SEMANTICS OF MULTIMEDIA IN MPEG-7

Ana B. Benitez [a], Hawley Rising [b], Corinne Jörgensen [c], Ricardo Leonardi [d], Alesandro Bugatti [d], Koiti Hasida [e], Rajiv Mehrotra [f], A. Murat Tekalp [g], Ahmet Ekin [g], Toby Walker [b]

[a] Dept. of Electrical Engineering, Columbia University, New York, ana@ee.columbia.edu
[b] Media Processing Division, Sony Electronics, San Jose, {hawley.rising@am, tobyw@usrl}.sony.com
[c] School of Informatics, University at Buffalo, SUNY, Buffalo, cjorgens@acsu.buffalo.edu
[d] University of Brescia, Brescia, {leon, bugatti}@ing.unibs.it
[e] Cyber Assist Research Center, AIST and CREST, JST, Tokyo, hasida.k@aist.go.jp
[f] Research & Development, Eastman Kodak Company, Rochester, mehrotra@image.kodak.com
[g] Dept. of Electrical and Computer Eng., University of Rochester, Rochester, {tekalp, ekin}@ece.rochester.edu

ABSTRACT

In this paper, we present the tools standardized by MPEG-7 for describing the semantics of multimedia. In particular, we focus on the Abstraction Model, entities, attributes and relations of MPEG-7 semantic descriptions. MPEG-7 tools can describe the semantics of specific instances of multimedia such as one image or one video segment but can also generalize these descriptions either to multiple instances of multimedia or to a set of semantic descriptions. The key components of MPEG-7 semantic descriptions are semantic entities such as objects and events, attributes of these entities such as labels and properties, and, finally, relations of these entities such as an object being the patient of an event. The descriptive power and usability of these tools has been demonstrated in numerous experiments and applications, these make them key candidates to enable intelligent applications that deal with multimedia at human levels.

1. INTRODUCTION

In recent years, there has been a major increase in available multimedia and in technologies to access the multimedia. However, the extraction of useful information from multimedia and the application of this information in practical systems such as multimedia search engines are still open problems. The most important barrier has been the lack of a standard, comprehensive, and flexible representation of multimedia that enables scalable, intelligent and interoperable multimedia applications.

MPEG-7 has standardized tools for describing different aspects of multimedia at different levels of abstraction. It has the potential to revolutionize current multimedia representation and applications [3]. The MPEG-7 framework consists of Descriptors (Ds) and Description Schemes (DSs) that represent features of multimedia, and more complex structures grouping Ds and DSs, respectively. In particular, the MPEG-7 standard includes tools that describe visual features (e.g., color), audio features (e.g., timbre), structure (e.g., moving regions and video segments), semantics (e.g., objects and events), management (e.g., creator and format), collection organization (e.g., collections and models), summaries (e.g., hierarchies of key frames) and, even, user preferences (e.g., for search) of multimedia.

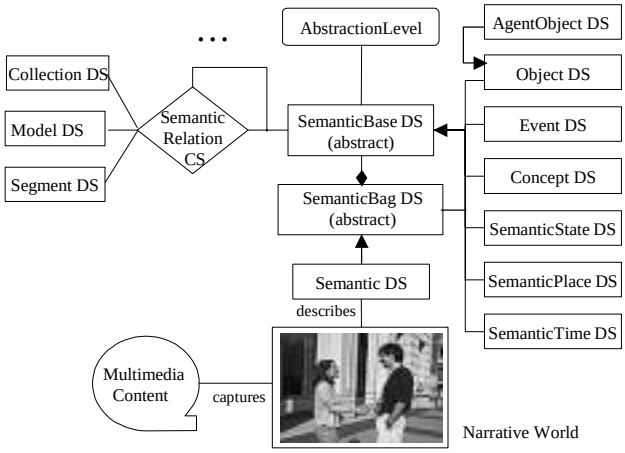

Figure 1: MPEG-7 tools for describing semantics of multimedia.content. The figure uses UML and ER conventions.

