Optimal Video Adaptation and Skimming Using a Utility-Based Framework

Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University
New York, NY 10027, USA
Ph.: + 1 212-749-5998 email: sfchang@ee.columbia.edu,
Web: http://www.ee.columia.edu/dvmm

ABSTRACT

Video adaptation uses various spatio-temporal adaptation techniques to transform video content in order to satisfy diverse resource constraints and user preferences. To explore systematic solutions, we present a new conceptual framework to model content, adaptation processes, utility, resources, and relations among them. The key to the framework is definition of the notion of utility and the formulation of the problem as one of constrained utility optimization. In this paper, we use the framework to form a unified view towards problems and solutions we have developed in several media adaptation projects, ranging from video skimming, transcoding, to adaptive streaming. We discuss several important open issues, including ambiguity in adaptation specification, representation of resource and utility distributions in high-dimensional spaces, modeling and classification of resource/utility, and efficient solution search strategies under multiple complex constraints.

1 INTRODUCTION

In this paper, we tackle the problem of optimal adaptation of multimedia in order to satisfy user preferences in consuming multimedia content on heterogeneous platforms with resource constraints. Resources include diverse factors, including bandwidth, display capability, CPU speed, and even user's available time. Such a problem is important in pervasive media applications in which media content may have to be transformed before being rendered on user device. It's also useful for interactive media access, in which users may wish to see a condensed version before retrieving the complete stream. In the paper, we focus on video, although the framework and methodology can be easily extended to other modalities of media.

Several works have studied media content adaptation for universal Web access. In [1], Internet content is modified by a real-time distiller at the proxy to meet resource constraints at various devices and improve end-to-end performance. In [2], an InfoPyramid framework is used to manage Internet multimedia content, model relations among various versions and modalities, and choose the allowable transcoding paths. However, the framework focuses on the content variation space only. The required resources and quality scores are modeled as simple signatures of the content before and after transcoding.

In this paper, we focus on adaptation of digital video at multiple levels within a single stream or across multiple streams. We present a new framework for defining video entity, the adaptation space, the utility space, the resource space, and explicitly modeling the relationships among them. We call the framework, *utility-base*d, and use it to formally formulate several optimal video adaptation problems as those of *constrained utility maximization*. By using several case studies, we show that such a multi-space framework is useful for formulating optimal video adaptation problems, particularly for those that need to consider multiple types of resources and multiple dimensions of quality.

We describe the utility-based framework in the next section. In Section 3, we present several case studies illustrating the realization of the utility framework in video skim generation, MPEG-4 video transcoding, and video streaming. Section 4 includes discussion of important issues in representation and modeling of resource/utility distributions. Finally, conclusions are given in Section 5.

2 THE UTILITY-BASED FRAMEWORK

In this section, we formally define the utility-based framework for modeling the relationships among content entity, adaptation, resource, and utility.

2.1 ENTITIES

We define entity to be a chunk of video data that shares a certain consistent property. For example, a shot is a sequence of image frames that are captured by a continuous camera take and share consistent color and motion. Entities exist at many levels — objects, frames, shots, syntactical elements, as well as semantic elements. This definition is easily extended in a hierarchical fashion (e.g. a dialog sequence is an entity, whose constituents are shots that share a topological property — adjacent shots differ, while every second shot is alike.). Such definition is similar to the one that we introduced in [3].

Complex entities can be defined by more sophisticated properties. For example, syntactic entities like recurrent anchor shots in news, pitching shots in baseball, and structured dialog sequences in films can be defined by syntactic relations among elements in the data. Semantic entities like scoring events in sports and news stories are

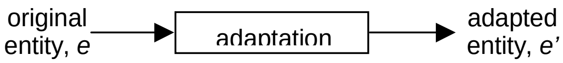


Figure 1. The entity adaptation process

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 romantic, rhythmic) shared by elements of the content. Finally, the audio-visual synchronous entities require tight synchronization between audio and video, such as close-up views of talking face, thunderstorms, explosions, etc.

An entity is the data unit undergoing the adaptation operation, as shown in Figure 1. The simple adaptation process can be extended in a hierarchical way, with constituent entities undergoing individual adaptation operations (e.g., reduce the size of individual frames) and the overall composite entity undergoing a different adaptation operation (e.g., condense the duration of the whole video shot).

