Conceptual Modeling of Audio-Visual Content

John R. Smith and Ana B. Benitez

IBM T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532
{jrsmith,ana}@watson.ibm.com

ABSTRACT

The explosive growth of applications involving digital audio-visual material is driving the need for more rich content description. In this paper, we present a design process for audio-visual content description. The process involves constructing a conceptual model in which entities, attributes and relationships important for content description are identified and modeled. The procedure provides special handling of different types of spatio-temporal relationships important for audio-visual content modeling. In this paper, we describe the audio-visual content description conceptual modeling methodology in the context of the evolution of MPEG-7 description schemes.

KEYWORDS: Audio-visual data modeling, conceptual modeling, digital libraries, image databases, MPEG-7.

1 INTRODUCTION

Two well known obstacles for audio-visual information search and retrieval result from the insufficiency of keyword indexing and the limitation of automated content-based analysis methods. Although much progress has been made in analyzing image and video data, such as to enable content-based retrieval, the unconstrained recognition of objects and understanding of scenes remain elusive goals. Regardless, one important challenge remains – providing a sufficiently rich framework for *describing* audio-visual content, regardless of whether content extraction methods are automatic, semi-automatic or manual.

Recently, the Moving Picture's Experts Group (MPEG) has begun work on a new standard related to multimedia content description, known as MPEG-7. The goal of MPEG-7 is to enable fast and efficient searching and filtering of audio-visual material. The effort is being driven by specific requirements taken from a large number of applications related to image, video and audio databases, media filtering and interactive media services (radio, TV programs), scientific image libraries, and so forth. Among the tasks of MPEG-7 is the definition of a number of specific description schemes for audio-visual material. The description schemes, which are related to each other through linking mechanisms, form a data model or schema for audio-visual content description.

Typically, in audio-visual applications the audio-visual description data needs to be stored, queried and retrieved. This encourages us to consider also the relevance of traditional database and software design methodologies in the modeling process. By distinguishing the separate tasks of conceptual, logical and physical modeling as in database design, we breakdown the content description process into separate steps with different dependencies on the eventual implementation.

1.1 Related Work

There have been a number of recent attempts at modeling image, video and multimedia content. Leung, et. al, developed a picture description language (PDL) based on an entity-attribute-relationship model [1]. PDL allows structured annotating of images that emphasizes the relationships among depicted objects. Gutpa et al, developed a four-layer model for picture description, which includes an image representation layer, an image object layer, an semantic object layer and a semantic event layer [2]. Gandhi and Robertson developed an algebraic formalism for describing and organizing continuous media data such as video [3].

Other efforts have focussed on audio-visual content modeling formalisms. Lahlou developed the Extended Model for Information Retrieval (EMIR) which models objects, relationships, and concept categories that are comprised of descriptions, compositions and topologies [4]. EMIR differs from classical Object-Oriented (O-O) models in that categories abstract only a minimal part of the object structures, where the objects are allowed more individual features. The Modeling Object-Oriented Data Semantics (MOODS) system was developed by Grifioen, et al, to model image objects, domain objects and the mappings between them [5]. MOODS extends the O-O model to include dynamic data semantics, abstract function types, and a built-in history mechanism. Woelk et al, examined an O-O approach to multimedia databases, which emphasized specialized roles of aggregation and relationships for multimedia applications [6].

For video databases, the OVID video-object system extends the O-O model to be schema-less, that is, to allow arbitrary attribute structures for the video objects, and to allow attribute-value inheritance based on temporal interval inclusion relationships [7]. In OVID, a video object is arbitrary video frame sequence meant to correspond to a meaningful scene. Each video object is

