MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Digital Video Transcoding

Jun Xin, Chia-Wen Lin, Ming-Ting Sun

TR2005-006 January 2005

Abstract

Video transcoding, due to its high practical values for a wide range of networked video applications, has become an active research topic. In this review paper, we outline the technical issues and research results related to video transcoding. We also discuss techniques for reducing the complexity, and techniques for improving the video quality, by exploiting the information extracted from the input video bitstream.

Proceedings of the IEEE

Video Adaptation: Concepts, Technologies, and Open Issues

Shih-Fu Chang, *Fellow, IEEE,* and Anthony Vetro, *Senior Member, IEEE*

***Abstract* — Video adaptation is an emerging field that offers a rich body of techniques for answering challenging questions in pervasive media applications. It transforms the input video(s) to an output in video or augmented multimedia form by utilizing manipulations at multiple levels (signal, structural, or semantic) in order to meet diverse resource constraints and user preferences while optimizing the overall utility of the video. There has been a vast amount of activities in research and standard development in this area. This paper first presents a general framework that defines the fundamental entities, important concepts (i.e., adaptation, resource, and utility), and formulation of video adaptation as constrained optimization problems. A taxonomy is used to classify different types of adaptation techniques. The state-of–the-art in several active research areas is reviewed with open challenging issues identified. Finally, support of video adaptation from related international standards is discussed.**

***Index Terms* — Video Adaptation, Universal Multimedia Access, Pervasive Media, Transcoding, Summarization, MPEG-7, MPEG-21**

I. Introduction

IN pervasive media environments, users may access and interact with multimedia content on different types of terminals and networks. Such an environment includes a rich variety of multimedia terminals such as PC, TV, PDA, or cellular phones. One critical need in such a ubiquitous environment is the ability to handle the huge variation of resource constraints such as bandwidth, display capability, CPU speed, power, *etc*. The problem is further compounded by the diversity of user tasks – ranging from active information seeking, interactive communication, to passive consumption of media content. Different tasks influence different user preferences in presentation styles and formats.

Video adaptation is an emerging field that includes a body of knowledge and techniques responding to the above challenges. A video adaptation tool or system adapts one or more video programs to generate a new presentation with a video or multimedia format to meet user needs in customized situations. Fig. 1 shows the role of video adaptation in pervasive media environments. It takes into account information about content characteristics, usage environments, user preferences, and digital rights conditions. Its objective is to maximize the utility of the final presentation while satisfying various constraints. Utility represents users' satisfaction towards the final presentation and is defined based on application contexts and user preferences.

Video adaptation differs from video coding in its scope and intended application locations. There are a wide variety of adaptation approaches – signal-level vs. structural-level vs. semantic-level, transcoding vs. selection vs. summarization, or bandwidth- vs. power- vs. time-constrained. Adaptation typically takes a coded video as input and produces a different coded video or an augmented multimedia presentation. Another difference is that adaptation is typically deployed in the intermediate locations such as proxy between server and client, although they may be included in the servers or clients in some applications.

There have been many research activities and advances in this field. Earlier work such as [1][2] has explored some interesting aspects of adaptation like bandwidth reduction, format conversion, and modality replacement for Web browsing applications. Recently, international standards such as MPEG-7 [20], MPEG-21 [21][24], W3C [21], and TV-Anytime [22] have developed related tools and protocols to support development and deployment of video adaptation applications.

Despite the burgeoning interest and advances, video adaptation is still a relatively less defined field. There has not been a coherent set of concepts, terminologies, or issues defined over well-formulated problems. This paper serves as a preliminary attempt in establishing part of the foundation that can be used to unify and explore various issues and approaches in this field.

Specifically, in section II we present a general conceptual framework to define the entity, concepts (resource, utility, and adaptation), and their relations from the perspective of video adaptation. Based on the framework, we present a straightforward but systematic procedure for designing video adaptation solutions, as well as a taxonomy of different classes of adaptation technologies. In section III, we review current active research areas in video adaptation, with important open issues discussed in section IV. Support from related international standards is discussed in section V. Finally, conclusions are presented in section VI.

Manuscript received January XX, 2004; revised May XX, 2004.

S.-F. Chang is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (sfchang@ee.columbia.edu).

A. Vetro is with Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA (avetro@merl.com).

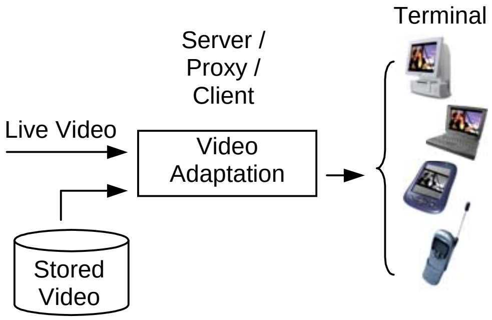


Fig. 1 Role of video adaptation in pervasive media environments to support heterogeneous terminals and networks.

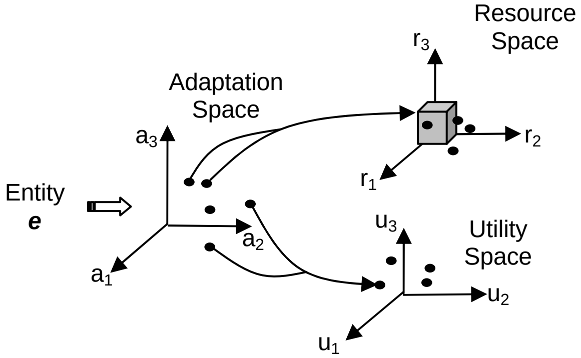


Fig. 2 A general conceptual framework for video adaptation and associated concepts of resources and utility.

II. A Unified Conceptual Framework and Technology Taxonomy

Design of video adaptation systems involves many complex issues. In this section, we first present a general conceptual framework to clarify and unify various interrelated issues, as illustrated in Fig. 2. The framework was based from the one we presented in [3], with extended description of a systematic design procedure and a taxonomy for classifying different adaptation techniques.

First, "***entity***" is defined to refer to the basic unit of video that undergoes the adaptation process. Entities may exist at different levels, such as pixel, object, frame, shot, scene, syntactic components, as well as semantic components. Different adaptation operators can be defined for different types of entities. For example, a video frame can be reduced in resolution, spatial quality, or skipped in order to reduce the overall bandwidth. A semantic component (such as a story in a news program) can be summarized in a visual or textual form. A subset of shots in a sequence may be removed in order to generate a condensed version of the video, i.e., video skims.