2.2 THE ADAPTATION SPACE

Audio-visual content can be adapted in various ways. The following lists a few possible categories:

Temporal:

- *Frame rate reduction*: Some frames of video are removed from transmission and playback. For example, all even frames are dropped to cut the frame rate to half; all B frames in a MPEG video sequence are dropped; or all frames except the key frames in a shot are removed. The first two adaptation processes drop frames in a uniform way, while the last one non-uniformly. Note here we assume the final playback duration is not changed. Therefore, the intervals corresponding to the dropped frames have to filled by duplicated or interpolated frames. The average bandwidth required for transmission is reduced.

- *Time condensation*: Assuming the playback frame rate is kept the same (e.g., 30 frames per second), image frames are removed to shorten the playback duration of the video. Dropped frames needs not to be uniformly distributed. In general, we call this adaptation operation, *video skimming*, in which any chunk or subset of frames can be removed in order to shorten the total viewing time. For example, a chunk of frames can be removed from the tail of a shot; or a subset of shots can be removed from a sequence based on semantic criteria or user preferences.

Spatial:

- *Resolution*: The resolution of the image is reduced so that the image is rendered with a smaller size, along with a smaller bandwidth requirement.

- *Signal-Noise-Ratio (SNR*): SNR is usually used to measure the objective quality of a signal. SNR adaptation operations are those that remove frequency-domain coefficients or less important layers to trade signal quality for lowered bandwidth. It is typically achieved by dropping transform coefficients in compressed streams. Note the resolution of the final displayed video is not changed.

- *Object:* In object-based video representations, such as MPEG-4 and JPEG-2000, low-priority video objects can be dropped to reduce the overall bandwidth requirement.

2.3 THE RESOURCE SPACE

Resources are available supports from *communication, computing, storage*, *display,* as well as *viewer* for receiving, rendering, and presenting the entity. The following includes some examples.

- Bandwidth
- Computation capability
- Memory
- Power
- Display (resolution, color depth, dynamic range *etc*)
- Viewer's available time

One may consider that the *viewer's time* is better categorized as a user preference or task requirement. But here we take a general view to categorize it as resource requirement on the viewer part.

Some of the required resources may depend on actual implementations. For example, the required computational cost for rendering a video depends on the decoder implementation and the hardware/software architecture.

2.4 THE UTILITY-BASED FRAMEWORK

2.4.1 Utility

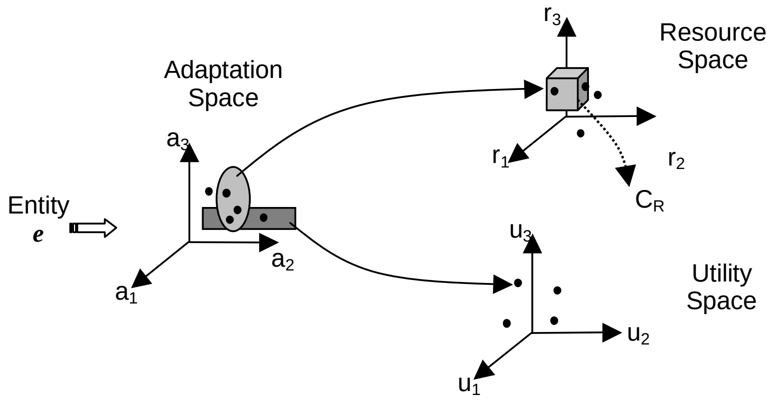
Utility is the quality of an entity when it's rendered on a user device through a delivery channel. When a content entity is adapted, its utility value is usually changed accordingly.

Utility can be measured in an objective manner (such as SNR) or a subjective one (such as subjective quality index). Subjective measures of videos are often sensitive to types and conditions of display devices, delivery channels, and viewers. Being free of such dependence, objective measures are often used.

Utility may include attributes in multiple dimensions. In addition to SNR, other dimensions include temporal smoothness, audio-visual rhythm, comprehensibility (see Section 3.2), coherence, etc. The last three have been included in the utility model used in our prior work of optimal audio-visual skim generation [3].

2.4.2 Relations among Adaptation, Resource, and Utility

Figure 2 shows the relationships among the adaptation space, the resource space, and the utility space. Given a content entity, *e*, the adaptation space represents the conceptual space of all possible adaptation operations.