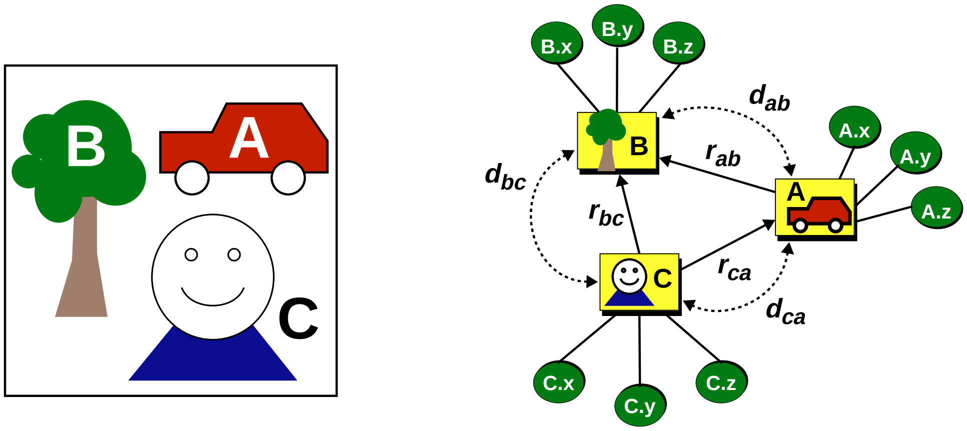


Figure 1: Audio-visual content description example: (a) Image showing three objects (A, B, C), (b) modeling of the corresponding entities, and their attributes and relationships – both extensional (r_{ab}, r_{bc}, r_{ca}) and intensional (d_{ab}, d_{bc}, d_{ca}).

assigned its own attributes and attribute values to describe its contents.

In contrast to modeling, several systems have been developed for annotating audio-visual content directly using iconographic languages. ICONCLASS provides a alphanumeric-based iconographic classification system, which includes ready-made definitions of objects, persons, events, situations and abstract ideas for describing art pictures [8]. Davis developed an iconographic video annotation language, which is used to annotate arbitrarily overlapping temporal segments of video [9].

2 CONCEPTUAL MODELING

The goal of the audio-visual conceptual modeling is to build a high-level model of the audio-visual domain based on content description and requirements. The conceptual modeling process consists of identifying entities and relationships in the audio-visual domain but not their behavior or methods of manipulation (functional analysis).

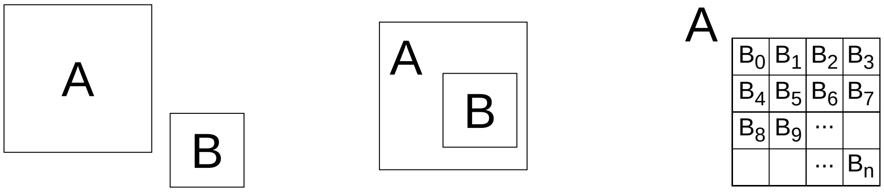
There have been a number of methodologies developed for conceptual modeling. In practice, conceptual modeling has found roles in both database design and software design using many powerful notions from Entity-relationship (E-R) and object-oriented (O-O) modeling. E-R modeling was developed to provide an implementation independent view of the data and has been instrumental in providing a unified representation for different types of relational, hierarchical and semantic databases [10]. With the growth of O-O methodologies, the E-R model has been extended to include important O-O features such as generalization and aggregation [11].

2.1 Audio-visual content description

We apply notions from both E-R and O-O methodologies in modeling audio-visual content. The conceptual analysis is based on the decomposition of the audio-visual domain into key abstractions that exhibit well-defined behavior [6]. As identified in the work on MPEG-7, audio-visual content may be described at many levels, such as structure, semantics, features and meta-data. Many key-abstractions or principal concepts originate from these different levels, such as regions, segments, spatio-temporal organization, objects, events, color, texture, shape, motion, title, author, and so forth.

To illustrate some of the complexity in describing audio-visual content, Figure 1 shows an example description of image content. The example shows the relevance of some of the MPEG-7 principal concepts (objects, spatio-temporal relationships, features). The image depicts three objects (A, B, C). Each object has distinct properties or features (i.e., $A.x$, $A.y$, $A.z$), which correspond to location, size, shape, color, texture and so forth. In addition, each object has relationships (spatial and semantic) to the other objects. Some of the relationships may be stored explicitly – for example, information may state that object C ("person") is associated with ("is the driver of") object A ("the car"). Other information is derived – for example, the relative spatial location of object A to object C may indicate that A "is near" C.