Complex entities can be defined by using additional properties. For example, *syntactic entities* like recurrent anchor shots in news, pitching shots in baseball, and structured dialog shot sequences in films can be defined by syntactic relations among elements in the video. *Semantic entities* like scoring events in sports and news stories are caused by real-world events, created by the producer or formed by expectations of the viewers. *Affective entities* are those defined by affect attributes (such as emotion and mood) conveyed by the video elements.

The space of feasible adaptations for a given video entity is called the ***adaptation space***. Note we use the term "space" in a loose way – the coordinates in each dimension represent particular adaptation operations and a point in the space represents a combination of operations from different dimensions. For example, a popular method for transcoding inter-frame transform coded video includes two dimensions: (1) dropping a subset of transform coefficients in each frame and (2) skipping a subset of frames in the video sequence.

Each entity is associated with certain resource requirements and utility values. An adaptation operation transforms the entity into a new one and thus changes the associated resources and utility values. Like the adaptation space, there are multiple dimensions in the ***resource space*** and the ***utility space***. Resources may include transmission bandwidth (i.e., bit rate), display capabilities (e.g., resolution, color depth), processor speed, power, and memory. Here we focus on the resources available in the usage environment or the delivery network. The information describing the usage environment resources (e.g., maximal channel capacity) can be used to derive implicit constraints and limitations for determining acceptable adaptation operations.

The ***utility value*** represents the quality or users' satisfaction of the video content. Utility can be measured in different levels – the objective level (e.g., peak signal-to-noise ratio, PSNR), the subjective level (e.g., subjective scores), and the comprehension level. The comprehension-level utility measures viewers' capability in comprehending the semantic information contained in a video. Measurement of such semantic-level comprehension is difficult as it depends on many factors including users' knowledge, tasks, and domain contexts. In some restricted scenarios, however, it might be possible to come up with measures of generic comprehensibility without deep understanding of the content. Such generic semantics may include generic location (indoor vs. outdoor), people (portrait vs. crowd), time (day vs. night), etc. Again, we use the term, utility space, to represent the multiple-dimensional characteristics of video utility measures.

The utility value of a video entity is not fixed and is heavily affected by the ***user preferences***. This is particularly true for the subjective and semantic-level utilities. The subjective relevance of a video entity depends on the user needs for his current task. The user preferences may also be used to set explicit constraints on the feasible adaptation operations, in addition to the implicit constraints set by the resource limitations described above. For example, if the user prefers to receive video summaries not longer than certain lengths, temporal condensation operations will be needed. Or the user may prefer to view videos within a window no larger than a fraction of the screen size, though the actual display resolution is not a limiting factor for the full-sized video.

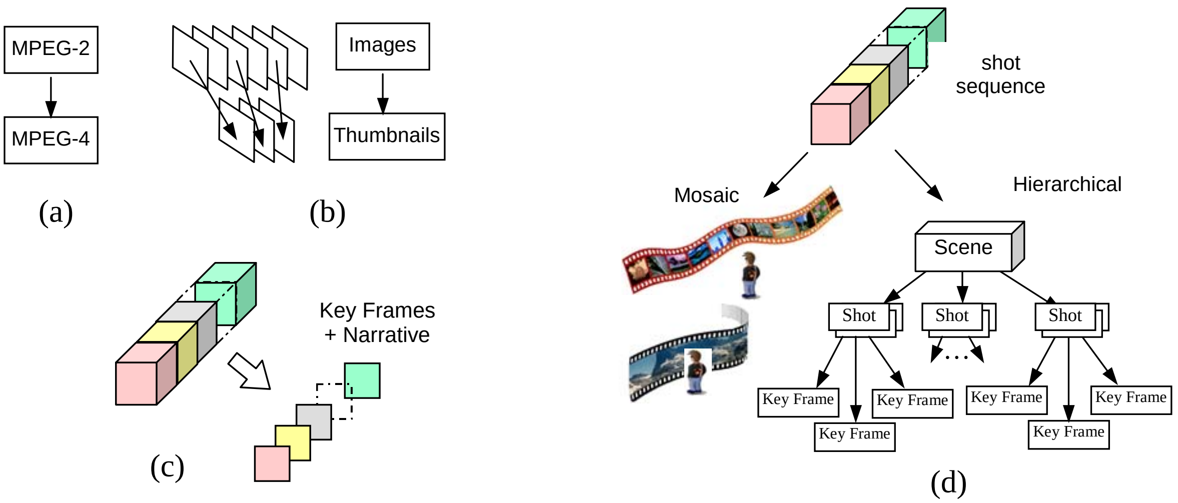


Fig. 3 Taxonomy of video adaptation operations: (a) Transcode, (b) Select/Reduce, (c) Replace, (d) Synthesize.

Given a video entity, the relationships among the adaptation space, the resource space, and the utility space represent critical information for designing content adaptation solutions. For example, in Fig. 2 the shaded cube in the resource space represents the resource constraints imposed by the usage environment. There exist multiple adaptation solutions that satisfy the constraints – we refer to these solutions as the *resource-constrained permissible adaptation set*. Similarly, different adaptation operators may result in the same utility value. Such operators are said to form an *equal-utility adaptation set*. It is such a multi-option situation that makes the adaptation problem interesting – our objective is to choose the optimal one with the highest utility or the minimal resource while satisfying the constraints.

A. Systematic Procedure for Designing Video Adaptation Technologies

The above conceptual framework can be used to guide the design process of practical adaptation solutions. Below, we discuss a systematic procedure that utilizes the concepts and relations of adaptation, resource, and utility.

1. Identify the adequate entities for adaptation, e.g., frame, shot, sequence of shots, etc.
2. Identify the feasible adaptation operators, e.g., re-quantization, frame dropping, shot dropping, replacement, etc., and their associated parameters.
3. Develop models for measuring and estimating the resource and utility values associated with video entities undergoing identified operators.
4. Given user preferences and constraints on resource or utility, develop strategies to find the optimal adaptation operator(s) satisfying the constraints.

With the above procedure, many video adaptation problems can be formulated as follows. Given a content entity (**e**), user preferences, and resource constraints (C_R), find the optimal adaptation operation, $\mathbf{a}_{\mathbf{opt}}$, within the feasible adaptation region so that the utility of the adapted entity **e’** is maximized.

Similar to the above, we can formulate other problems in a symmetric way – exploring the utility-constrained permissible set to find the optimal adaptation operator to satisfy utility constraints while requiring minimal resources.

B. Video Adaptation Taxonomy

Many interesting adaptation operations have been reported in the literature. To help readers develop a coherent view towards different solutions, we present a simple taxonomy based on the type of manipulations performed. Fig. 3 shows illustrative examples of each class of adaptation.