Problem Formulation:
*Given entity **e** and resource constraints C_R → find optimal adaptation operation $\boldsymbol{a_{opt}}$ so that utility of the adapted entity **e'** is maximized.*

Figure 2. The utility framework for content adaptation and problem formulation as constrained utility maximization. Note the formulation can be easily extended to a utility-constrained version.

Each dimension of the adaptation space represents one type of adaptation, such as frame dropping, time condensation, or transform coefficient dropping. Each dimension has a certain cardinal index but there are no meaningful distance metrics defined in the adaptation space. For example, the uniform frame dropping dimension can be indexed based on the percentage of frames dropped. The SNR adaptation dimension can be indexed by the percentage of transform coefficients dropped or the size of the quantization step. Each point in the adaptation space represents a combination of adaptations from constituent dimensions. The origin of the space represents the null operation, namely, the original entity is kept intact.

We use lower-case letters to denote the dimensions in each space. Bold-face lower-case letters denote points in the space. Upper-case letters denote sets of points or regions in a space.

If an entity consists of multi-level sub-entities, each node in the hierarchy is associated with an individual adaptation space defining the operations applicable to the node. Often, adaptations of sub-entities may affect the utility and permissible adaptations of sibling sub-entities under the same parent node. For example, removing the odd shots in a video dialogue will make the remaining shots meaningless. Removing I frames in a MPEG sequence will seriously affect the utility of subsequent B and P frames.

Each point in the adaptation space represents a function that maps the input entity to an adapted entity. Given a delivery link and a rendering device, each adapted entity has a corresponding point in the resource space and a corresponding point in the utility space. Such mapping relations are shown by the directed arcs in Figure 2. The shaded cube in the resource space represents the resource constraints specified by applications. Note there may exist multiple adaptation points that satisfy the same resource requirement. The oval shaped region in the adaptation space shows such a *constant-resource set*. Also, different adaptation operations may result in the same utility value. The rectangle-shaped region in the adaptation space represents such a *constant-utility set*.

Relations shown in Figure 2 can be simplified to classical forms in certain fields. For example, if the adaptation space is a one-dimensional space specifying only the step size for transform coefficient quantization, the resource space includes only the bit rate, and the utility space includes only the SNR measure of the compressed video, the relationship between bit rate and utility can be reduced to the rate-distortion (R-D) curve that is often used for video coding optimization.

2.5 RESOURCE-CONSTRAINED UTILITY MAXIMIZATION FORMULATION

Using the utility framework, many problems in video skimming and adaptation can be formulated as the following:

*Given a content entity **e** (either elementary or composite), and resource constraints (C_R), find the optimal adaptation operation, $\boldsymbol{a_{opt}}$, within the permissible adaptation region so that the utility of the adapted entity **e'** is maximized.*

Note the above formulation can be easily extended to one that imposes constraints in the utility space and aims at overall resource minimization or system efficiency improvement. The case study described in Section 3.2 includes more discussions about this.

Constraints in resources vary in different applications. For example, the device resolution and color depth may be limited on handheld devices. The available user viewing time may be limited in a pervasive media environment or an interactive video retrieval system. The network bandwidth may be constrained in a dynamic mobile environment.

The search region in the adaptation space may be limited by practical implementations. For example, an adaptation engine that only drops B frames in MPEG sequences cannot drop an arbitrary number of frames within a video segment. The number of droppable frames (i.e., B frames) depends on the specific coding parameters used in the source encoder.

3 CASE STUDIES

In this section, we present a few cases that can be modeled as realizations of the above framework and the constrained utility optimization approach. Modeling these works after the utility framework provides a unified view towards solving problems in video adaptation and skimming.

3.1 AUDIO-VIDEO SKIM GENERATION

In the audio-video skimming project [3], we applied the utility-based approach to automatic generation of video skim generation. A *skim* is a short video clip obtained by condensing the original sequence to satisfy the user's information needs (or tasks) as well as the capabilities of the rendering device. Skims come with different types of flavors- semantic, affect, event, and information centric. Here we focus on the information centric skims that can be robustly computed.

