Utility-Based Video Adaptation for Universal Multimedia Access (UMA) and Content-Based Utility Function Prediction for Real-Time Video Transcoding

Yong Wang, *Member, IEEE*, Jae-Gon Kim, *Member, IEEE*, Shih-Fu Chang, *Fellow, IEEE*, and Hyung-Myung Kim, *Senior Member, IEEE*

***Abstract*—Many techniques exist for adapting videos to satisfy heterogeneous resource conditions or user preferences, whereas selection of the best adaptation operation among various choices usually is either *ad hoc* or inefficient. To provide a systematic solution, we present a conceptual framework based on utility function (UF), which models video entity, adaptation, resource, utility, and the relations among them. In order to support real-time video adaptation, we present a content-based statistical paradigm to facilitate the prediction of UF for real-time transcoding of live videos. Instead of modelling the UF through analytical models, as in the conventional rate-distortion framework, the proposed approach formulates the prediction as a classification and regression problem. Each video clip is classified into one of distinctive categories and then local regression is used to accurately predict the utility value. Our extensive experiment results based on MPEG-4 transcoding demonstrate that the proposed method achieves very promising performance—up to 89% accuracy in choosing the optimal transcoding operation (in both spatial and temporal dimensions) with the highest quality over a diverse range of target bit rates.**

***Index Terms*—Content-based prediction, universal media access, utility function, video adaptation.**

I. Introduction

AN EMERGING multimedia framework, in which multimedia content is accessed from heterogeneous networks and terminals in a seamless way, is often referred to as universal multimedia access (UMA) [9]. In UMA, media content adaptation is considered to be a core technology for coping with variations of environment resources and user preferences. Media adaptation is a process that transcodes the original encoded media into a new version in order to match resource constraint (e.g., bandwidth and resolution) or user preference. Many adaptation methods exist for adjusting the bit rate of compressed video streams. For example, requantization of transform coefficients [19], frame dropping (FD) [13], DCT coefficients dropping (CD) [6], and resolution reduction [20] are commonly used. More discussion involving transcoding for UMA can be found in [2]. To address heterogeneous resources and user conditions, some recent developments in scalable video coding have been made with greater flexibility and improved video quality [5]. Nevertheless, most existing adaptation techniques have a common problem—they concentrate on optimization of pre-selected adaptation operations, rather than systematically choosing the optimal adaptation operation from multiple options. The issue becomes more prominent when the number of adaptation dimension increases, including spatial, temporal, and signal-no-noise ratio (SNR). In the literature, there are a few efforts to address this issue. In [15], a rate-distortion (R-D) optimization method was proposed by modelling the mean-squared-error distortions caused by quantization and frame skipping. In [12], a dynamic programming scheme was used to achieve optimal rate control where frame rate, spatial resolution and quantization step size were jointly considered in modelling the distortion. A distortion measurement was used to estimate the video quality in the full resolution, while some weights were assigned to address the perceptual effects of spatio-temporal scale variation. In [7], variable frame rate coding was realized, where the quantization step size was determined by an analytical distortion model for each frame, and the frame with quantization step size exceeding some threshold was skipped. Nevertheless, all of these approaches rely on the availability of some analytical models. Construction of adequate analytical models of R-D relationship is known to be nontrivial. It is even more difficult when video undergoes multidimensional adaptation.

In this paper we present a general framework, called utility-based video adaptation, as a systematic solution for the issue of spatio-temporal combined adaptation. Specifically three key aspects involved in adaptation problems—adaptation (A), resource (R), and utility (U)—are modelled and represented using a utility function (UF), which describes the tradeoff relationship between resources and utilities along each adaptation dimension. UF plays a key role in choosing the optimal adaptation among multiple options that meet resource constraints or

Manuscript received March 28, 2004; revised May 1, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ching-Yung Lin.

Y. Wang and S.-F. Chang are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: wangyong@gmail.com; sfchang@ee.columbia.edu).