1) Format Transcoding: A basic adaptation process is to transcode video from one format to another, in order to make the video compatible with the new usage environment. This is not surprising when there are still many different formats prevailing in different application sectors such as broadcasting, consumer electronics, and Internet streaming. One straightforward implementation is to concatenate the decoder of one format with the encoder of the new format. However, such implementations may not be feasible sometimes due to the potential excessive computational complexity or quality degradation. Alternate solutions and complexity reducing techniques can be found in [9].

2) Selection/Reduction: In resource-constrained situations, a popular adaptation approach is to trade some components of the entity for saving of some resources. Such schemes usually are implemented by selection and reduction of some elements in a video entity like shots and frames in a video clip, pixels in an image frame, bit planes in pixels, frequency components in transformed representation, etc. Some of these schemes typically are also considered as some forms of transcoding – changing the bit rate, frame rate, or resolution of an existing coded video stream. Reduction involves a selection step to determine which specific components to be deleted. Uniform decimation sometimes is sufficient, while sophisticated methods further explore the non-equal importance of different components based on psychophysical or high-level semantic models. For example, in several video summarization systems, key events (such as scoring in sports) are defined based on user preferences or domain knowledge. During adaptation, such highlight events are used to produce condensed video skims.

3) Replacement: This class of adaptation replaces selected elements in a video entity with less expensive counterparts, while aiming at preserving the overall perceived utility. For instance, a video sequence may be replaced with still frames (e.g., key frames or representative visuals) and associated narratives to produce a slide show presentation. The overall bandwidth requirement can thus be dramatically reduced. If bandwidth reduction is not a major concern, such adaptation methods can be used to provide efficient browsing aids in which still visuals can be used as visual summaries as well as efficient indexes to important points in the original video. Note the replacement content does not have to be extracted from the original video. Representative visuals that can capture the salient information in the video (e.g., landmark photos of a scene) can be used.

4) Synthesis: Synthesis adaptation goes beyond the aforementioned classes by synthesizing new content presentations based on analysis results. The goal is to provide a more comprehensive experience or a more efficient tool for navigation. For example, visual mosaics (or panorama) can be produced by motion analysis and scene construction. The extended view provides an enhanced experience in comprehending the spatio-temporal relations of objects in a scene. In addition, transmission of the synthesized stream usually requires much less bandwidth than the original video sequence since redundant information in the background does not have to be transmitted. Another example of adaptation by synthesis is the hierarchical summary of video, as shown in Fig. 3(d). Key frames corresponding to highlight segments in a video sequence are organized in a hierarchical structure to facilitate efficient browsing. The structures in the hierarchy can be based on temporal decomposition or semantic classification.

In practical applications of adaptation, various combinations of the above classes can be used. Selected elements of content may be replaced with counterparts of different modalities, encoded with reduced resolutions, synthesized according to practical application requirements, and finally transcoded to a different format.

III. Active Research Areas

In this section, we review several active research areas of video adaptation and show how the proposed resource-utility framework can be used explicitly or implicitly to help formulate the optimization of adaptation processes at different levels – semantic, structural, and signal. The chosen areas are not meant to be exclusive. Many interesting combinations or variations exist.

A. Semantic Event-Based Adaptation

Detecting semantic highlights or events in video has attracted much interest in many applications, such as personal multimedia information agent, video archive management, and security monitoring. In the context of video adaptation, the important events are usually defined by the content providers or derived from user preferences, for example, the scoring points in sports video, the breaking news in broadcast programs, and the security breaking events in surveillance video. Following the framework defined in Section II, we can interpret such events as the segments in the video that have the highest semantic-level utilities.

Video analysis for event detection has been an active research area in the community of image processing, computer vision, and multimedia. In [4], information in metadata streams (e.g., closed captions and sports statistics) is combined with video analysis to detect important events and players. Sports statistics are provided by commercially available services. Such data have specific information about the scores, player names, and outcomes of events. However, they may not give complete information about content shown in the video. Recognition of scenes and objects in the audio-visual streams adds complementary information, and more importantly, helps detecting the precise start/end time of events reported in the statistics streams.

In [6], canonical views in sports (e.g., pitching in baseball and serving in tennis) were recognized through joint feature-layout modeling. Because of the fixed convention used in the production syntax, major events in some sports domains usually start with the canonical views, detection of which can be used to find the event boundaries. Semantic labels of the detected events were further extracted by recognizing the score text box embedded in the image [7], resulting in the development of a video summarization system that automatically captures all the highlight points in the video such as scoring and last pitch for each player. The above two systems serve as excellent examples of the necessity of combining multi-modality information in detecting high-level semantic events in video.

Results of video event analysis can be utilized to produce different forms of adaptation. In live video applications such as sports broadcast, detected event information can be used to dynamically determine the optimal encoding and transmission formats of the video. [6] demonstrated a real-time adaptive streaming system in which non-important segments of the video were replaced with still visuals, text summaries, and/or audio only. Fig. 4 illustrates the concept of adaptive streaming. Such replacements facilitate great saving of the bandwidth or condensation of the total viewing duration. Important segments (those showing key events) in the video can be encoded with high quality or delivered as alerts depending on the user preferences. Because of the variable bit rate used in live video streaming, special transmission scheduling and buffer management methods are needed in order to handle the bursty traffic.

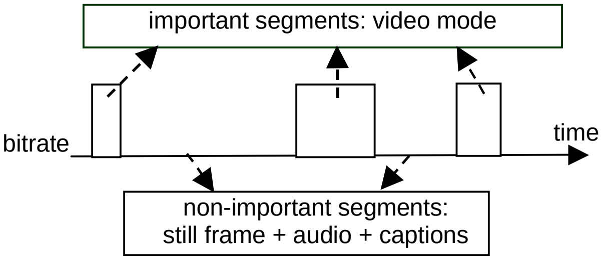

important segments: video mode
bitrate
time
non-important segments:
still frame + audio + captions

Fig. 4 Event-based adaptive streaming of videos over bandwidth limited links

The performance gains of the above event-adaptive streaming scheme depend on the actual video content, e.g., the percentage of important segments in the whole stream. In an experiment using baseball videos, we found non-important segments occupy more than 50% of duration. Such a significant ratio provides a large room for bandwidth reduction or time condensation. The speed and accuracy also depend on the complexity of events in each domain. For canonical views in sports, we realized a real-time software implementation with detection accuracy higher than 90% [6].

