A Semantic View Mechanism for User-centric Video Adaptation

Dawei Ding[1] Wang Feng[2] Chong Wah Ngo[3] Qing Li[1]
dwding@cityu.edu.hk wfeng@ust.hk cwngo@cs.cityu.edu.hk itqli@cityu.edu.hk
[1]Dept. of Computer Engineering and Information Technology, City University of Hong Kong, HKSAR, China
[2]Dept. of Computer Science, The Hong Kong University of Science & Technology, HKSAR, China
[3]Dept. of Computer Science, City University of Hong Kong, HKSAR, China

ABSTRACT
The concept of video adaptation includes not only adaptation to terminals or networks but also adaptation to user interests in order to maximize their satisfaction. The most emerging trend in video adaptation is *er entri* content adaptation, instead of terminal centric adaptation. To address this problem, a semantic scheme is required to efficiently filter and index the video content. In this paper, we advocate the concept of *Medi iew*, which represents video content using an object-oriented view mechanism. Advanced media views can be customized with view operators and probabilistic decision-tree (PDT) on a set of system-defined basic views. Such a collaborative view-accumulation mechanism provides an approach to model the high-level semantics of video content from low-level features with human aid, and makes semantic video adaptation feasible. Preliminary experimental studies are conducted in the context of personalized video summarization.

Keywords
Multimedia Database, MediaView, Video Adaptation, User-centric Access, Personalized Video Summarization

1. I TR D CTI
Nowadays, the increasing availability of multimedia data creates the need for managing, integrating, retrieving and navigating them using a database approach to fulfill the efficiency and flexibility requirements. In a video management system, efforts should be put into adapting and delivering the content in a semantic way that maximizes the user satisfaction. Unfortunately, although researchers on multimedia databases have addressed the indexing and querying of media contents, relatively little progress is achieved on the semantic modeling of multimedia, which is of essential importance to most multimedia applications. Most existing data models are unable to capture adequately and accurately the semantic aspect of multimedia, which features the following two unique properties:

Context-dependency. Semantics is not a static and inherent property of a media object. (In this paper, a media object refers to a video shot.)

Media-independency. Media objects of different types of modality (i.e., multi-modal objects) may suggest the related semantic meaning.

The dynamic nature of multimedia is fundamentally different from that of the traditional alphanumeric data whose semantics is explicit, unique, and self-contained. The context-dependent nature of media objects needs the ability to switch dynamically among various semantics depending on specific contexts. Earlier in [1], we thus have proposed an advanced view mechanism, MediaView, which models the context-dependent nature well so that multimedia content is easy to be incorporated into different semantic contexts. (Since we focus on the scenario of video adaptation, the media-independency nature of multimedia will not be addressed in the rest of this paper.) To address the problem of video adaptation we have mentioned before, specifically, we propose to incorporate MediaView mechanism into semantic video adaptation that allows users to access contents with individual preference from a semantic level.

In the rest of this paper, we first review some related work in section 2, and present the framework of MediaView for video adaptation in section 3. We present some preliminary implementation results on real-world video collections in section 4. We conclude the paper with directions for future work in section 5.

2. RE ATED RK
The research community has made a lot of efforts in developing more effective methods for accessing video databases recent years. Several content-based video database systems have been proposed ([6], [7]). Most of those systems partition videos into a set of basic units such as shots, and then follow the approach of representing video via a set of feature attributes, such as color, texture, shape, and motion. However those systems could hardly avoid the gap between low-level features and high-level semantics. Since high-level similarity may not correspond to low-level feature based similarity, this should involve understanding the semantics of multimedia content. To address this problem, some work has been done in the field of multimedia modelling. Candan K.S. et al. suggest that a non-interactive multimedia presentation is a set of virtual objects with associated spatial and temporal presentation constraints [3]. The approach could deal with building multimedia presentations whose content changes dynamically based upon queries. Some systems also benefit from the user profiles [4] and the database feedback [5] in deciding which objects are the most relevant for the user and that these objects should be fetched from the database. Boll S. et al. s approach [2] allows for flexible on-the-fly composition of multimedia fragments in order to create individualized multimedia

* The work presented in this paper has been partly supported by a grant from City University of Hong Kong (Project No. 7001564), and substantially by a grant from the Research Grants Council of the HKSAR, China (Project No. CityU 1125/04E).

documents, and for the personalization of multimedia presentations depending on the user profiles. Comparing to these works, however, MediaView provides a more **general purpose framework** to organize the semantic aspects of a multimedia database in a higher level, in addition to the spatio-temporal relationships, and utilize these semantic relations into the database navigation and multimedia document authoring.