We focus on skims of computable scenes. A computable scene is a sequence of shots that include long-term consistence in chromaticity, lighting, and ambient sound. Such a computable scene typically corresponds to a physical location or a semantic event (e.g., a journey of the same characters). The research objective is to solve the following problem: given an input video scene in films and certain resource constraints (user's affordable viewing time in this case), how do we generate a skim with maximal comprehension utility?

Entities, Utility, and Constraints

We detect entities in both video and audio. We detect the basic video shot entity and then detect the syntactical entities such as dialogues. We define comprehension utility for such entities based on visual complexity and duration. For audio, we detect the basic audio entities (speech, non-speech segments, and significant phrases) using SVM classifiers.

In addition, we construct audio-visual synchronous entities via tied multimedia segments. A multimedia segment is said to be fully *tied* if the corresponding audio and video segments begin and end synchronously, and in addition cannot be condensed. Examples of tied multimedia segments are those speech segments that contain significant phrases.

Syntax plays an important role in assigning utility and developing search strategy. To ensure the coherence of the skims, maximal syntactic entities have to be preserved. For example, to preserve the perception of video dialogues, a minimal number of shots in the dialogue have to be retained. To preserve the establishment and ending of a scene, a few shots (e.g., 3) at the start and the end of the scene must be retained. Tied multimedia segments should not be altered in order to retain the end-to-end audio-visual synchrony. Such syntactic properties are usually imposed by the production rules in specific domains. To preserve maximal syntax, we can either assign large utility values to syntactic elements or adopt a biased strategy in searching the adaptation space.

To model the utility of entities, we measure the comprehensibility of a video shot and an audio segment as a function of its duration and complexity. The non-negative utility function of a video shot $S(t, c)$, where t is the duration of the shot and c is its complexity, is modeled to be a bounded, differentiable, separable, concave function:

$$S(t,c) = \beta c(1-c)\cdot(1-\exp(-\alpha t)). \qquad <1>$$

The utility function for the sequence of shots is the sum of the utilities of the individual shots:

$$U_v(\vec{t}_v, \vec{c}, \phi_v) = \frac{1}{N_{\phi,v}}\left(\sum_{i:\phi_v(i)=1} S(t_{i,v}, c_i) - \sum_{j:\phi_v(j)=0} P(t_{p,j}) \right) \qquad <2>$$

where $\vec{t}_v : t_0, t_1 ... t_N$ and $\vec{c} : c_0, c_1 ... c_{N.}$ represent the durations and complexities of the shot sequence and where P($t_{p,j}$) represents a negative shot dropping utility. The utility function for an audio segment is defined in a similar fashion. Further details can be found in [3].

In addition, we incorporate a utility component to model the "film rhythm." This component computes the utility penalty associated with deviation from the original affect entity. The penalty is increased if the duration ratios among shots deviate from the original ratios.

We do not explicitly model the utility of the elements of syntax or the entities due to synchronous elements. The utilities are *implicitly* maximized by the syntax-preserving search strategy described below.

The Search Strategy

We focus on the generation of passive information centric summaries that have maximum coherence. Since we deem the speech segments to contain the maximum information, we achieve this goal by biasing the audio utility functions in favor of the clean speech class.

In order to ensure that the skim appears coherent, we do two things: (a) ensure that the principles of visual syntax are not violated and (b) have maximal number of synchronous entities. These entities ensure synchrony between the audio and the video segments. The target skim duration is met by successively relaxing the synchronous constraints. Relaxing the synchronization constraints has two effects: (a) the corresponding audio and video segments are no longer synchronized (b) they can be compressed and if necessary dropped.

The information-centric skims are generated using the resource-constrained utility maximization framework described above. We conduct experiments using multiple genres of films and human subjects. Results indicate that the optimal skims are better than those generated using heuristics in a statistically significant sense at high compression rates (80% ~ 90%).

3.2 UTILITY-BASED MPEG-4 TRANSCODING

In the universal media access project [4], we address the issue of optimal video transcoding to satisfy the dynamic bandwidth resource constraint. We define the video entity as a MPEG-4 compressed video sequence. Resource is defined to be the network bandwidth available for transmitting the video over the network link. We focus on mobile applications in which network bandwidth may be dynamically changing.