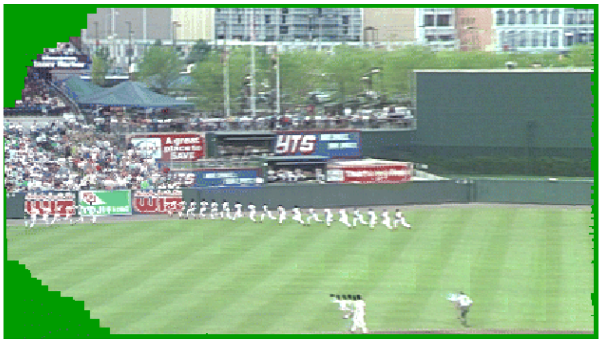


Fig. 5 Synopsis mosaic as visual summary of baseball events (from [8]).

B. *Structural-Level Adaptation*

Video is a linear medium capturing the real-world events and scenes that occur in space and time. The structures in video are caused by event occurrence orders, camera control patterns, and the final editing process. Exploration of relations of structural elements provides great potential for video adaptation. Such adaptations differ from those described in the previous subsection in the utility measure used – structural vs. semantic.

First, representative frames, or *key frames*, in each shot can be used to summarize the information in the shot. There has been a lot of work in key frame extraction based on detection of feature discontinuity, statistical characteristics, or syntactic rules. The adaptation process takes the original full-length video as input and produces a sequence of key frames, which can be sequentially played along with audio as a slide show, or organized in a hierarchical interface as navigation aids. In practical designs, there is tradeoff between the number of key frames and information completeness. In addition, ideal positions of key frames are usually difficult to determine – leaving the evaluation to some subjective criteria. The utility-optimization design procedure proposed in section II offers a systematic solution – given the constraints on the transmission bandwidth or the screen real estate in the user interfaces, determine the optimal set of key frames adaptively so that the largest amount of information utility can be achieved.

Another interesting technique for video adaptation at the structural level is mosaicing, which transforms image frame sequences captured by continuous camera takes (usually pan and zoom) into a panoramic view [8]. Background pixels captured in different frames are aligned and “stitched” together by estimating camera motions and pixel correspondence. The foreground moving objects are detected, and their moving trajectories are shown on top of the mosaiced background to highlight the long-term movement of the objects. An example of video mosaic for soccer video from [8] is shown in Fig. 5.

C. *Transcoding*

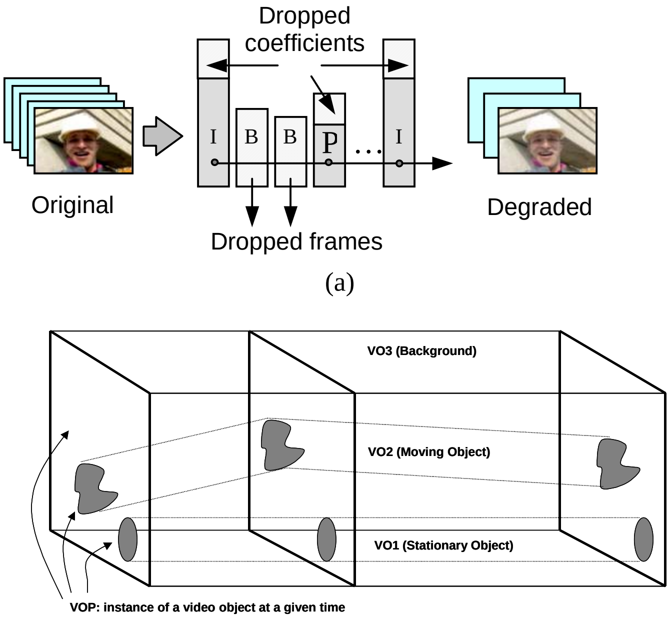
Below the semantic and structural levels comes the signal level adaptation, involving various manipulations of coded representations and issues of bit allocation. As mentioned in the adaptation taxonomy, the most straightforward way of transcoding is to decode video from a format to a new one, usually with change of bit rate as well. In applications that involve real-time transcoding of live videos for multiple users, design of the video transcoding system requires novel architectural- and algorithm-level solutions in order to reduce the hardware complexity and improve video quality (see a companion paper in this special issue on transcoding [9]).

In addition to format and basic bitrate transcoding, signal-level video adaptation may involve manipulation of video signals in the following dimensions:

- Spatial - change spatial resolution, i.e., frame size
- Precision - change the bit plane depth, color depth, or the step size for quantizing the transform coefficients
- Temporal - change the frame rate
- Object - transmit a subset of objects

Multiple dimensions of adaptation form a rich adaptation space as described earlier in section II. For example, Fig. 6(a) illustrates a system that combines frame dropping (FD) and coefficient dropping (CD), which can be implemented in most compression standards such as MPEG-2 and MPEG-4 [10]. Fig. 6(b) shows another example varying the frame rates for encoding different objects in a scene according to their importance [11]. Both methods can be used to meet tight bandwidth or storage constraints while optimizing the overall utility of the adapted video. If the spatio-temporal resolution of the video is unchanged, conventional quality measures such as PSNR can be measured at a fixed resolution. But, if the spatio-temporal resolutions are different, perceptual-level quality measures are needed. In [5], we conduct user studies to compare the subjective quality of videos transcoded at different spatio-temporal rates. We find distinctive patterns of users’ preferences of the temporal rate under different bandwidth conditions and content types.

Support of multi-dimensional video transcoding may be readily available if the source video is already encoded in a scalable format, i.e., a single encoded stream that can be truncated at different points to generate compatible substreams. The truncation results in different spatio-temporal rates and different bandwidth. Fixed-layer scalable coding usually consists of a small number of layers, targeting typical usage scenarios. On the other hand, continuous scalable coding provides much higher flexibility by allowing arbitrary truncation points. Interested readers are referred to a companion paper in this special issue on scalable video coding [12].



(a)

(b)

Fig. 6 (a) Video transcoding using combination of frame dropping and coefficient dropping (b) Video transcoding that varies the frame rate of different objects according to their importance (from [11]).

D. Rapid Fast-Forward – Drastic Temporal Condensation

Rapid fast-forward, sometimes referred to as video skimming [13][14], is a very useful adaptation tool when users' preferred viewing time is severely limited, while other resources may not be restricted. For example, users may want to rapidly complete viewing of a 10-min video within 1 minute. Such function resembles the typical fast-forward feature in the VCR player. However, here we are interested in a much higher time reduction ratio (e.g., 10x) compared to that of typical fast-forward (e.g., 2x to 3x).

Due to the drastic time condensation, simply increasing the frame rate of the player is not feasible, neither is the uniform subsampling of the frames in the original sequence. The former requires a playback throughput that's beyond the player's capability and will make the audio track unrecognizable. The latter will result in poor perceptual quality (e.g., important video frames skipped and audio content unrecognizable).

Instead of uniform frame subsampling, keyframes, as described in the previous section, can be extracted to form a much shorter image sequence. However, with such a frame-based subsampling scheme, we will lose the synchronization between video and the associated audio track.