This paper presents the MPEG-7 tools for describing the semantics of multimedia [2], the Semantic DS and its components, which are shown in Figure 1. These tools can be best understood by analyzing how general semantic descriptions of anything are constructed. One way to describe semantics is to start with events, understood as occasions upon which something happens. Objects, people and places can populate such occasions, as well as the times at which they occur. Furthermore, these entities can have properties, and states through which they pass as what is being described transpires. There are the interrelations among these entities. Finally, there is the world in which all of this is going on, the background, the other events and other entities, which provide context for the description. Since the ultimate goal is to describe the semantics of multimedia, and since much of

these semantics are narrative in quality (as in a movie that tells a story), in MPEG-7 we refer to the participants, background, context, and all the other information that makes up a single narrative as a "narrative world". One piece of multimedia may have multiple narrative worlds, or vice versa. An example of a semantic description of an image is illustrated in Figure 2. In this example, two persons, an event, a place, a time and a concept depicted or symbolized in the image are described together with several relationships among them.

The components of the semantic descriptions roughly fall into entities in narrative worlds, their attributes and their relations. In this paper, sections 3, 4 and 5 describe the MPEG-7 tools for describing such entities, attributes and relations, respectively. We start by introducing in section 2 the Abstraction Model of MPEG-7 semantic descriptions.

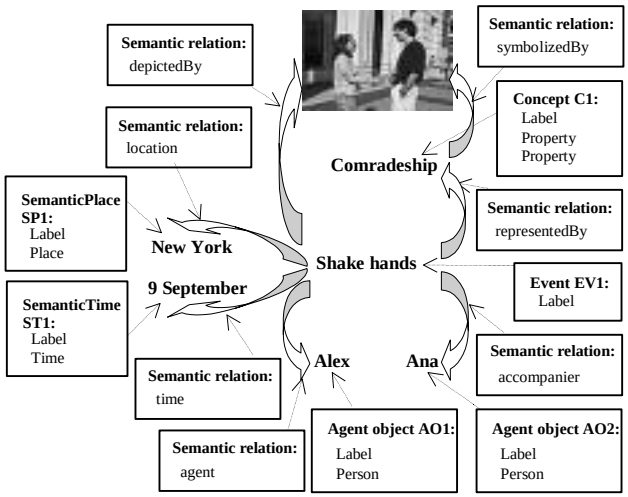

Figure 2: Example of a semantic description of an image.

2. THE ABSTRACTION MODEL

Besides the semantic description of specific instances of multimedia (see Figure 2), MPEG-7 semantic description tools also allow the description of abstractions and abstract quantities. When dealing with the semantics of multimedia, it becomes necessary to draw upon descriptive techniques such as abstraction because these techniques are part of the way the world is understood and described by humans. An abstraction, or more precisely, a lambda abstraction, in logic, replaces one or more of the constant expressions in a statement by a variable [4]. Abstraction thus replaces a single instance with a more general class. In describing media, two types of abstractions can occur: media abstractions and formal abstractions (see AbstractionLevel datatype in section 4). Moreover, MPEG-7 also provides ways of representing abstract properties and concepts that do not arise by an abstraction.

A media abstraction is a description that has been separated from a specific instance of multimedia, and can describe all instances of multimedia that are sufficiently similar (similarity depends on the application and on the detail of the description). If we consider the description of the image shown in Figure 2 ("Alex is shaking hands with Ana in New York on the 9th of September"), a media abstraction is the description "Alex is shaking hands with Ana in New York on the 9th of September" with no links to the image and that is, therefore, applicable to any image or video depicting that event. The variable in a media abstraction is the media itself.

A formal abstraction describes a pattern that is common to a set of examples. The common pattern contains placeholders or variables to be filled in that are common to the set. A formal abstraction may be formed by gathering a set of examples, determining the necessary placeholders from these examples, and thus deriving the pattern, or the other way around. The description is a formal abstraction as long it contains variables, which, when filled in, using a semantic relation (see section 5), would create either media abstractions or concrete descriptions. A formal abstraction of the semantic description in Figure 2 could be "Alex is shaking hands with any woman in New York on the 9th of September", where "any woman" is the variable of the formal abstraction.