The adaptation space includes frame dropping and transform coefficient dropping. It can be easily extended to include other adaptation dimensions. To minimize the computational cost, we consider dropping B or P frames in the compressed stream without decoding the video to the uncompressed source format. Coefficient dropping operation is indexed by the percentage of rate reduction by dropping coefficients over the entire frame. As discussed in Section 2.6, to avoid definition ambiguity, we adopt some restrictions and allow only dropping the same percentage across different frames. However, within a single frame, we allow the flexibility in choosing either an optimal non-uniform dropping scheme or a simple uniform dropping scheme among blocks

The objective is as follows- given the available bandwidth and a MPEG-4 video entity, find the optimal adaptation operations satisfying the target bandwidth and achieving the maximal SNR utility. The optimal adaptation specifies the optimal combination of temporal adaptation (i.e., number of B and P frames dropped in each group of pictures, GOP) and spatial adaptation (i.e., coefficient dropping percentage).

In [4], we proposed a utility-based descriptor to describe the optimal adaptation operation for each given bandwidth over a range. Such descriptor information is useful in a three-tier proxy-based network architecture, in which the transcoding proxy can determine the best adaptation operation without needing to exhaustively comparing all possible adaptation operations. The bandwidth range is discretized to a finite set of points. Associated with each bandwidth point, the descriptor includes information about the permissible adaptation points capable of achieving the target bandwidth, the ranking of such permissible adaptation points based on their SNR utility values, and a flag to indicate the consistence of the ranking information over different implementations of adaptation engines. The consistence flag is set at the server based on empirical simulations or prediction by some models. In the case that there are no variations of utility values among different implementations, the absolute values of the utility are also included in the descriptor.

Figure 3 shows experiment results of transcoding a MPEG-4 test sequence "coastguard" over a bandwidth range up to 1.5 Mbps. It's important to see that for a given target bandwidth, there are multiple adaptation operations satisfying the same target bandwidth. The optimal operation with the highest video utility is selected.

The above utility-based descriptor can be easily extended to spaces with higher dimensions. For example, if the resource space includes both bandwidth and computational cost, the two-dimensional resource space can be sampled to discrete points, each of which is associated with all possible adaptation operations satisfying the target resource values. Ranking of these permissible adaptation points can be assigned based on utilities of their adaptation results. As mentioned in Section 2.4.1, the utility space can be extended to multiple dimensions as well.

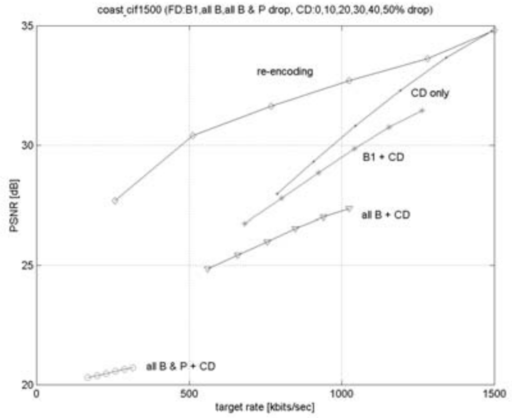


Figure 3. Rate-Distortion relations for different adaptation operations. CD: coefficient dropping (with optimal allocation within a frame), B1: drop one B frame within a GOP, B &P: drop all B and P frames within a GOP.

Note the utility-based framework provides general information about relationships between resource and utility. Such information can be utilized in a flexible way. The above utility-based descriptor addresses resource-constrained problems and is intended primarily for the wireless video transcoding application. A different application scenario involves the utility-constrained issue: given the total network bandwidth and a minimal acceptable utility for each video, what's the maximal number of videos that can be simultaneously served in the same network? The scenario can be further extended to add sophisticated utility-pricing models to seek for maximum overall system revenue.

3.3 MPEG-4 FGS JOINT SPATIO-TEMPORAL QUALITY OPTIMIZATION

In [5], we used subjective measures to compare the utilities of videos compressed by a MPEG-4 fine-grained-scalable (FGS) coding scheme. FGS allows for scalable coding in the temporal dimension (different frame rate) as well as the spatial dimension (layers with different numbers of bit planes). Using a scalable coder derived from MPEG-4 FGS reference software, we conducted subjective tests to compare the quality of multiple videos coded at the same bit rate but with different frame rates. Users were shown multiple videos on the screen and asked to indicate their preferences in the optimal frame rate. Figure 4 shows the average optimal frame rate for different videos over a bandwidth range.

