A COMPONENT-BASED MULTIMEDIA DATA MODEL

Alejandro Jaimes
FXPAL Japan, Fuji Xerox Co., Ltd.

ABSTRACT

In this position paper I propose a conceptual model for multimedia that consists of *physical* (from layout to cameras), *conceptual* (e.g., meeting types, actors), *sensory* (audio-visual capture), and *content* (syntax and semantics) components. I argue that solving the multimedia content analysis problem requires a model such as the one presented, that includes the interrelationships between sensors, physical space, conceptual structure, and content, with context as the underlying foundation that determines the parameters of the components as well as their interrelationships. This model should constitute a starting point in considering the different components that affect multimedia content production and analysis.

Categories and Subject Descriptions

H.3.1. [Content Analysis and Indexing]: Abstracting methods

General Terms

Algorithms, Design, Theory

Keywords

Ergonomics, Computer Vision, posture, ergonomics.

1. INTRODUCTION

The field of multimedia has grown tremendously in the last few years, spanning several related research areas (computer vision, networking, human-computer interaction, etc.). Recently, however, lower hardware costs have made the integration of multiple kinds of sensors a reality (e.g., cameras, motion sensors, etc.). In spite of this, little work has been done on constructing models that help us understand the relationship between the different elements involved in multimedia content creation. I will argue that considering all of these components is important because they strongly affect the multimedia data that is obtained, therefore placing constraints that can be used by content-analysis algorithms.

The basic model proposed in this paper assumes that in the content creation chain, there is a *scene* we wish to record using a combination of *sensors*. In the model, therefore, we are only concerned with content creation and the factors that influence the outcome. We define a scene as any physical space, and capture, as the action of recording anything in the scene using any combination of sensors (camera, microphone, motion sensor, haptic sensor, etc.).

CONVEY '05, November 6–11, 2005, Singapore.

2. A MULTIMEDIA MODEL

We define multimedia content as content that includes 2 or more modalities (communication channels). The proposed multimedia model (Figure 1) consists of four components: *physical layout, conceptual structure, sensory acquisition*, and *content*. The physical component models the objects and layout of the scene. The conceptual component models the domain-specific structure of the actors and events in the scene. The sensory component models the capture of the scene using multiple sensing devices (cameras, microphones, etc.). The four components of the model are directly linked by a contextual mesh, defined as the set of conditions under which the content is captured. The circle in the center indicates that the semantics of the content are directly influenced by all of the components.

The model was first applied to meeting videos [4], but the discussion here extends to general multimedia capture.

2.1. Conceptual Component

The *conceptual structure* element models information about the conceptual aspects of the scene: the types of events that take place and their structure, and the actors and their possible actions. For example, in baseball, a pitching event is followed by only one of a limited number of possible events (e.g., strike, base hit, etc.). The pitcher has a particular *role* that constraints his actions (cannot hit the ball, can only throw it). As another example, in a panel, the structure (who speaks and when)

is well defined, and the moderator or chair controls the floor.

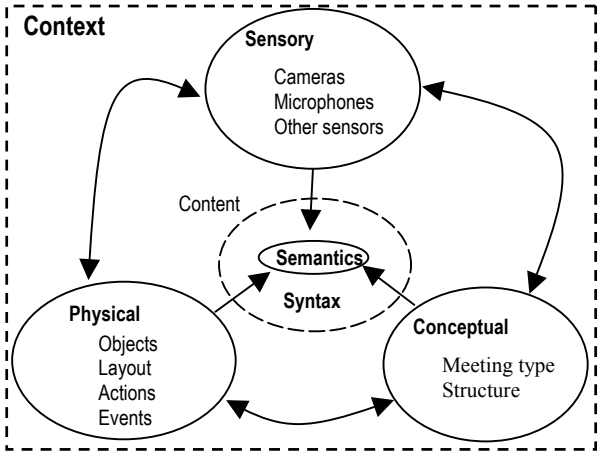

Figure 1. The multimedia model.