MPEG-7 also provides methods of describing abstract quantities such as properties (see the Property element in section 4) and concepts (see the Concept DS in section 3). For instance, the property "Ripeness" can be an attribute of the object "Banana". Concepts are collections of properties that define a category of entities but cannot fully characterize it, i.e., abstract quantities that are not the result of an abstraction such as "Comradeship" in Figure 2. An abstract quantity may be describable only through its properties, and therefore is not abstract because it comes from replacement of elements of a description by generalizations or variables. Note that the same abstract quantity could appear as a property or as a concept in a description. In order to describe relationships of a property, or allow multiple semantic entities to be related to a single property, or specify the strength of a property, the property or group of properties must be described as a concept using the Concept DS.

3. SEMANTIC ENTITIES

The MPEG-7 semantic entity tools describe narrative worlds and semantic entities such as objects, events, concepts, states, places and times. The semantic entity tools are derived from the SemanticBase DS, which is an abstract type that represents any semantic entity including a narrative world (see Figure 1). Semantic entities involved in narrative worlds can change status or functionality, and therefore be described by different semantic entity tools. For instance, a picture can be an object (when it hangs in a museum) or a location ("In the picture"). If the same entity is described by multiple semantic entity tools, the descriptions can be related with a semantic relation (see section 5).

A narrative world in MPEG-7 is represented using the Semantic DS, which is described by a number of semantic

entities and of graphs of their relationships. The Semantic DS is derived from the SemanticBag DS, which is an abstract type representing any kind of collection of semantic entities and their relationships. The Semantic DS represents one type of those collections describing a narrative world (e.g., world "Maya Mythology" in the engravings on a Maya vessel).

The SemanticBase DS is an abstract type that holds common functionalities shared in describing any semantic entity: labels used for categorical searches, a textual description, properties, links to and descriptions of the media, and relationships to other entities. Although the Semantic DS allows textual descriptions of content, the real power of the Semantic DS lies with the ability to describe semantic entities and their relationships in a manner that supports reasoning and inference [4]. In order to facilitate the use of a description of one narrative world as context for another, or to allow a narrative world to be embedded in another, the Semantic DS (via the SemanticBag DS) is also derived from the SemanticBase DS. Other specialized SemanticBase DSs are the Object, AgentObject, Event, SemanticPlace, SemanticTime, SemanticState and Concept DSs. These represent entities that populate the narrative world such as an object, agent and event; the where and when of things; a parametric entity, and a concept, respectively.

The Object and Event DSs describe perceivable semantic entities (objects and events) that can exist or take place in time and space in narrative worlds. The Object DS and Event DS are recursive to describe the subdivision of objects and events into other objects and events. The AgentObject DS extends from the Object DS to describe an object that acts as a person, a group of persons or an organization in a narrative world. For example, the semantic description in Figure 2 shows one event and two agent objects corresponding, respectively, to "Shake hands" (EV1), and to "Alex" (AO1) and "Ana" (AO2).

The SemanticPlace and SemanticTime DSs describe a location and time in a narrative world, respectively. The event "Shake hands" in Figure 2 has an associated location, “New York”, and time, “9 September”. The SemanticPlace and SemanticTime DSs can also describe lengths and durations. Notice that these are semantic descriptions of place or time, not necessarily numerical in any sense (e.g., “Near New York” and “Around midnight”). They serve to give extra semantic information beyond pinpointing a location or instant. The SemanticPlace and SemanticTime DSs can describe locations and times that are fictional or that do not correspond to the locations and times where the multimedia was recorded. As an example a movie could have been recorded in Vancouver during winter but the action is supposed to take place in New York during fall.