The above preference curves are similar to the ranking information of permissible adaptation operations described in Section 3.2, except here a subjective utility measure is used. In [5], we take advantage of such information and determine the optimal bit rates for the motion prediction reference layer. We called this scheme, *MPEG-4 FGS+*. At any given bit rate, we know the preferred frame rate based on the subjective model. The model indicates that neither the option of a lower frame rate with higher spatial quality in individual frames nor the one of a higher frame rate with lower spatial quality gives viewers higher subjective satisfaction. Conversely, from the same model, we know

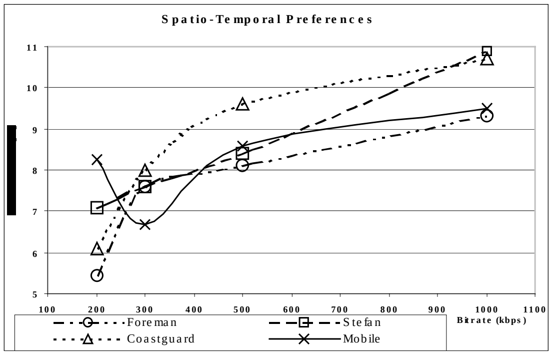


Figure 4. Subjective preference in the optimal frame rate over a bandwidth range. Different curves are results for different videos.

the minimal available bit rate for each frame at any given frame rate. Therefore, we can set the motion prediction reference layer to such a bit rate when we further increase the frame rate.

The original FGS coding in MPEG-4 uses a preset constant rate for the motion prediction reference layer. Our approach described above adapts the reference layer bit rate for different B and P frames based on the subjective utility model. Our experiments show up to 1.5 dB improvement over the MPEG-4 FGS coding scheme.

3.4 CONTENT-ADAPTIVE VIDEO STREAMING

In [6], we adaptively change the display modality of a video segment according to its semantic utility class. For semantically important classes, we assign full-motion video and audio. For non-important segments, we just display static key frames plus audio or textual captions. An adaptive streaming prototype system for sports video was developed in [6]. Important segments correspond to pitching and follow-up events after pitching in baseball, or serving and active play segments in tennis. Figure 5 shows such an adaptive streaming and playback scheme.

The above utility-adaptive streaming scheme optimizes the subjective utility while satisfying the bandwidth resource constraint. It’s intended for bandwidth limited applications such as wireless video transmission. In order to satisfy the bandwidth constraint, non-important segments are first adapted in the temporal dimension to reduce full motion video to static key frames. If further bandwidth reduction is needed, spatio-temporal adaptation operations such as frame rate reduction or SNR quality reduction are applied to the frames during the important segments.

The performance of the above content-adaptive streaming scheme depends on the video characteristics, e.g., the percentage of important segments in the whole stream. In the baseball experiment, we found non-important segments occupy more than 50% the content (in terms of time). Such a significant ratio provides a large room for bandwidth reduction by using the adaptive scheme. For real-time implementation, we also include automatic tools to detect important segments from incoming video. Our prototype shows feasibility of real-time performance and satisfactory detection accuracy (higher than 90%).

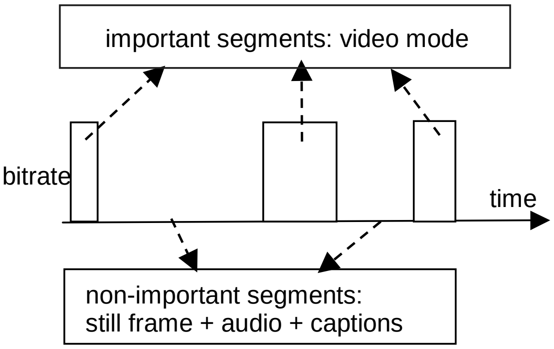

important segments: video mode
bitrate
time
non-important segments:
still frame + audio + captions

Figure 5. Content-based adaptive streaming of videos over bandwidth limited links.

4 OPEN ISSUES

The utility framework is general and can be used to model diverse problems in media adaptation. However, several issues arise- some require careful consideration while others need further investigation.

4.1 AMBIGUITY IN SPECIFYING ADAPTATION OPERATIONS

Some adaptation operations may not be uniquely defined. For example, an operation of “remove the tail half of each shot” seems clear. But in practice, the shot boundaries may not be exactly defined and need to be computed using shot change detection tools, which often are imperfect. For another example, an operation of “drop 10% of transform coefficients in a frame” does not specify the exact set of coefficients to be dropped. Different implementations may choose different sets and result in slightly different resource values and utility values.