2.2. Physical Component

The physical component models the *physical structure* of the scene being captured. This is important because the physical configuration places important constraints on the acquired content. In a smart meeting room scenario, for instance, the location and layout of tables could be modeled here. In a soccer video, for example, there are physical elements that are an intrinsic part of the scene: the goal posts, the lines, the grass, and the audience.

2.3. Sensory Component

The sensory component models all of the sensors used to capture the data. This includes parameters (e.g., camera settings, microphone settings, etc.), number, and locations. The placement of the sensors has a great impact on the content, and often follows known production rules. In the baseball example, again, cameras are usually placed in similar locations in different stadiums by different broadcasters.

2.4. Multimedia Content Layers

The captured content can be interpreted at many different levels. Although it is well known that there are important differences between syntax and semantics, the relationship between the layers is not well understood. One option to represent such levels is the pyramid of Figure 2 [4] (see other alternatives in [3] and details in [5]), in which content can be classified into ten levels: type (e.g., color, b/w), global distribution (i.e., measures taken globally, such as color histogram over an entire image), local structure (i.e., individual components such as lines or circles), global composition (i.e., the leading angle or leading line in an image), generic objects (i.e., car, house, etc.), generic scene (e.g., indoor, outdoor), specific objects (i.e., individually named objects), specific scene (i.e., individually named scene), abstract object (i.e., what the objects are *about*), and abstract scene (i.e., what the scene is about).

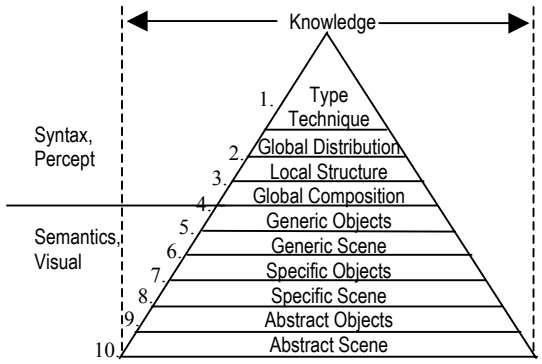

Figure 2. Multi-level indexing pyramid from [4].

One way to use the pyramid is to consider the changes generated by the different components of the model at different levels in each medium (*or* in the content as a whole). For example, a small difference in the local structure of two images can mean a large semantic difference, thus the layer of largest change would be one of the scene layers.

The main point is that measures of similarity can vary greatly at different levels; two images of a tree might look the same to the naked eye, but any motion of the branches due to wind can mean a big dissimilarity between the two images at the pixel level. The same is true if the camera parameters are changed. An in-depth discussion on such changes, in the case of determining similarity between near duplicate images can be found in [5]. The same principles apply to other types of content.

2.5. Context & Memory

Context is defined as a mesh that connects all of the underlying components. In particular, it is as a set of conditions under which the different components are used to create or process multimedia content.

Context is crucial because it defines the constraints for each of the elements of the model. In this sense, context is included in the constraints of each of the subcomponents. One important notion, however, is that of history—while the elements represent the current structures (sensory, physical, conceptual, and content), context includes information about past (and future) events. This may not be represented in the elements themselves, but it may be included in other information sources (e.g., metadata, etc.). Again, in the baseball example, one could use the average batting area statistic for a particular player to assign a probability that he will

hit a homerun. This would obviously affect the captured content, but such statistic would not necessarily be included in any of the components above—it would be included in the "contextual mesh" that ties all the elements together.

3. INTERCONNECTIONS

One important aspect of the model that must be addressed is the relationship between the different components and which levels of the content pyramid are affected when there are changes in the components. There are many open issues, and the types of influences between components depend on the specific application. Some discussion on the different interconnections follows (see arrows in Figure 1).

