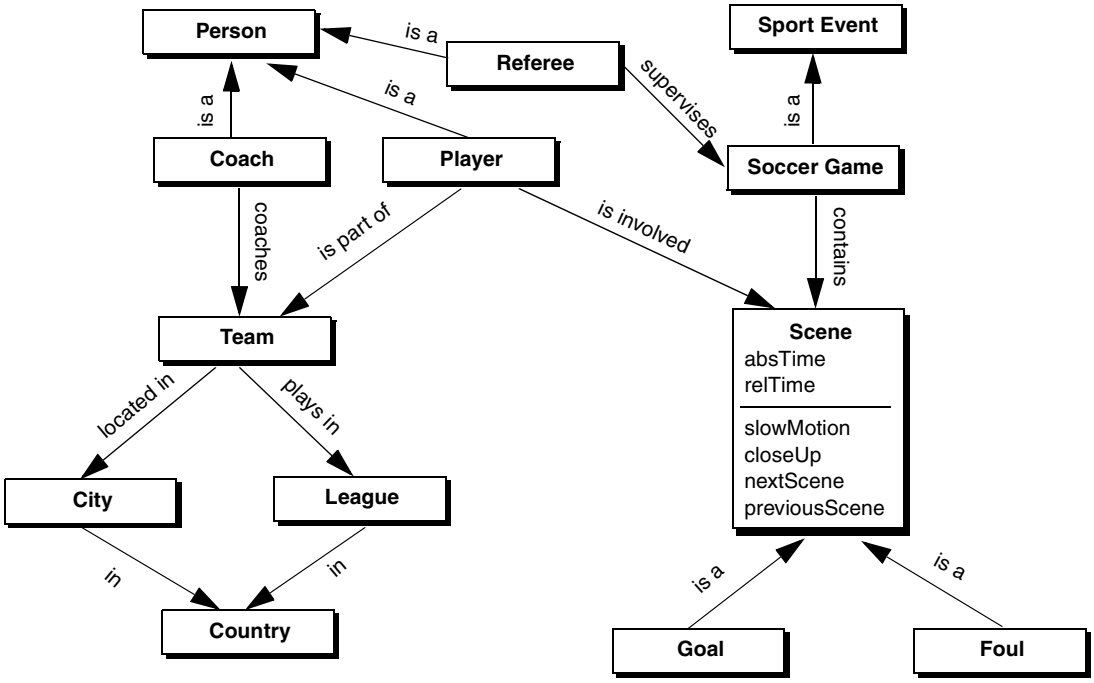

Fig. 4: Example of an object model for sport events with a specification of one class

Object-oriented models allow to describe application domains using concepts that correspond closely to the conceptualizations preferred by the relevant users. At the same time they are based on a solid software technological foundation which fosters implementation and maintenance. One problem however remains to be dealt with in every single modelling project: How to identify the relevant objects? While Meyer, one of the protagonists of object-oriented software development, states "The objects are there for the picking" it is a matter of fact that merely asking the domain

experts for objects or concepts usually does not result in a comprehensive model. Instead it is usually recommended to use a heuristic approach that helps with identifying the relevant objects. There are two promising heuristics. One is to concentrate on processes - which is always preferable when processes are a preferred way to describe a domain. For instance: If you want to design an application for an editor at a TV station you could ask him for the relevant tasks or processes he is involved in. Thereafter you would decompose the tasks or processes into smaller units and then ask for the information objects that are required there. The other heuristics is scripting: By interviewing people or using existing documents you get a script of the relevant domain. This script can then be preprocessed in order to produce a list of nouns (candidates for objects) and predicates (candidates for relationships), which would then be used as the input for designing an object model.

3.5 Instantiation as the Second Major Challenge

Even if the - sometimes tremendous - effort to design comprehensive semantic models can be accomplished there is still one major problem to be solved. A model is an abstraction of a specific case. This is for a good reason: We talk about concepts rather than about instances. However, when it comes to derive a specific application from a model we look at instances: Looking at a video on a particular soccer game we are not only interested in the concepts "player" or "goal" but also in the specific instances. While deriving instances from concepts is a well known procedure - called instantiation - for data processing it marks an outstanding challenge for applications including videos.

