Multimedia Content Modeling and Personalization

Marios C. Angelides
Brunel University, UK

There's a wealth of possiblities in the realm of personalized multimedia. Personalizing multimedia content, however, is an extensive process that involves extracting and modeling semantic and structural information about the content as well as metadata; building user profiles either manually or automatically through direct obersvation of user behavior; retrieving and filtering content through the user profiles; and adapting the filtered content to fit usage conditions and user preferences.

This is a time-consuming and laborious process, and researchers adopt various approaches to achieve this. However, regardless of which route they follow, the end goal is a more satisfying and personalized experience for the user.

This special issue offers a collection of techniques and applications for multimedia content modeling and personalization. The rest of this article introduces the field of content modeling and personalization.

Content modeling

Most researchers agree that these are the basic ingredients of a content model (not in order of importance or modeling):

- objects depicted in the media stream and their visible and known properties,
- spatial relationships between those objects,
- video segments or clusters that depict the events involving one or more of these objects, and
- the temporal order between those segments or clusters.

Video segmentation or clustering is traditionally the first step toward multimedia content interpretation and modeling. Segmentation or clustering is usually driven by competing factors, such as visual similarity, time locality, coherence, color, motion, and so on. Some researchers adopt cinematic definitions and rules, such as the 180-degree, montage, and continuity rules.

Often we group together video segments or clusters with similar low-level features or frame-level static features (such as keyframes). Over the years, researchers have suggested a plethora of schemes to describe the segments or clusters: scenes with shots, events with actions, events with subevents, and sequences with subsequences, to name just a few.

The segments are then mapped into some kind of structure, such as a hierarchical structure or decomposition with incremental top-to-bottom semantic granularity. Regarding sound content, unless the sound is of primary importance—such as when changing a speaker drives the segmentation or clustering—the audio element accompanies each video cluster or segment.

It's somewhat common to have parts of the original stream appearing in more than one segment. This isn't necessarily a fault in boundary detection techniques; it may be the case that a video sequence serves a dual purpose. If you attempt to segment the video footage of Arnold Schwarzenegger promoting simultaneously himself as an actor in *Terminator 3* and as a politician, it would be impossible to draw fine lines between the scene boundaries of the actor footage and the politician footage.

This kind of footage saw the introduction of the video lens metaphor (that is, alternative user perspectives of the same content) to cope with multiple interpretations or perspectives of multimedia content, thereby adding multiple perspectives to a content model. In the footage, clearly one video lens would be of a user interested in the actor being simply an actor and another video lens would be of a user interested in the actor becoming a politician. The use of lenses eliminates the need to remodel the content each time a different perspective of the same content arises.

However, a video lens isn't an alternative to representing the *temporality between segments or clusters*—rather, it complements it. Once we've

1070-986X/03/$17.00 © 2003 IEEE

Published by the IEEE Computer Society

Table 1. MPEG-7 temporal and spatial relationships.

Temporal Relationships			Spatial Relationships	
Binary	**Inverse Binary**	***N*-ary**	**Relation**	**Inverse relation**
Precedes	Follows	Contiguous	South	North
Co-occurs	Co-occurs	Sequential	West	East
Meets	Met by	Co-being	Northwest	Southeast
Overlaps	Overlapped by	Co-end	Southwest	Northeast
Strict during	Strict contains	Parallel	Left	Right
Starts	Started by	Overlapping	Right	Left
Finishes	Finished by	—	Below	Above
Contains	During	—	Over	Under

segmented or clustered the video according to one or more perspectives, the original video sequence loses its meaning in the new and often nonsequential structure. Thus, it's important to describe the new chronological order of the segments or clusters. MPEG-7's multimedia description scheme expands on the temporal relationships suggested by J.F. Allen[1] (see Table 1). In the previous example, if segment A is about the actor's future acting career and segment B is about running for governor, then if B is contained within A we can represent this as "A contains B." If the two are identical, then we can say "A co-occurs B."

However, the actual media stream content is the objects and their absolute and relative locations in a frame—that is, their spatial relationships to other objects. There are two competing camps with respect to modeling objects:

- Those who use automated image-tracking tools to identify objects in each segment and then use automated extraction tools to separate them from the background.

- Those who undertake the painstaking task of manual or semiautomated modeling.

With the latter, no limit exists for what we can model: what's visible, what isn't, what we can derive, what we know, and so on. It's laborious but it usually ends in rich and multifaceted content models. The former tracks quickly and extracts what's visible and what may be obscured, but it doesn't map or derive from what we know unless it's visible in the footage. Hence, some alternative modeling must be done even with the former. Determining which of the two is the more efficient is currently the subject of debate and many research projects.