We consider that an "object" in Figure 1 is an entity in the conceptual model. The properties of the "object" correspond to attributes of the entity in the model, while its relationships to other "objects" correspond to relationships in the model. In general, the relationships can be extensional (stored), r_{ab}, r_{bc}, r_{ca}, or intensional (derived), d_{ab}, d_{bc}, d_{ca}. Due to the potential combinatorial explosion in the number of relationships among entities (i.e., "objects" in an image), not all of the relationships are typically stored, but are inferred at query time. In addition, many of the relationships in the audio-visual domain are of special spatio-temporal types as shown in Figure 2. For example, a "region" within the image has spatial association relationships (i.e., "near", "left of") with other "regions" as shown in Figure 2(a). How-



(a) (b) (c)

Figure 2: Different types of spatio-temporal relationships are distinguished in the conceptual model: (a) STA - spatio-temporal association, (b) STE - spatio-temporal aggregation, (c) STC - spatio-temporal composition.

ever, a "region" can also have spatial aggregation, i.e., "encapsulation," as shown in Figure 2(b), and spatial composition relationships, as shown in Figure 2(c), with other regions. It is important to distinguish between these special relationship types in the audio-visual conceptual model.

2.2 Audio-visual content modeling

In order to construct the audio-visual conceptual model, we first carry out the task of identifying the principal concepts as described above. The next task is to classify each concept as follows:

1. **Entity** – a "thing" in the audio-visual domain that can be distinctively identified, i.e., based on the principle objects in the domain. An object class can be used to describe the set of things that have the same characteristics. Examples include *shot*, *frame*, *video*, *segment*, *image*, *region*, *event*.

2. **Attribute** – information obtained about an entity or relationship that has descriptive properties. There are two types of attributes: identifiers – uniquely identify entities, and descriptors – describe entities. Examples include *color*, *texture*, *shape*, *motion*.

3. **Relationship** – an association among one or more entities, described as follows:

(a) **Generalization** – a relationship that partitions the entity class into mutually exclusive subclasses. Generalizations are also called "a kind of" or "is a" or "category" or "specialization" relationships. An example includes: a *frame* is an *image*.

(b) **Aggregation** – is an assembly-component relationship, also called a "part of" relationship. An example includes: a *face object* is part-of a *person object*.

(c) **Association** – relates two or more independent entities that do not exhibit existence dependency. An example includes: a *region* depicts an *object*.

(d) **Spatio-temporal relationship** – corresponds to a relationship among entities in space and/or time, as follows:

   i. **Spatio-temporal association (STA)** – an association of independent entities along the spatio-temporal dimension. An example includes: a *region* is related to other *regions* via spatial associations such as "near".

   ii. **Spatio-temporal aggregation (STE)** – a relationship along the spatio-temporal dimension in which one entity encapsulates or aggregates other entities. An example includes: a *shot* is related to a *key-frame* via temporal aggregation in that it temporal encapsulates the *key-frame*.

   iii. **Spatio-temporal composition (STC)** – a relationship along the spatio-temporal dimension in which one entity is strictly composed of other entities. An example includes: a *video* is related to *shots* via temporal composition as a sequence of *shots*.

(e) **Qualification of relationships** – adds information about the many end of the relationship, as follows:

   i. **Cardinality** – indicates the number of entities of one class that are related to entities of the other class, which is characterized by degree of connectivity (one to one, one to many, many to many). An example includes: a *video* contains one or more *shots*.

   ii. **Existence** – indicates the degree of participation in relationships, such as by allowing optional participation. An example includes: a *shot* may optionally contain a *key-frame*.

2.3 Example

We provide an example construction of an audio-visual conceptual model using nine of the principal concepts from MPEG-7 in Figure 3. The conceptual model shows the complex relationships among entities and the important of modeling the spatio-temporal relationships. The model defines entities and relationships as follows:

- A *video* is temporally composed of a sequence of *shots*
- A *shot* is temporally composed of a sequence of *frames*
- A *frame* is an *image*,
- A *key-frame* is a *frame*,
- A *shot* is a *segment*,
- A *video* temporally aggregates *segments*
- A *segment* is temporally associated with other *segments*
- A *segment* is temporally composed of and aggregates other *segments*