3. MEDIAVIE FRAME RK

The incapability of semantic multimedia modeling undermines the value of a database in supporting semantics-intensive applications. This semantic gap between databases and multimedia applications constitutes the major motivation of *Medi iew* as an extended object-oriented view mechanism. This mechanism bridges the semantic gap by introducing above the traditional database architecture an additional layer constituted by a set of modeling constructs named *Medi iew* [1]. Each MediaView, defined as an extended object view, formulates a *t i ed ntext* in which the dynamic and subjective semantics of media objects are properly interpreted. In this section, we present this uniform framework to represent semantic information of video contents in the databases.

3.1 View anguage

The basic unit in the MediaView is a *iew*. *Medi iew* is essentially an extension built on top of an object-oriented knowledge representation model. As in an object-oriented model, a *edi iew* is defined just like a *l* . The formal definition of *Medi iew* as an extended object-oriented view is given as follows:

Definition 1.: *A edi iew n ed M i i repre ented t ple tw ele ent*

$$M_i \quad M_i \; P_i$$

M_i *i et je t t t re in l ded int* M_i *it e er*

P_i *i et iew le el pr pertie ttri te nd et d w i re pplied n* M_i

A sample *iew* is indicated in Figure 1, which describes players in a sports competition, with three view-level properties: country, name and score.

Player View
Country
Name
Score

Figure 1: A sample media view of Player in sports

3.2 View Modeling

The advantage of MediaView, obviously, exists in that it avoids the invocation of the expensive media-processing algorithm each time a query is processed; instead, it accumulates and learns the semantic knowledge among different contexts. However, in many cases, a complex context may be given by users for preferred content, for example, Team China s diving performance in Men s Springboard . In this regard, the MediaView framework provides a bottom-up fashioned mechanism for users to dynamically construct those complex context-based media views, based on existing basic ones.

Oper t r

Besides the operators defined in [1], several user-level operators are devised to support more complex contexts, as follows.

- **Inheritance peration**

This operator creates a media view, which inherits from a MediaView set. When executed successfully, it returns the reference to the created media view, which has all the properties inherited from its super views .

- **nion peration**

Union operations are used to combine the contents of a set of views into a new media view. From the context viewpoint, it acts as an OR logic.

- **Intersection peration**

In MediaView, the intersection operation is used to get the common contents from a set of media views. From the context viewpoint, it acts as an AND logic.

Ad n ed iew e initi n

The above set of operators can at best, however, provide a limited flexibility to define advanced media views using existing simple ones. A subjective context such as the best diving collection could not be deduced only from basic contexts with those operators. Obviously, a user could always make his subjective decision that whether a piece of video content belongs to any uncertain context. Actually, people solve this problem by 1) enumerating the related media views, which affect the decisions, and 2) indicating how to select the video data based on their property values. The function of uncertainty modeling hereby is to replace human decision-making processes with automatic decision-making algorithms. We use a probabilistic decision-tree to take this work.

Definition 2.: *A pr ili ti de i i n tree P T i de i i n tree t t i n pr ility t e t e p i le i e ed n t e ntext t e de i i n P w ere i p i le i e nd i i t ry i e t e ntext t e de i i n*

By assigning a probability distribution to the possible choices, probabilistic decision-trees not only specifies the order of preference for the possible choices, but also gives a measure of the relative likelihood that each choice is the one which should be selected.

4. SER CE TRIC VIDE ADAPTATI

To demonstrate the effectiveness and elegancy of our MediaView framework, we have implemented a prototype based on some real-world sports video collections. The

experimental data was collected from the video collections of 2004 CHAMPIONS DIVING TOUR , which contains 157 minute long video clips of 1663 Mbytes.

Firstly, we segment the videos into shots, which are considered as the *edi je t* in MediaView. Then those shots are annotated according to several different clues, including text, motion and pattern of flag. To complete this task, some video processing technologies ([8], [9]) are used to recognize the text and flag pattern on the video. Simple semantics of the video shots, such as the type of diving, the name of the player, and the score (see Figure 2), are therefore gained to define basic media views. As Figure 2 illustrates, we have defined 4 classes of basic views as the foundation: *Te* , *Pl yer*, *R nd*, and *M t* . Most of the aforementioned basic views are cross-referenced by each other s. Indeed, each view is associated with a set of video shots, and on the flip side, each shot could also associate more than one media views. This properly embodies the context-dependency property of video contents.

4.1 Case Studies

To cater for the interests of different users, we discuss here the problem of semantic content selection for some cases of utilizing the MediaView mechanism. In each case, the procedure of view customization is presented as a diagram, since it provides an intuitive and flexible way for users to manipulate MediaViews with view operators and PDT.