Spatial relationships between objects describe the relative location of objects in relation to other objects (rather than their absolute screen coordinates) within the segment. Spatial representations aren't an alternative to screen coordinates; they complement them. Sometimes when it's difficult to derive screen coordinates, a spatial relationship is the only way to model an object's presence.

The spatial relationships between two objects may differ over time within the same segment. In the previous example, the actor giving an interview will be an obvious object to track and have information extracted about it (the actor). Using a semantic description scheme, the object description may include

- what's visible (height, weight, color of hair, color of eyes, clothes worn, and so on),

- what's not but can be made visible (a heart surgery scar on the chest),

- what we can derive (he may be running for governor of California), and

- what we already knew (he's Austrian, in his mid-50s, married with kids, and so on).

The growth of digital video often produces multiple media streams relevant to a content model that, in turn, may cause filtering to yield more than one relevant video segment. *Video summarization* and *video indexing* may solve both the raw digital video and finer tuning of video segment filtering. The former generates video summaries or skims either as a storyboard of video segment images and keyframes or as a dynamic skim of shortened video segments. The latter generates content-based indexes by using low-level visual features such as color and texture and high-level semantic features such as objects and events. We can use MPEG-7's structure description scheme and semantic description scheme to describe the structural and semantic information of multimedia content.

Content filtering

Filtering techniques analyze content information and prepare presentation of content recommendations using either one or a combination of rule-based, content-based, and collaborative filtering agents. *Rule-based filtering* works with rules derived from statistics such as user demographics and initial user profiles. The rules determine the content that a user receives. Both the accuracy and the complexity of this filtering increase proportionally with the number of rules and the richness of the user profiles. Hence, a major drawback is that it depends on users knowing in advance what content might interest them. Consequently, with this filtering, the accuracy and comprehensiveness of both the decision rules and the user modeling are critical success factors.

Content-based filtering chooses content with a high degree of similarity to the content requirements expressed either explicitly or implicitly by the user. Content recommendations rely heavily on previous recommendations. Hence, a user profile delimits a region of the content model from which all recommendations will be made. This filtering is simple and direct but it lacks serendipity; content that falls outside this region (and the user profile) could be relevant to a user but it won't be recommended. As with rule-based filtering, a major drawback is that the user requirements drive the process. Hence, this filtering combines the challenges of knowledge engineering and user modeling.

With *collaborative filtering* every user is assigned to a peer group whose members' content ratings in their user profiles correlate to the content ratings in the individual's user profile. Content is then retrieved on the basis of user similarity rather than matching user requirements to content. The peer group's members act as recommendation partners. With this filtering, the quality of filtered content increases proportionally to the user population size, and since the matching of content to user requirements doesn't drive filtering, collaborative recommendations don't restrict a user to a region of the content model. One major drawback is the inclusion of new, and hence, unrated content in the model. It may take time before other users see and rate the content. Also sometimes users who don't fit into any group end up being included because of unusual requirements.

Some researchers are developing hybrid-filtering techniques on an ad hoc basis with the goal of combining strengths and solving weaknesses. For example, a collaborative content-based hybrid eradicates the problems of new, unrated content with collaborative filtering and content diversity via content-based filtering.

Content adaptation

Adaptation may require communicating filtered multimedia content through different interconnected networks, servers, and clients that assume different quality of service (QoS), media modality, and content scalability (spatial and temporal). This will either require real-time content transcoding—if what's required is changing a multimedia object's format on the fly into another—or prestored multimodal scalable content with variable QoS (or a hybrid).

We can achieve this through a combination of MPEG-7 and MPEG-21 capabilities. MPEG-7's variation description scheme enables standardized scalable variations of multimedia content and metadata for both summarization and transcoding. While transcoding may transform the spatial and temporal relationships as well as an object's code, color, and properties (or even remove completely nonessential objects), it seeks to preserve the content model semantics because it's semantic-content sensitive.

With *intramedia transcoding* content semantics are usually preserved, because no media transformation takes place. However, content semantics preservation then guides the process, because media are being transformed from one form to another (for example, from video to text). In this case, while the visual perception of an object might change as a result, the object's semantics should be preserved in the new medium. MPEG-21's Digital Item adaptation (DIA) enables standardized description of a digital object—including metadata—as a structured Digital Item independent of media nature, type, or granularity. Consequently, we can transform the object into, and communicate it through, any medium. MPEG-21 supports standardized communication of Digital Items across servers and clients with varied QoS.