The SemanticState DS describes the state or parametric attributes of a semantic entity at a given place and/or time in the narrative world. The SemanticState DS can describe the changing of an entity’s attributes’ values in space and time. An example of the state of an object is the piano's weight; an example of the state of a time is the cloudiness of a day. It can also be used to parametrize the strength of a relationship between entities (e.g., the ripeness of a banana).

Finally, the Concept DS forms part of the abstraction model (see section 2). The Concept DS describes concepts as collections of one or more properties. A concept is a collection of properties that does not characterize any category of entities, i.e., is not the result of some abstraction. An example of a concept is any affective property of the media, for instance, “Suspense”, or a quantity not directly portrayed by the media, like “Ripeness”. A concept can also represent the embodiment of an adjectival quality such as "Happiness” and "Comradeship" (see Figure 2).

4. SEMANTIC ATTRIBUTES

In MPEG-7, semantics entities can be described by labels, by a textual definition, or in terms of properties or of features of the media segments where they occur. Other semantic attribute tools describe abstraction levels and semantic measurements in time and space.

The label has a similar functionality to a "descriptor" or "index term" in Library and Information Science. It is a type used for classifying and retrieving the descriptions of semantic entities (e.g., "Man" for object “Alex” in Figure 2). A semantic entity can have multiple labels, one for each "index term". The textual definition of a semantic entity can be done with either free text or structured text (e.g., the definition of “Alex” in Figure 2 could be "bipedal primate mammal anatomically related to the great apes but distinguished by notable brain development "). The Property element describes a quality or adjectival property associated with a semantic entity (e.g., "Tall" and "Slim" for object “Alex” in Figure 2). This is one way of describing abstract quantities in MPEG-7 (see section 2). The MediaOccurrence element describes a specific appearance or occurrence of a semantic entity in multimedia with the spatial and/or temporal location of the semantic entity in the media together with audio-visual features of these media segments. Three types of semantic entity occurrences are considered: a semantic entity can be perceived, referenced (the subject of), or symbolized in the media.

The AbstractionLevel datatype in the SemanticBase DS describes the kind of abstraction that has been performed in the description of a semantic entity. When it is not present in the description of a semantic entity, then the description is concrete – it describes the world depicted by a specific instance of multimedia and references the multimedia by the MediaOccurrence element or relations to media segments. If the AbstractionLevel element is present in the description, a media abstraction or a formal abstraction is in place. Media abstractions have the dimension attribute of the AbstractionLevel element set to zero. Formal abstractions that abstract specific semantic entities get a dimension attribute one. Higher values of the dimension attribute are used for abstractions of abstractions, such as structural abstractions in which the elements of a graph are not variables, but classes of variables. Examples of concrete and abstract descriptions are discussed in section 2.

Finally, the Extent and Position datatypes in the SemanticPlace and SemanticTime DSs describe semantic

measurements in space or time, respectively. The Extent datatype describes the size or extent of an entity with respect to a measurement type (e.g., "4 miles" and "3 days"). The Position datatype describes the position of an entity in semantic terms by indicating the distance and direction or angle from an origin (e.g., "4 miles from New York City" and "The 3rd day of April").

5. SEMANTIC RELATIONS

MPEG-7 has standardized common semantic relations such as *agent* (normative relations are in italics) but it allows the description of non-normative relations. The semantic relation tools include the SemanticRelation CS, which specifies semantic relations that apply to entities that have semantic information. As shown in Figure 2, the normative semantic relations may apply to entities such as semantic entities, collections, models, and segments, among others. Figure 1 shows examples of several semantic relationships between the objects, the event and the concept in the description.

Normative semantic relations may describe how several semantic entities relate in a narrative or story (e.g., *agent*, *patient*, and *accompanier*). For example, semantic relations may describe relations between objects and/or events that draw upon typical linguistic relations such as an object being the agent, patient, experiencer, accompanier, stimulus, cause, goal, or instrument of an event, or an event being the result or a summary of another event, among others.