An alternative approach to drastic video condensation is by intelligent dropping of a subset of continuous segments of video like shots or part of shots from the whole sequence. Simple heuristic rules like dropping from the end of each shot or random dropping of shots does not work because the perceptual quality will be severely undermined. In [14], a theoretical approach based on the utility-based conceptual framework discussed in section II was developed to find the optimal condensation scheme. First, video shots and syntactic structural units are identified as adaptation entities. Adaptation operations include length trimming or dropping of individual shots. The problem was formulated as constrained optimization, using the target viewing time as the main constraint. Other constraints were also used to take account of important production syntax used in films. For example, establishing shots at the beginning and syntactically critical structures such as dialogs cannot be changed. At emphasis points (e.g., key phrases or key audio-visual events), synchronization between audio and visual cannot be altered. In addition, psychophysical models based on subjective studies were used to estimate the relation between perceptual quality and audio-visual segment length. The subjective experiments confirmed the user preference of the optimized fast-forward schemes over the alternatives using fixed sub-sampling methods.

IV. Open Issues

Despite the many exciting advances discussed in the previous section, many open issues require future investigation in order for video adaptation to become a viable field. Some of the issues identified below are related to the analytical foundation, while others mainly address practical aspects.

A. Define Utility Measures and User Preferences

The most challenging part of quantitative analysis of video adaptation is to define adequate measures or methods for estimating utility. Conventional signal level measures like PSNR need to be modified when video quality is compared at different spatio-temporal resolutions. In [15], the signal-level distortion for videos coded at different spatio-temporal scales is computed at the full resolution, while some weighting factors are incorporated to account for the perceptual effects. Similarly, weights that account for motion masking effects are discussed in [16]. However, signal-level measures are often inadequate since the adaptation space involves many high-level operations such as shot removal, modality replacement, etc. Such operations cause complex changes to the content beyond the signal level and thus affect quality at other levels (such as perceptual, semantic, and comprehensiveness). Each level of quality may also involve multiple dimensions. For example, the perceptual level may involve spatial, temporal, or resolution dimensions.

Given the complex nature of utility, it will be difficult to define a universal measure for different levels or dimensions. In practice, input from user preferences can be used to set multiple constraints and optimization objectives. For example, a practical approach is to find an adaptation solution maximizing the comprehension-level utility while keeping the signal-level utility (e.g., SNR) above some threshold. However, asking users to unambiguously specify their preferences of some dimensions (e.g., temporal) over others (e.g., spatial) is impractical. In addition, user preferences often vary with content, task, and usage environment.

One possible alternative is to infer user preferences based on the usage history. Analysis of such data can be used to predict user preferences in similar contexts. Sharing of analysis results among different participating users may also be used to derive common criteria in collaborative filtering, provided that privacy concerns are adequately addressed.

Another direction is to correlate subjective preferences with content characteristics. In [10], we assume users have similar preferences of transcoding options (e.g., spatial vs. temporal scaling) for videos of similar characteristics. Based on this, automatic tools were developed to extract content features, cluster video data, and predict the utility values of each transcoding option, and thus automatically select the optimal transcoding option satisfying the resource constraints. The prediction accuracy was promising – about 90% of time the optimal transcoding option was correctly predicted.

Despite several potential approaches mentioned above, understanding what factors contribute to the overall video utility before and after adaptation and what components are computable/predictable still remains as a wide open issue. In [17], a relevant, broad concept called universal media experience is proposed to emphasize the need of optimizing the overall user experience instead of just enhancing the accessibility of content as in most existing UMA systems.

B. Resolve Ambiguity in Specifying Adaptation Operation

Due to the flexible formulation and implementation, some adaptation operations are not unambiguously defined. For example, an operation “remove the second half of each shot” appears to be clear. But in practice, the shot boundaries may not be exactly defined because of the use of automatic, imperfect shot detection tools. For another example, an operation “drop 10% of transform coefficients” does not specify the exact set of coefficients to be dropped. Different implementations may choose different sets and result in inconsistent resource and utility values.

There are several possible ways to address this problem. First, we can restrict adaptation operations to be based on unambiguous representation formats. For example, some scalable compression formats, such as JPEG-2000 and MPEG-4 fine-grained scalable schemes, provide unambiguously defined scalable layers. Subsets of the layers can be truncated in a consistent way as long as the codecs are compliant with standards.

The second approach is to allow for an ambiguity margin tolerating implementation variations, and estimate the bound of the variations in resource and utility. Theoretical estimate of such bounds is hard if not impossible. But assuming there exists some consistence among implementations, empirical bounds of such variations may be obtained. For example, it can be reasonably assumed that shot segmentation tools are relatively mature and bounds of shot boundary variations from different detection algorithms can be estimated through empirical simulations. Imposing further restrictions on implementations can tighten the bounds. For example, in the case of transform coefficient dropping, a uniform dropping policy can be used to restrict each block in a frame to drop the same percentage of coefficients.

Third, in some applications, the absolute values of resource and utility of each adapted entity are not important. Instead, the relative ranking of such values among different adaptation options are critical. In such cases, the chance of achieving ranking consistence is higher than consistence in individual values.

C. Relations Among Adaptation, Utility, and Resource

Relations among adaptation, resource, and utility are often complex, as described in section II. The complexity is especially high when the dimensionality of each space is high. Choices of the representation schemes for such complex relations will greatly affect flexibility and efficiency of the design of video adaptation.

One potential approach to tackling such complexity is to sample the adaptation space and store the corresponding resource and utility values as multi-variable high-dimensional tables. If a certain scanning scheme is adopted in the sampling process, elements of the tables can be represented by a one-dimensional sequence.

Another option is to decompose the adaptation space into low-dimensional spaces and sample each subspace separately. However, such schemes may lose the chance of exploring correlations among different dimensions.

Adequate representations vary with and depend on actual applications. For example, in a case that the adaptation space has a single dimension of varying quantization step size, the classical representation of rate-distortion curves is appropriate and has proven to be powerful. If the application only requires the information about ranking among adaptation operations satisfying certain resource (or utility) constraints, then sampling in the resource (or utility) space and representing the ranking among feasible adaptation options is an adequate solution.

D. Search Optimal Solutions in Large Spaces

Exploration of the above multi-space relations often leads to formulation of constrained optimization, some of which analytical solutions exist. For example, in most video coding systems, the rate-distortion (R-D) models have been used to represent resource-utility relations of video signals and achieve optimal coding performance. Such models are usually used for low-dimensional cases, e.g., quantization in the adaptation space, bit rate in the resource space, and SNR in the utility space. Joint optimization in multi-dimensional adaptation space including spatial, temporal, and SNR adaptation dimensions has been addressed in [10][15]. In the general cases, each space may have high dimensionality and the relations across spaces may be complex. It remains a challenging issue to find analytically optimal solutions or efficient search strategies under such complex conditions.