There are several ways to address this issue. First, we can restrict that adaptation operations need to be defined 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 decoders are compliant with standards.

The second approach is to leave a certain window for implementation variations, but attempt to bound the resulting variations in resource and utility. Theoretical estimates of such bounds are hard if not impossible. Assuming there exist some consistence among practical implementations, empirical bounds of such variations may be obtained. For example, it can be reasonably assumed that shot segmentation tools are mature and results of shot boundaries should not be drastically different from each other. In the above case that suffers ambiguity in transform coefficient dropping, some restrictions may be imposed on the implementations. For example, a uniform dropping policy will ensure each block in the frame drops approximately the same percentage of coefficients. Imposing such constraints will help improve the consistence of the adapted output and the corresponding resource and utility values.

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 operations are critical. In such cases, the likelihood of achieving ranking consistence is higher than absolute value consistence. Such issues are also discussed in Section 3.2.

4.2 HETEROGENEOUS TYPES OF UTILITIES

The objective of maximizing the utility of an adapted entity implies that we need a ranking metric for comparing utilities. However, the utility space may include multiple dimensions, which are often of different types and difficult to integrate. For example, how do we compare an entity with high SNR, low comprehension utility with another with low SNR, but high comprehension utility?

The first option in addressing the above issue is to come up with a fusing model to combine different types of utilities to a single measure. However, such models often depend on user preferences and tasks, and are not linear. Another solution to the problem is to choose a subset of the utilities for maximization and impose constraints on the rest. For example, one possible objective is to maximize the comprehension utility subject to the constraint that the SNR utility is not less than some preset bound.

4.3 REPRESENTING DISTRIBUTIONS AND RELATIONS

Given distributions of points in the three high-dimensional spaces and mapping relationships among them, what are the appropriate representation schemes? Choices of the representation schemes will affect flexibility and efficiency of usage of the framework.

One approach is to sample the distribution in the adaptation space and store the corresponding resource and utility values as multi-variable high-dimensional matrixes (or tensors strictly speaking). If a certain scanning scheme is adopted, elements of a matrix can be represented by a one-dimensional list. Sampling and scanning schemes are often dependent on practical application considerations.

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

Appropriate representations vary and depend on actual applications. For example, in the case that the adaptation space has a single dimension of varying quantization step size, the classical representation of R-D curves will be appropriate. If the adaptation space has a high dimension but all the permissible operations reside on a highly correlated subspace (e.g., sum of adaptation operations in different dimensions is fixed), then an efficient representation is to sample the subspace and represent the resource/utility values using the sampled points as indexes. 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 adaptation ranking information as descriptor metadata is an adequate solution, as discussed in Section 3.2.

4.4 SEARCH STRATEGY IN LARGE SPACE

In video coding, the rate-distortion theory has been used to find optimal solutions in achieving optimal coding performance (in the rate-distortion sense). As we mentioned earlier, the rate-distortion model in coding correspond to one-dimensional cases in all three spaces-adaptation (quantization), resource (bit rate), and utility (SNR). In the general cases with high-dimension spaces, finding analytical optimal solutions or efficient computational strategies remains a critical issue.

4.5 MODELING AND PREDICTION OF DISTRIBUTIONS

Given a video and an adaptation engine, how do we obtain the distributions of points in each space? If off-line processing of video is feasible and computational cost is not a concern, distributions of *utility* and *resource* for a family of defined adaptation operations can be computed by testing all possible operations. Otherwise, efficient computational models or prediction methods need to be developed.

Analytical models can be used to compute the utility or resource based on computable attributes of videos. For example, in Section 3.1, utility of a video shot is modeled as a concave function of duration and visual complexity of a shot. Such analytical methods are derived based on human visual psychophysical models.

In some cases, analytical models are difficult to obtain. For example, it is difficult to model the SNR or subjective utility of an image frame when different coefficient quantization tables are used[1]. Brute-force simulations are often needed to obtain the resource and utility values of the video using each possible quatization table.