Advancements in this special issue

One year after submission and after two rounds of blind reviewing, four articles have finally made it through to the special issue. Each of these four articles contributes in its own and unique way to the advancement of research and development in the field and to the growth of the community.

Geigel and Loui demonstrate the use of content modeling and the application of genetic algorithms for personalizing multimedia content in an interactive, Web-based, photo album prototype. First, they model multimedia content as events and subevents. The genetic algorithm then distributes the modeled content among album pages and adapts the layout of each album page according to the user preferences.

In a similar fashion, Lim, Mulhem, and Tian propose a content modeling scheme and learning technique for content personalization, demon-

Further Reading

The following materials (listed alphabetically) should be helpful to anyone wanting to learn more about multimedia content modeling and personalization:

- N. Adami et al., “The ToCAI Description Scheme for Indexing and Retrieval of Multimedia Documents,” *Multimedia Tools and Applications*, vol. 14, no. 2, 2001, pp. 153-173.
- N. Bryan-Kinns, “VCMF: A Framework for Video Content Modelling,” *Multimedia Tools and Applications*, vol. 10, no. 1, 2000, pp. 23-45.
- A. Cavallaro, O. Steiger, and T. Ebrahimi, “Multiple Video Object Tracking in Complex Scenes,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 523-532.
- I. Cingil, A. Dogac, and A. Azgin, “A Broader Approach to Personalization,” *Comm. ACM*, vol. 43, no. 8, 2000, pp. 136-141.
- R. Cucchiara, C. Grana, and A. Prati, “Semantic Transcoding for Live Video Server,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 223-226.
- S.H. Ha, “Helping Online Customers Decide through Web Personalization,” *IEEE Intelligent Systems*, vol. 17, no. 6, 2002, pp. 34-43.
- H. Hirsh, C. Basu, and B.D. Davison, “Learning to Personalize,” *Comm. ACM*, vol. 43, no. 8, 2000, pp. 102-106.
- J. Kramer, S. Noronha, and J. Vergo, “A User-Centred Design Approach to Personalization,” *Comm. ACM*, vol. 43, no. 8, 2000, pp. 45-48.
- Y.F. Ma et al., “A User Attention Model for Video Summarization,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 533-542.
- P. Maglio and R. Barrett, “Intermediaries Personalize Information Streams,” *Comm. ACM*, vol. 43, no. 8, 2000, pp. 96-101.
- M.R. Naphande and T.S. Huang, “A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval,” *IEEE Trans. Multimedia*, vol. 3, no. 1, 2001, pp. 141-151.
- W. Ngo, T.C. Pong and H.J. Zhang, “On Clustering and Retrieval of Video Shots,” *Proc. 9th ACM Int’l Conf. Multimedia*, ACM Press, 2001, pp. 51-60.
- H. Sundaram, L. Xie, and S.F. Chang, “A Utility Framework for the Automatic Generation of Audio-Visual Skims,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 189-198.
- T. Syeda-Mahmood, “Retrieving Actions Embedded in Video,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 513-522.
- D. Teixeira and Y. Faihe, “In-Home Access to Multimedia Content,” *Proc. 10th ACM Int’l Conf. Multimedia*, ACM Press, 2002, pp. 49-56.

strating their use for home photos. They construct user-labeled event models from home photos and, using relevance feedback, propagate event labels to unlabeled photos.

Doulamis, Doulamis, and Varvarigou explore online learning strategies for personalizing multimedia content. They investigate relevance feedback for developing similarity measures for ranking multimedia content according to user preferences.

Likewise, Wallace et al. suggest neural networks for personalizing multimedia content. They apply neural networks for content filtering and retrieval according to user preferences.

Finally, the “Further Reading” sidebar lists those publications that influenced the writing of this article and provide useful reading in this growing field. **MM**

Reference

1. J.F. Allen, “Maintaining Knowledge about Temporal Intervals,” *Comm. ACM*, vol. 26, pp. 832-843.

Marios C. Angelides is a professor of computing in the Department of Information Systems and Computing at Brunel University. He has more than 10 years of research experience in multimedia information systems. Angelides holds a BSc in computing and a PhD in information systems, both from the London School of Economics and Political Science. He’s the author of *Multimedia Information Systems* (Kluwer, 1997). He’s a member of the ACM, the IEEE Computer Society, and the British Computer Society.

Readers may contact guest editor Marios C. Angelides at the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex UB8 3PH, UK; angelidesm@acm.org.