- ***Case 1: China's Springboard performance***

In this first case, user Sue is supposed to only have interests in Team China and she likes the springboard matches. The user interests, therefore, could be conveniently described as a customized view via the view operators, as Figure 2 illustrates. Firstly a *Te C in* view and a *S M t* view are inherited from the basic views *Te* and *M t* , respectively. Secondly we perform an INTERSECTION operator on the views of *Te C in* and *S M t* to construct the target view *C e* . The video shots associated with *C e* are then presented to the user (Sue). Once defined, the new view *C e* is eligible to be stored in the database for further reference.

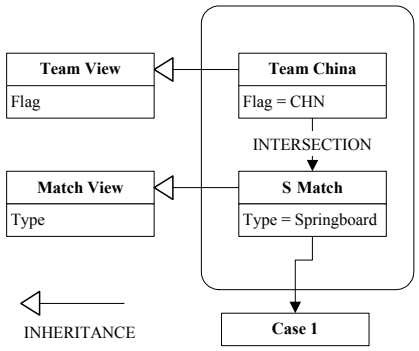

Figure 2: China s Springboard Performance

- ***Case 2: Best performance in each round***

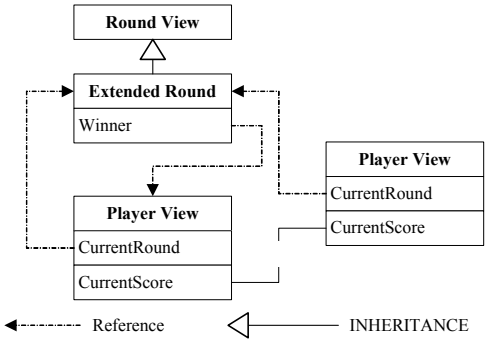

Figure 3: Definition of the *Winner* in a *Round*

In the second case, another user Tom wants to see the best performance in each round. This actually corresponds to the video clips with the highest score in each round. Before defining the target view, to simplify the view

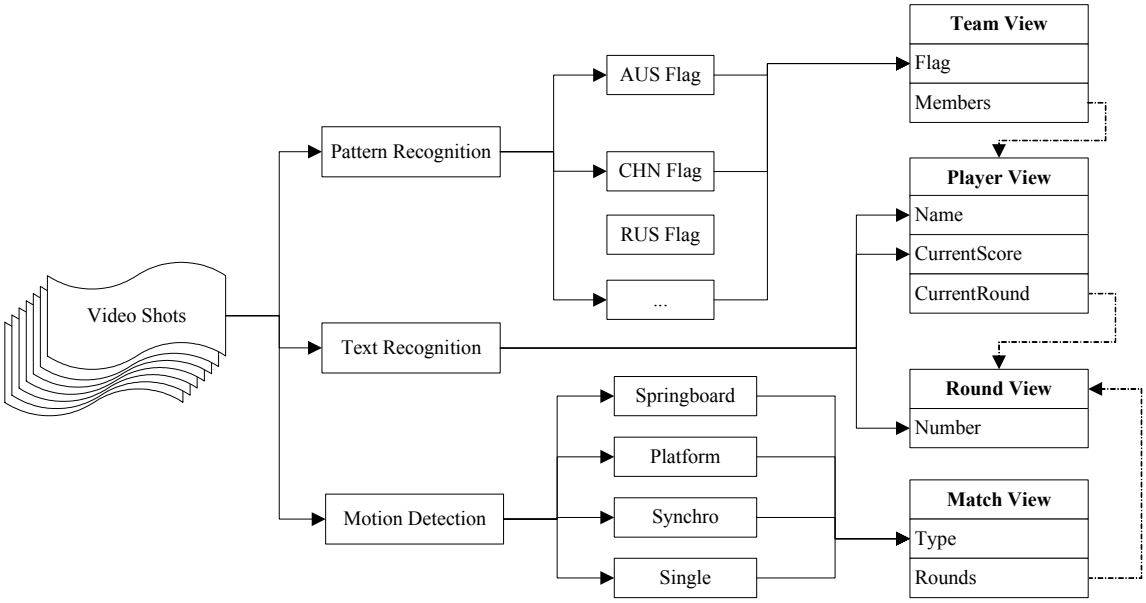

Figure 4: Basic MediaViews based on ow level Feature Clustering

construction, we add a customized reference property *winner* to *R nd* view, which links to a *Pl yer* view. This property indicates the best player, the one with the highest score, in any particular round. As Figure 3 depicts, the formal definition of the *inner* in a *R nd* is as follows:

$$xtendedR\ \ nd.\ \ inner = Pl\ \ yer,$$

$$\text{when } \begin{cases} Pl\ \ yer.C\ \ rrentS\ \ re \geq P.C\ \ rrentS\ \ re \\ Pl\ \ yer.C\ \ rrentR\ \ nd = \ \ xtendedR\ \ nd \end{cases}'$$

where $\forall P \Rightarrow P.C\ \ rrentR\ \ nd = \ \ xtendedR\ \ nd$.