E. Design End-to-End Integrated Systems

Design of effective video adaptation solutions requires joint consideration of the adaptation subsystem with other subsystems such as content analysis, transmission, or usage environment monitoring. For example, many adaptation methods require recognition of structural elements or semantic events in the video. How do we design robust adaptation systems to accommodate the inconsistent, imperfect results from content analysis? Or sometimes it might be desirable to include users in the loop and use semi-automatic recognition methods in lieu of fully automatic ones. Adaptation solutions are often designed to satisfy various constraints or user preferences, which may be dynamically varying. What are the mechanisms and protocols for acquiring and monitoring such dynamic conditions? How should the adaptation process be designed in order to tolerate imprecise or imperfect information about usage environments?

In some applications that require live adaptation of embedded implementation, the computational resources are limited. We need to optimize resource allocation not only among components of adaptation but also between adaptation and other subsystems mentioned above. In such cases, the utility-resource framework described earlier offers an adequate conceptual basis that can be extended to address multi-subsystem resource allocation.

Another critical issue that affects the feasibility of video adaptation is related to the rights management. Many adaptation applications are hindered in practice today due to the restriction imposed by content owners on video content altering. Such restrictions may be placed through the use of proprietary formats or explicit legal limitations on manipulating the video content.

The first partial response to the above issues is to adopt modular designs of subsystems and provide well-defined abstraction of requirements and performance of each subsystem. For example, each content recognition subsystem can be abstracted in terms of the detection accuracy, the input content format, and the implementation complexity, etc. Similarly, each usage monitoring subsystem is abstracted based on the accuracy, the complexity, and the frequency of the measurement. With such modular abstraction, system-level integration and performance optimization can be made more tractable.

Another potential solution is to adopt international standards that define protocols and tools for describing important attributes required for designing an end-to-end video adaptation system. Such descriptions may address content adaptability, adaptation options, usage environment, and user preferences. In addition, standards are needed for describing information related to media rights management. In the next section, we will briefly review several international standards that are closely related to video adaptation.

V. Support Of Adaptation in International Standards

Recognizing the importance of media adaptation applications, several international bodies have recently developed standards to facilitate deployment and interoperability of adaptation technologies. Most notable ones include MPEG-7 [19][20], MPEG-21 [21][24], W3C [22], and TV-Anytime [23]. Different standards are targeted at different applications. For example, TV-Anytime focuses on adaptation of content consumption in high-volume digital storage in consumer platforms such as PVRs. W3C and IETF focus on facilitating server/proxy to make decisions on content adaptation and delivery. Its approach is based on a profile framework, called composite capabilities/preferences profile (CC/PP), and is mainly used to describe terminal capabilities and user preferences. In the following, we focus on a select set of tools provided by the MPEG-7 and MPEG-21 standards, and illustrate how these tools could be used together in a standardized adaptation framework that is consistent with the concepts put forward in this paper.

A. MPEG-7 Content Descriptions

MPEG-7 has standardized a comprehensive set of description tools, i.e., descriptors (Ds) and description schemes (DSs) to describe information about the content (such as program title and creation date) and information present in the audio-visual content (such as low-level features, mid-level features and structures, and high-level semantics). Such Ds and DSs are encoded using an extensible language based on XML and XML schema. In the area of video adaptation, MPEG-7 provides comprehensive support by specifying a wide variety of tools for describing the segmentation, transcoding hints, variations, and summaries of multimedia content. An excellent review of such components along with some application scenarios is presented in [18]. We include a brief summary of the tools and their use for adaptation here.

MPEG-7 provides tools for describing *user preferences* and *usage history*, which can be combined with description about content in personal content filtering/selecting applications. Specifically, the usage history DS consists of lists of actions performed by the user over some periods of time. A variety of actions (e.g., PlayStream, Record, etc) have been defined in an extensible dictionary. The UserPreferences DS describes user preferences related to different categories of attributes such as creation (creators, time periods, locations, etc), classification (genre, language, etc), dissemination (delivery type, source, and disseminator), media format, and format of navigation or summarization. Each preference attribute may be associated with a numerical weight, indicating the relative importance of each attribute compared to others.

MPEG-7 also provides *summary descriptions* that define the summary content, its relation to the original content, and the way the summary content is used to synthesize the final summary presented to the user. Summary content specifies the parts or components of the source content such as key segments or key frames of video or audio. The final synthesized form of summaries can be based on hierarchical organization of key components, sequential display of key components, or some customized presentations defined by practical applications.

The *variation description* is used to describe alternative versions derived from the original version. The type of the derivation process is specified by the variation relationship attribute. General types of processing may include revision by editing/post processing, substitution, or data compression. Transcoding types of processing involve reduction of bit rate, spatio-temporal resolution, spatial detail, color depth, or change of color format. Other processing types include summarization, abstraction, extraction, and modality conversion. Each variation is given a fidelity value and a priority value – the former indicates the quality of the alternative version of the content compared to the original version, while the latter the relative importance of the variation compared to other options.

In many applications of transcoding, low-delay, low-complexity, and quality preservation is required. To facilitate satisfaction of such requirements, MPEG-7 defines *transcoding hints* to provide metadata for guiding practical transcoding implementations. Such descriptions contain specifications of importance, priority, and value of segments, objects, and regions in audio-visual content, as well as descriptions of behaviors of transcoding methods. Some examples are motion hints (for guiding motion-based transcoding methods), difficulty hints (for bit rate control), semantic importance hints (for guiding rate control), spatial resolution hint (for specifying the maximum allowable spatial resolution reduction), etc. Transcoding hints descriptions are associated with compressed videos and can be stored in the server or transmitted to proxies where the transcoding operations take place.

B. MPEG-21 Digital Item Adaptation

An extended scope of issues related to adaptation of digital multimedia content is addressed by Part 7 of the MPEG-21 standard, Digital Item Adaptation (DIA) [24]. In the following, specific tools related to the adaptation conceptual framework presented in section II are briefly outlined and discussed.

Given that adaptation always aims to satisfy a set of constraints, tools that describe the usage environment in a standardized way are essential. As a result, the DIA standard specifies tools that could be used to describe a wide array of user characteristics, terminal capabilities, network characteristics and natural environment characteristics. As a whole, this set of usage environment descriptions (UED's) comprise the resource space discussed in section II.