In principle there would be the chance to apply sophisticated pattern matching and natural language recognizing algorithms to analyze a video: The patterns would have to be identified as instances of concepts defined in a related semantic model. For instance: A foul in a soccer game. Furthermore it would be desirable to identify the state of this instance. For example: Where on the field did the foul occur? Experience gathered in Artificial Intelligence research on less complex tasks however indicates, that such an approach will only lead to poor (if any) results and is extremely expensive at the same time:

> "I recognized the depth of the difficulties in getting a machine to understand language in any but a superficial and misleading way, and am convinced that people will be much better served by machines that do well-defined and understandable things that those that appear to be like persons until something goes wrong (which won´t take long), at which point there is only confusion."
>
> *Winograd (in Bobrow/Hayes)*

Instead it seems to be much more promising to adapt the production of videos to the needs of multimedia applications. In other words: It is easier and safer to reduce ambiguity by appropriate measures than to apply automatic procedures to resolve it. How could such an approach look like? For all videos which base on events organized for being filmed (like movies or sport events) there is the chance for electronically marking at least some of the relevant instances. Marking means to allow for an affordable and save automatic detection of the instances. That does not have to affect the way a human viewer perceives the video. For instance: Actors in a play could be marked as instances of certain classes (like detective, hero, lover, father, mother, etc.). In order to characterize the relationships between the figures the script that has to be prepared anyway could be enhanced to express important relationships in a way that they would allow for automatic interpretation.[1] A similar approach could be applied to sport events. For instance: The various players in a soccer game could be marked as well as segments of the field.

1. It is no doubt that this would certainly require a major cultural change - and that many people are probably not fond of this idea at all. It is remarkable however that some companies have already adapted the process of writing technical documentation to the requirements of language parsing procedures.

4. Concluding Remarks

For a new generation of multimedia applications to become reality a number of technological challenges have to be overcome. However, the crucial challenges are of another kind: Only if artists, video experts, and computer scientists succeed in cooperating on a conceptual level, the synergy necessary for redesigning current production processes will emerge - without the risk to reinvent the wheel every other day. Cooperation requires communication which in turn stresses the importance of models that foster a common understanding of the essential problems. Therefore the emphasis has to be on conceptual modelling, not on implementation details which are not stable anyway.

Furthermore it is evident that we do not only face an intellectual/cultural challenge: Not everything that could be done, will be affordable. This is the case for semantic modelling as well as for instantiation. Similar to software we already have today, the effort spent for a particular system very much depends on the economy of scale: Since copying is cheap, the price mainly depends on the size of the market. So we will probably encounter both rather simple systems and systems that caused a high amount of production costs but are sold on mass markets.

References

Appelt, W.: Document Architecture in Open Systems. The ODA Standard. Berlin, Heidelberg, New York etc. 1991

Bobrow, D.G.; Hayes, P.J.: Artificial Intelligence - Where Are We? In: Artificial Intelligence 25, 1985, pp. 375-415

Booch, G.: Object-Oriented Analysis and Design with Applications. Redwood City 1994

Chen, P.P.: The Entity-Relationship Model: Towards a Unified View of Data. In: ACM TODS, Vol. 1, No. 1, 1976, pp. 9-36

Dean, T.; McDermott, D.V.: Temporal Data Base Management. In: Artificial Intelligence, Vol. 32, 1987, pp. 1-55

Loomis, M.E.S.: OODBMS - The Basics. In: Journal of Object-Oriented Programming. Vol. 3, No. 1, 1990, pp. 77-88

Meyer, B.: Object-Oriented Software Construction. Englewood Cliffs, N.J. 1988

Rumbaugh, J. et al.: Object-oriented modeling and design. Englewood Cliffs, N.J. 1991

Schank, R.; Abelson, R.: Scripts, Plans, Goals and Understanding. Hillsdale 1975

Sripada, S.M.: A logical framework for temporal deductive databases. In: Banchilhon, F.; De Witt, D.J. (Hg.): Proceedings of the 14th International Conference on Very Large Database Systems. Los Altos 1988, pp. 171-182

Winograd, T.; Flores, F.: Understanding Computers and Cognition - a new Foundation for Design. Norwood, N.J. 1986

Winston, P.H.: Learning and Reasoning by Analogy. In: Communications of the ACM. Vol. 23, 1980, pp. 689-703