Then, the best performance in each round could be defined as a new view (*C e*) shown in Figure 5.

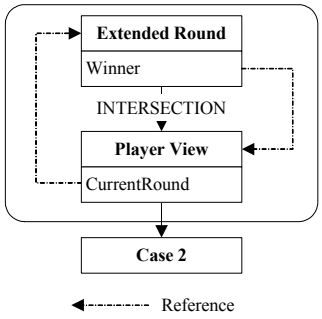

Figure 5: The Best Performance in each Round

- ***Case 3: Highlights in platform matches***

In this third case, user Joe is going to select the contents by a fuzzy concept *i li t* . As it is a subjective concept and there is no general criterion to decide which content belongs to highlights, Joe needs to customize the view with his specific understanding.

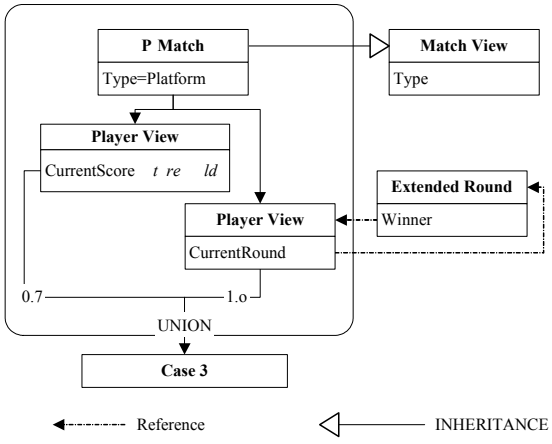

Figure 6: A definition of *Platform Highlights*

Figure 6 gives a **possible** customization of *Pl t r i li t* . The meaning of the definition is described with the aid of PDT, as follows:

1. If the *Pl yer* in a *Pl t r* match is a *R nd inner*, which means the player performs the best in a round, Joe would regard the performance must belong to highlights. Thus the associated contents are included as *i li t* for sure (possibility = 1.0).
2. If the *Pl yer* in a *Pl t r* match gets a high score (bigger than a constant *t re ld*), Joe supposes the performance should likely be part of the highlights to some extent. Thus the possibility of associated contents being included as *i li t* is 0.7.

5. C C SI A D F T RE RK

Aiming at a user-centric model for efficient video adaptation, this paper proposed MediaView, an object-oriented view mechanism that accommodates video content representation using bottom-up inference. Users of MediaView can construct their own media views based on existing views, and/or accumulate customized views into the video database for others reference. We have demonstrated the effectiveness on sports videos.

There are still open research issues remaining in supporting user-centric video adaptation. From the examples presented in the paper, we can see that this mechanism is suitable for sports video because their semantic views can be easily represented. However, videos like movies may not be efficiently represented with several semantics views. It also depends on the development of more accurate and efficient video processing technologies to retrieve the correct semantics. Moreover, customized media views could be collaboratively refined with recorded user feedbacks.

6. REFERE CES

[1] Qing Li, Jun Yang, Yueting Zhuang. MediaView: A Semantic View Mechanism for Multimedia Modeling, IEEE Pacific Rim Conference on Multimedia, 729-736, 2002.

[2] Boll S. et al., ZyX A Semantic Model for Multimedia Documents and Presentations, Semantic Issues in Multimedia Systems, 189-209, 1999.

[3] Candan K.S. et al., View Management in multimedia databases, The VLDB Journal, Vol. 9, 131-153, 2000.

[4] R. Fagin and E.L. Wimmers, Incorporating User Preferences in Multi-media Queries, Int. Conference on Database Theory, 1997.

[5] Y. Rui, T.S. Huang, and S. Mehrotra. Content-based Image Retrieval with Relevance Feedback in MARS, IEEE Int. Conf. on Image Processing, 1997.

[6] Flickner, M., et al. Query by image and video content: the QBIC system. IEEE Comput. 38, 1, 23 31, 1995.

[7] Humrapur, A. et al. Virage video engine. In Proceedings of the 5th Conf. on Storage and Retrieval for Image and Video Databases. 188 197. 1997.

[8] F. Wang, C. W. Ngo & T. C. Pong, Synchronization of Lecture Videos and Electronic Slides by Video Text Analysis , ACM Multimedia Conf., 2002.

[9] C. W. Ngo, T. C. Pong & H. J. Zhang, On Clustering and Retrieval of Video Shots Through Temporal Slices Analysis, IEEE Trans. On Multimedia, Vol. 4, No. 4, 2002.