User characteristics include several tools imported from MPEG-7 (e.g., user preference), as well as a number of newly developed tools. Among the new tools are presentation preferences, which describe preferences related to audio-visual rendering, or to the format/modality a user prefers to receive, accessibility characteristics, which enable one to adapt content according to certain auditory or visual impairments of a user, and location characteristics, which describe the mobility and destination of a user. Terminal capabilities include encoding and decoding capabilities, display and audio output capabilities, as well as power, storage and input-output characteristics of a device. Network characteristics include static capabilities of a network such as its maximum capacity, as well as dynamic conditions of a network such as the available bandwidth, error and delay. The natural environment pertains to physical environmental conditions such as the lighting condition or auditory noise level, or a circumstance such as the time and location that content is consumed or processed.

While the usage environment description tools may be used in a standalone manner to convey implicit constraints to a server or proxy, they may also be used to provide a richer form of expression through the Universal Constraints Description (UCD) tool. With the UCD tool it is possible to formulate explicit limitation and optimization constraints. In this way, additional guidance is provided to an adaptation engine in a standardized way so that a more satisfactory adaptation could be provided and/or to limit the space of feasible adaptations so that the required effort to search for an optimal solution is reduced. As an example, consider an input image to be adapted according to the following: maximize the adapted image quality, such that (i) the output rate is less than the average available network bandwidth, (ii) the adapted width is greater than 50% of the display width, and (iii) the aspect ratio of the adapted image is equal to that of the input image.

It should be noted that such expressions may be provided not only by the user, but the content provider as well to enforce some level of control as to how their content is adapted and the form it is ultimately delivered. As part of ongoing work in DIA, the link to such constraints with adaptation rights and other digital rights management tools is being explored.

Also worth noting in the above example is that descriptions of both the usage environment such as network bandwidth, and descriptions of media characteristics such as output rate of the source, are required to describe both ends of the system. This reinforces the inherit dependency between MPEG-7 and MPEG-21 towards solving UMA related problems.

To complete the adaptation framework, the DIA standard also specified a means to describe the relationship between the above constraints, the feasible adaptation operations satisfying these constraints and associated utilities that result from adaptation. The tool enabling this is referred to as the AdaptationQoS tool. The relations that this tool describes could be specified at various levels of granularity (e.g., frame, group-of-pictures), which is consistent with the concept of an entity and the adaptation-resource-utility relations introduced in section II. With this information, the adaptation problem becomes a well-defined mathematical problem to which the optimal adaptation strategies described earlier could be applied.

C. Standardized Adaptation Framework

Fig. 7 illustrates how the above concepts fits together into a standardized form of the conceptual adaptation framework presented in this paper. Several inputs are provided to an adaptation decision engine, including media characteristics as given by MPEG-7, along with the constraints and relations as given by the UED/UCD and AdaptationQoS tools of MPEG-21. It is the function of the adaptation decision engine to use this input to find an optimal set of adaptation parameters that satisfy all the given constraints. These parameters are then passed to a bitstream adaptation engine, where the actual adaptation of the input bitstream occurs.

From the above, it is clear that both MPEG-7 and MPEG-21 are well aligned with the conceptual adaptation framework presented in this paper and could provide solutions that address some of the end-to-end design concerns raised in section IV.E.

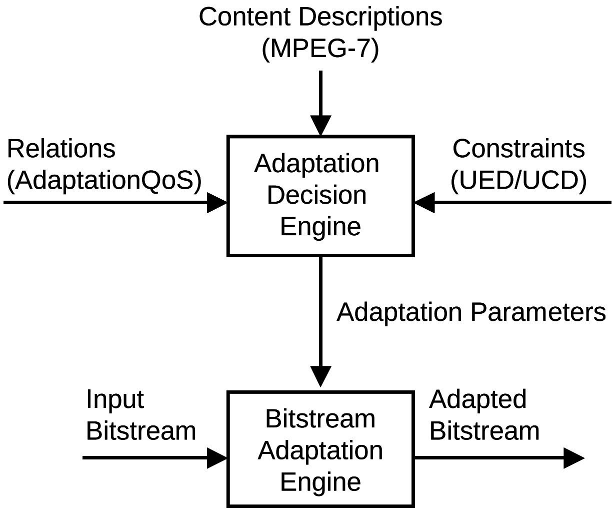


Fig. 7 Diagram illustrating adaptation framework according to MPEG-7/21 description tools.

VI. Conclusions

Video adaptation is an emerging field that encompasses a wide variety of useful technologies and tools for responding to the need of transmitting and consuming multimedia content in diverse types of usage environments and contexts. Different from video compression or transcoding, adaptation offers a broader spectrum of operations at multiple levels ranging from signal, perceptual, to semantic. Recently, exploration of various adaptation techniques has facilitated development of many exciting applications such as event-adaptive streaming, personalized media variation and summarization, and multi-level multi-dimensional transcoding. Several international standards such as MPEG-7 and MPEG-21 also include tools to describe various information about the content, user and usage environment, which is necessary for video adaptation.

Despite the bourgeoning activities and advances, this field is in need of an analytical foundation and solutions to many challenging open issues. This paper offers a preliminary framework that characterizes fundamental entities and important concepts related to video adaptation. Introduction of such a framework allows for systematic formulation of many practical problems as resource-utility tradeoff optimization.

Critical open issues that call for further investigation include development of effective measures and estimation methods for utility (i.e., video quality in a general sense), adequate representation of relationships among concepts (i.e., adaptation, resource and utility), efficient search methods of optimal solutions satisfying diverse constraints, and finally systematic methodologies for designing complex end-to-end adaptation systems. The first issue related to utility measurement is of the foremost importance for the theoretical development in the field. In view of the difficulty in establishing universal computable metrics for utility, potential solutions may be derived by exploring the description, analysis, and prediction of user preferences of different adaptation options in each practical application setting.

It is worthwhile to note that solutions to most of the above identified open issues require joint consideration of adaptation with several other closely related issues, such as analysis of video content, understanding and modeling of users and environments, and rights management of digital content. Such cross-disciplinary exploration is critical to innovation and advancement of video adaptation technologies for next-generation pervasive media applications.

VII. Acknowledgements

The authors are grateful for anonymous reviewers for their valuable input, and for Yong Wang and Anita Huang for their editorial help.

References

[1] A. Fox and E. A. Brewer, "Reducing WWW Latency and Bandwidth Requirements by Real timer Distillation", *Proc. Intl. WWW Conf.*, Paris, France, May 1996.

[2] J. R. Smith, R. Mohan and C. Li, "Scalable Multimedia Delivery for Pervasive Computing", *Proc. ACM Multimedia*, Orlando, FL, Oct-Nov 1999.