- **Sensory-conceptual.** A change in the sensors does not necessarily affect the conceptual structure. But we can think of scenarios in which this may occur. For example, in a conference, an audience member may have to walk to a microphone to ask a question, or raise his hand in order to get a microphone if a wireless microphone is available.
- **Conceptual-physical.** These two components are tightly linked. For example, the room chosen for a meeting depends largely on the conceptual structure of the meeting (e.g., panel, presentation, brainstorming, etc.). The structure, in turn, may depend on the physical constraints (e.g., number of people at the meeting determine the type of meeting).
- **Physical-sensory.** The sensors used are often clearly chosen to fit the physical constraints of the scene (e.g., cameras in a smart meeting room). At the same time, the sensor parameters are constrained by the physical space (e.g., camera locations, settings, etc.).
- **Influence on content.** The direct connections between the different components and the content exist to emphasize the direct impact they have on the multimedia content. A change in the physical structure may create changes in the conceptual and sensory components. In many cases, however, it will imply direct changes in the content itself.

4. OPEN ISSUES

Although in [4] several of the open issues on content analysis are described, the model constitutes a starting point for multimedia content creation. Some of the issues to explore include the following:

- How do we model the cause-effect relationships between the different components (i.e., the arrows)?
- How do we model the levels at which content is affected with each change in each of the components and the levels at which information is exchanged between components?
- Can we define context as a set of variables or do we need structures (e.g., a graph) to define it?
- How do the user's interactions affect the model?
- Can we model the components manually using ontologies? (e.g., [2][1] or statistical approaches [8]).

5. APPLICATION

The model is currently being applied in the context of a smart meeting room project [4]. In particular, we have made a link between the physical component and the content component by exploiting the location of fixed cameras and the fixed structure of the scene. In the work described in [7], templates (for automatic content analysis) can be constructed via a graphical user interface to detect events in fixed locations (e.g., person standing by a board), or events that have similar visual structure if the cameras are fixed (e.g., we know where persons sit and the visual constraints on a "raise hand" action).

The physical layout is, of course, an important aspect of the framework in [7], as the location of the cameras is determined in conjunction with the location of the tables and chairs. Although we have experimented with multiple setups, we have found that having a fixed physical layout greatly reduces the complexity of the audio-visual analysis.

The conceptual component, in turn, dictates, in our case, that the meetings have a small number of participants (i.e., small research meetings). Thus, the tables are placed in a particular layout to accommodate this structure.

In general, application of the model's components will vary depending on the domain. In a meeting room scenario, for instance, most of the elements in the scene can be controlled, which facilitates exploiting physical constraints. While not all elements of the model may be easily modeled, having this structure does set a framework to consider main elements that have the strongest effect on multimedia content creation.

6. CONCLUSIONS AND FUTURE WORK

I have presented a multimedia content model that consists of *physical*, *conceptual*, *sensory*, and *content* components. The model is merely a starting point as many of the subcomponents have yet to be defined. Future work includes defining each of the components within a particular application and work on the open issues discussed in section 4.

7. REFERENCES

[1] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting Recorder Project: Dialog Act Labeling

Guide," *ICSI Technical Report TR-04-002*, Version 3, October 2003.

[2] A. Hakeem, M. Shah, "Ontology and Taxonomy Collaborated Framework for Meeting Classification," in proc. *ICPR 2004,* Cambridge, UK, August 2004.

[3] L. Hollink, A.Th. Schreiber, B. Wielinga, M. Worring, "Classification of User Image Descriptions," *International Journal of Human Computer Studies* 61/5, pp. 601-626, 2004.

[4] A. Jaimes, and J. Miyazaki, "Building a Smart Meeting Room: From Infrastructure to the Video Gap (Research and Open Issues)," in proc. *1st IEEE International Workshop on Managing Data for Emerging Multimedia Applications (EMMA)* in conjunction with *21th IEEE Conference on Data Engineering (ICDE)* Tokyo, April 2005.

[5] A. Jaimes. *Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information*, Ph.D. Thesis, Department of Electrical Engineering, Columbia University, February 2003

[6] A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual Information at Multiple Levels", in *Internet Imaging 2000, IS&T/SPIE*, San Jose, CA, January 2000.

[7] A. Jaimes, Q. Wang, N. Kato, H. Ikeda, and J. Miyazaki, "Visual Trigger Templates for Knowledge-Based Indexing", *Fifth Pacific Rim Conference on Multimedia (PCM 2004)*, Tokyo, Japan, December 2004.

[8] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models", *Pattern Recognition Letters*, 25(7), pp. 767-775, May 2004.