Content-Based Classification

In [7], we propose a content-based utility classification solution to the above problem. We assume that videos can be mapped to distinctive utility distribution classes based on their computable features, such as object complexity, motion, and coding parameters extracted from compressed streams. Video objects are segmented using semi-automatic tools and coded using a MPEG-4 encoder. A psychophysical modeled based quality measure was used to compute the utility of the encoded video. The utility distribution is sampled at a set of predefined quantization step sizes. Utility values at the sampled points are concatenated and used as a multi-dimensional feature vector to represent each specific distribution. Similar representation scheme are used for the resource distribution. Thus given a set of adaptation points (e.g., a set of N quantization step sizes), the corresponding distributions in resource and utility can be modeled as N-dimensional feature vectors.

[1] Some approximation models based on human vision models may exist.

Unsupervised clustering was used to obtained classes for utility and resource distributions both separately and jointly. Then automatic classifiers were learned to classify new videos to utility/resource classes based on the computed visual and coding features. Experiments showed a reasonable accuracy, 80%-85%.

Classification Targets

The above example uses computable features to directly predict the distribution classes of resource and utility. An alternative is to predict the preferred adaptation operations given a resource or utility constraint. For example, in the MPEG-4 optimal transcoding project described in Section 3.2, we may not be concerned with the distributions of the utility or resource. Instead, we are interested in knowing the best adaptation operation among multiple options that satisfy a certain constraint.

In such a case, the target variable of classification we are interested is the preferred adaptation operation or the ranking among candidate operations. We can first sample the resource space. For each sampled point or bin, we find the preferred adaptation class or the ranking among all candidate operations. Then the task is to select suitable features and develop classifiers that can predict the preferred operations for new videos based on the computed features.

5 CONCLUSIONS

Content adaptation is an important technique for maximizing accessibility and functionalities of audio-visual content in pervasive or interactive multimedia environments. There exist a rich set of adaptation methods in different levels and modalities. Many important factors of the resources and user preferences also have to be considered. A systematic methodology is needed to understand the problem space, formulate the objectives, and develop strategies for designing effective solutions.

We present a new conceptual framework to model the relations among adaptation, resource, and utility. We define the concept of entity, which can be abstracted at multiple levels and modalities. We define the utility in a multi-dimensional space. Every point in the adaptation space transforms the input entity to an adapted one, and thus changing the corresponding utility and required resources.

The framework is general and flexible. We use it to model several problems and solutions we have developed in several media adaptation projects, ranging from video skimming, tarnscoding, to adaptive streaming. All these problems are conveniently modeled as resource- or utility-constrained optimization issues as long as the entity, utility and adaptation space are adequately defined.

We also identify several important issues in applying the framework and associated optimization methodology to practical applications. Especially, interesting issues arise in developing automatic methods for modeling and predicting resource and utility, representing high-dimensional distributions, and effective strategies for searching optimal adaptations given constraints of resource, utility, and user preference.

5 ACKNOWLEDGEMENTS

The author thanks Raj Kumar Rajendran and Jae-Gon Kim for their kind help in reviewing the paper and discussing some concepts described in the paper.

REFERENCES

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

[2] J. R. Smith, R. Mohan and C. Li, "Scalable Multimedia Delivery for Pervasive Computing", ACM Multimedia Conference (Multimedia 99), Oct. -Nov., 1999, Orlando, Fl.

[3] H. Sundaram, L. Xie,, and S.-F. Chang, “A Utility Framework for the Automatic Generation of Audio-Visual Skims,” ACM Multimedia, Juan-les-Pins, France, Dec. 2002.

[4] J.-G. Kim, Y. Wang, S.-F. Chang, K. Kang, and J. Kim, “Description of utility function based optimum transcoding,” MPEG-21 DIA contribution, ISO/IEC JTC1/SC29/WG11 MPEG02/M8319, Fairfax, May 2002.

[5] R. Kumar Rajendran, M. van der Schaar, S.-F. Chang, “FGS+: Optimizing the Joint Spatio-Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding,” IEEE International Symposium on Circuits and Systems (ISCAS), Phoenix, Arizona, May 2002.

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

[7] P. Bocheck, Y. Nakajima and S.-F. Chang, “Real-time Estimation of Subjective Utility Functions for MPEG-4 Video Objects,” Proceedings of IEEE Packet Video Workshop (PV'99), New York, USA, April, 1999.