[3] S.-F. Chang, “Optimal Video Adaptation and Skimming Using a Utility-Based Framework,” *Proc. Tyrrhenian Intl Workshop on Digital Communications*, Capri Island, Italy, Sept. 2002.

[4] B. Li, J. Errico, H. Pan, and I. Sezan, “Bridging The Semantic Gap in Sports Video Retrieval and Summarization,” *Journal of Visual Communication and Image Representation, Special Issue on Multimedia Database Management*, 2004.

[5] Y. Wang, S.-F. Chang, A. C. Loui, “Subjective Preference of Spatio-Temporal Rate in Video Adaptation Using Multi-Dimensional Scalable Coding,” *Proc. IEEE Int'l Conf. Multimedia and Expo*, Taipei, Taiwan, June 2004.

[6] S.-F. Chang, D. Zhong, and R. Kumar, “Real-Time Content-Based Adaptive Streaming of Sports Video,” *Proc. IEEE Workshop on Content-Based Access to Video/Image Library, IEEE CVPR conference*, Hawaii, Dec. 2001.

[7] D. Zhang, S.-F. Chang, Event Detection in Baseball Video Using Superimposed Caption Recognition, *Proc. ACM Multimedia*, Jean Les Pins, France, Dec. 2002.

[8] M. Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Trans. on PAMI*, 86(5):905--921, May 1998.

[9] Transcoding paper in this special issue.

[10] Y. Wang, J.-G. Kim, and S.-F. Chang, "Content-based utility function prediction for real-time MPEG-4 transcoding," *Proc. IEEE Int'l Conf. Image Processing*, Barcelona, Spain, Sept 2003.

[11] A. Vetro, T. Haga, K. Sumi and H. Sun, "Object-Based Coding for Long-Term Archive of Surveillance Video", *Proc. IEEE Int'l Conf. Multimedia and Expo*, Baltimore, MD, July 2003.

[12] Scalable coding paper in this special issue.

[13] M. A. Smith and T. Kanade, "Video Skimming for Quick Browsing Based on Audio and Image Characterization," *Carnegie Mellon University, Technical Report CMU-CS-95-186*, July 1995.

[14] H. Sundaram, L. Xie, S.-F. Chang, "A Utility Framework for the Automatic Generation of Audio-Visual Skims," *Proc. ACM Multimedia*, Juan Les Pins, France, December 2002 .

[15] E.C. Reed and J.S. Lim, "Optimal multidimensional bit-rate control for video communications," *IEEE Trans Image Processing*, vol. 11, no. 8, pp. 873-885, Aug. 2002.

[16] J. Lee and B. W. Dickinson, "Temporally adaptive motion interpolation exploiting temporal masking in visual perception," *IEEE Transactions on Image Processing*, vol. 3, pp. 513–526, September 1994

[17] F. Pereira and I. Burnett, "Universal Multimedia Experiences for Tomorrow," *IEEE Signal Processing Magazine*, vol. 20, pp. 63-73, March 2003.

[18] P. van Beek, J. R. Smith, T. Ebrahimi, T. Suzuki, J. Askelof, "Metadata-Driven Multimedia Access," *IEEE Signal Processing Magazine*, pp. 40-52, March 2003.

[19] Public MPEG documents, http://www.chiariglione.org/mpeg/working_documents.htm

[20] MPEG-7 Overview v.9, ISO/IEC JTC1/SC29/WG11N5525, March 2003.

[21] MPEG-21 Overview v.5, ISO/IEC JTC1/SC29/WG11/N5231, Oct. 2002.

[22] W3C CCPP CC/PP, "Exchange protocol based on HTTP extension framework" [Online]. Available: http://www.w3.org/TR/NOTE-CCPPexchange

[23] TV-Anytime Forum. Available: http://www.tv-anytime.org

[24] Information Technology – Multimedia Framework (MPEG-21) – Part 7: Digital Item Adaptation, ISO/IEC 21000-7 FDIS, ISO/IEC JTC 1/SC29/WG11/N6168, December 2003.

Shih-Fu Chang (M'93–SM'02–F'04) is a Professor in the Department of Electrical Engineering of Columbia University. He directs the Digital Video/Multimedia Lab (http://www.ee.columbia.edu/dvmm) and ADVENT industry-university consortium at Columbia University, conducting research in video analysis, multimedia indexing, universal media access, and media authentication. Systems developed by his group have been widely used, including VisualSEEk, VideoQ, WebSEEk for image/video search, WebClip for networked video editing, and Sari for online image authentication. He has led major cross-disciplinary projects, such as a medical video digital library funded by NSF DLI initiative, an art image education project with Teachers College, and a large video search engine with stock footage companies in NYC. Through collaboration with industry partners, his group has made major contributions to the development of MPEG-7 multimedia description standard.

Prof. Chang served as a general co-chair of ACM 8th Multimedia Conference 2000, and will participate as a conference Co-Chair in IEEE Multimedia Conference 2004. He has also consulted at several media technology companies. He has been a Distinguished Lecturer of the IEEE Circuits and Systems Society, 2001-2002; a recipient of a Navy ONR Young Investigator Award, IBM Faculty Development Award, NSF CAREER Award, and three best paper awards from IEEE, ACM, and SPIE in the areas of multimedia indexing and manipulation. In recent years, his students have also received recognition with several best student paper awards in peer reviewed publications.

Anthony Vetro (S'92-M'96-SM'04) received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from Polytechnic University, Brooklyn, NY.

He joined Mitsubishi Electric Research Labs, Cambridge, MA, in 1996, and is currently a Team Leader and Senior Principal Member of the Technical Staff. His current research interests are related to the encoding and transport of multimedia content, with emphasis on video transcoding, rate-distortion modeling and optimal bit allocation. He has published more than 80 papers in these areas and holds 18 U.S. patents. Since 1997, he has been an active participant in MPEG, contributing to the development of the MPEG-4 and MPEG-7 standards. Most recently, he served as editor for Part 7 of MPEG-21, Digital Item Adaptation.

Dr. Vetro has been a member of the Technical Program Committee for the International Conference on Consumer Electronics since 1998 and has served the conference in various capacities. He has been a member of the Publications Committee of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS since 2002 and elected to the AdCom of the IEEE Consumer Electronics Society from 2001-2003. He is a member of the Technical Committee on Visual Signal Processing and Communications of the IEEE Circuits and Systems Society, and was a member of the Editorial Board for the *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* from 2001-2004. He served as Guest Editor (with C. Christopoulos and T. Ebrahami) for the special issue on Universal Multimedia Access of *IEEE Signal Processing Magazine*. He has also received several awards for his work on transcoding, including the 2003 IEEE Circuits and Systems CSVT Transactions Best Paper Award.