Other normative semantic relations may describe how the definitions of several semantic entities relate to each other (e.g., *combination*, *specializes* and *exemplifies*). As examples, semantic relations may describe a semantic entity being the combination of the meanings of two or more semantic entities, or a semantic entity being an specialization, similar in meaning to, opposite in meaning to, or an example of another semantic entity, among others.

There are semantic entities that may also describe the spatial, temporal or media localization of semantic entities in segments, models, collections or other semantic entities (e.g., *depicts*, *symbolizes* and *context*). For example, semantic relations may describe an event taking place at a specific place and time; an object being perceived, referenced, or symbolized in a segment or collection; or a narrative world being the context or interpretation of another, among others.

Finally, there are MPEG-7 tools for describing graph (e.g., *equivalent*), basic (e.g., *member*), spatial (e.g., *left*), and temporal (e.g., *before*) relationships that can also describe relations among semantic entities. See [2] for the complete list of normative relations.

Figure 3 illustrates the use of the semantic relation tools. This figure shows the semantic description of a Maya vessel [1]. The textual description of the picture could be as follows: “The vessel is an example of Maya art created in Guatemala in the 8th Century. The vessel’s height is 14 cm and it has several paintings. The paintings show the realm of the lords of death with a death figure that dances and another figure that holds an axe and a handstone. The paintings represent sacrifice”. The semantic description comprises several semantic entities described using DSs as indicated within parenthesis in Figure 3 and several semantic relations.

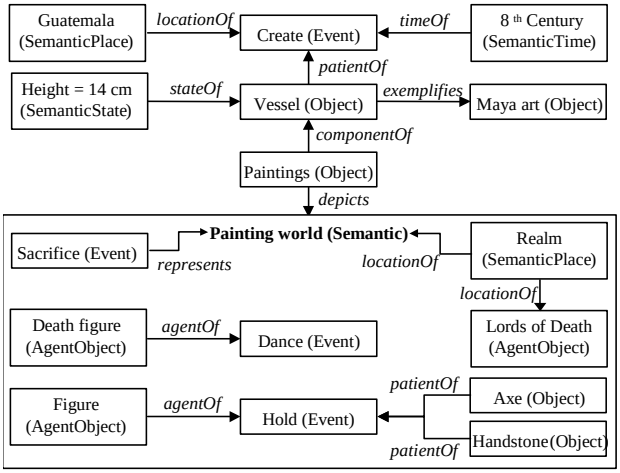

Figure 3: Semantic description of a Maya vessel [1].

6. CONCLUSIONS

MPEG-7, by providing a comprehensive suite of tools for describing the semantics of multimedia, has the potential to revolutionize current multimedia applications. In particular, these tools can describe the semantic entities, attributes, and relationships of multimedia at different abstraction levels and can support powerful and intelligent applications that search and filter multimedia at humanly understandable levels.

7. ACKNOWLEDGEMENTS

The Semantic DS presented in this paper is the result of the contributions and collaborative efforts of many people. The authors are particularly grateful to the members of the MPEG MDS Group for their many contributions in the development of the Semantic DS in MPEG-7.

8. REFERENCES

[1] Getty Research Institute, Getty Standards Program, Categories for the Description of Works of Art. J. Paul Getty Trust and the College Art Association, "Cataloging Examples: Maya Vessel", http://www.getty.edu/research/institute/standards/cdwa/3_catloging_examples/index.html, Accessed on 6/1/01.

[2] MPEG MDS Group, “Text of ISO/IEC 15938-5 FDIS Information Technology –Multimedia Content Description Interface – Part 5 Multimedia Description Schemes", ISO/IEC JTC1/SC29/WG11 MPEG01/N4242, Sydney, July 2001.

[3] MPEG Requirements Group, “MPEG-7: Context, Objectives and Technical Roadmap, V.12”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2861, Vancouver, July 1999.

[4] S. J. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach”, Prentice Hall, Englewood Cliffs, NJ, 1995.