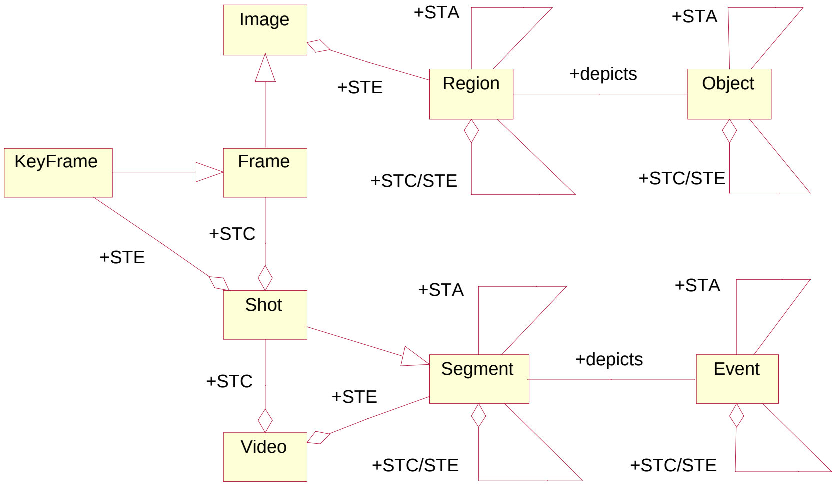


Figure 3: Example conceptual model showing audio-visual entities and their spatio-temporal and semantic relationships.

- A *segment* depicts an *event*
- An *event* is temporally associated with other *events*
- An *event* is temporally composed of and aggregates other *events*
- A *shot* temporally aggregates *key-frames*
- An *image* spatially aggregates *regions*
- A *region* is spatially associated with other *regions*
- A *region* is spatially composed of and aggregates other *regions*
- A *region* depicts an *object*
- An *object* is spatially associated with other *objects*
- An *object* is spatially composed of and aggregates other *objects*

Once the conceptual model is constructed, automatic mapping procedures may be used to generate appropriate structures for audio-visual content description (description schemes), storage (database schema) [11] and implementation (software classes).

3 SUMMARY

In this paper, we presented a methodology for audio-visual content description that involves conceptual modeling. We described the audio-visual modeling constructs including several special methods for handling different types of spatio-temporal relationships.

REFERENCES

1. C. H. C. Leung, D. Hibler, and N. Mwara. Picture retrieval by content description. *Journal of Information Science*, 18:111 – 119, 1992.

2. A. Gupta, T. E. Weymouth, and R. Jain. Semantic queries with pictures: The VIMSYS model. In *Proc. Conf. on Very Large Databases (VLDB)*, pages 69 – 70, September 3 – 6 1991.

3. M. Gandhi and E. L. Robertson. A data model for audio-video data. Technical Report 415, Computer Science Department, Indiana University, Bloomington, IN, August 1994.

4. Y. Lahlou. Modeling complex objects in content-based image retrieval. In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology – Storage & Retrieval for Image and Video Databases IV*, pages 104 – 115, San Jose, CA, February 1995.

5. J. Griffioen, R. Mehotra, and R. Yavatkar. An object-oriented model for image information representation. In *Proc. ACM Intern. Conf. on Information and Knowledge Management (CIKM)*, November 1993.

6. D. Woelk, W. Kim, and W. Luther. An object-oriented approach to multimedia databases. *ACM Databases*, 1986.

7. E. Oomoto and K. Tanaka. OVID: Design and implementation of a video-object database system. *IEEE Trans. Knowl. Data Engin.*, August, vol. 5 1993.

8. H. van de Waal. *ICONCLASS. An iconographic classification system.* Koninklijke Nederlandse Akademie can Wetenschappen, Amsterdam, 1973 – 1985.

9. M. Davis. Media Streams: An iconic language for video annotation. *Telektronik 4.93: Cyberspace*, 1993.

10. P. P.-S. Chen. The Entity-Relationship Model – toward a unified view of data. *ACM Databases*, 1(1):9 – 36, March 1976.

11. T. J. Teory, D. Yang, and J. P. Fry. A logical design methodology for relatioal databases using the Extended Entity-Relationship Model. *ACM Computing Surveys*, 18(2):197 – 222, June 1986.