J.-G. Kim is with the Department of Broadcasting Media Technology, Electronics and Telecommunications Research Institute (ETRI), Daejeon 305-350, Korea (e-mail: jgkim@etri.re.kr).

H.-M. Kim is with the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea (e-mail: hmkim@csplab.kaist.ac.kr).

Color versions of Figs. 4, 5, and 7–9 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2006.886253

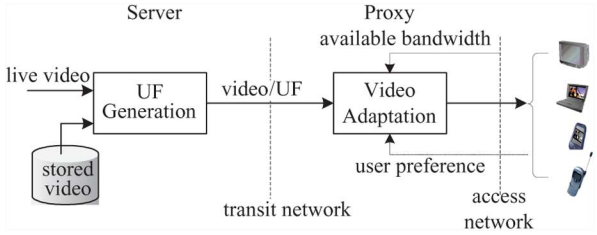


Fig. 1. Three-tier adaptation architecture using the utility-based framework.

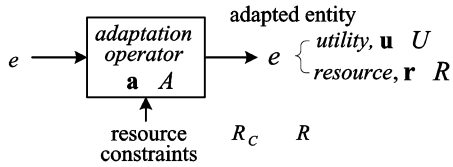


Fig. 2. Definition of adaptation, resource, and utility spaces involved in video adaptation problems in the utility-based framework.

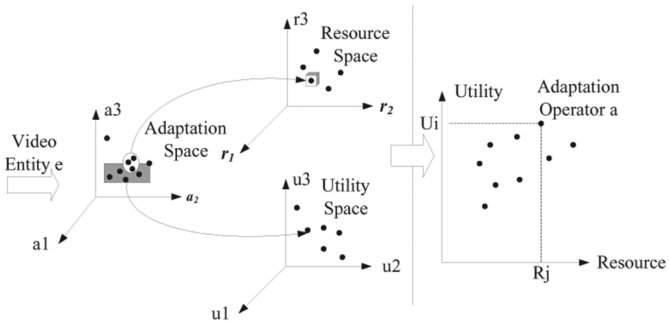


Fig. 3. Use UF to describe relations among adaptation, resource, and utility.

user preferences. This approach represents a simple extension of conventional R-D framework to allow incorporation of diverse types of resources (e.g., complexity and bandwidth) and adaptation operations. In the utility-based framework, one key issue is how to generate UF in real time to accommodate live videos. For stored videos in on-demand applications, UF can be generated by exhaustive off-line simulations. Such an approach may require significant computational complexity, which is unacceptable for live videos. In this paper we present a novel real-time UF-prediction method that utilizes the strong correlations between content features and the UF characteristics of a video. The prediction method, first described in [17], combines real-time compressed-domain feature extraction, pattern discovery, classification, and statistical regression. We formulate the problem as a pattern classification and prediction question, taking the automatically extracted content features as input and then making predictions about the UF. The only on-line computation required is for content feature extraction and pattern classification. Therefore, the proposed approach is fully automatic and can be done in real-time. Our extensive MPEG-4 transcoding experiment results show a very promising accuracy (up to 89%) in choosing the optimal operation from several competing options. The main contributions of our work include the formulation of UF for the joint spatio-temporal adaptation and the novel algorithms for predicting the optimal adaptation operation based on the content feature extracted from the compressed streams.

The rest of this paper is organized as follows. The framework of utility-based transcoding is introduced in Section II. In Section III, the statistical approach to UF prediction is described, including feature extraction, unsupervised, and supervised learning methods and statistical local regression. The experiment setup and results are presented in Section IV. The conclusion and future work are given in Section V.

II. Utility-Based Transcoding

The UF-based adaptation approach mentioned above fits a popular three-tier server-proxy-client architecture very well, as shown in Fig. 1. The adaptation engine deployed in the proxy adapts incoming videos to satisfy dynamic resource constraints that are not known *a priori*. The role of the UF is to describe the relationship between required resources and resulting video utilities when the video is subject to various adaptation operations in multiple dimensions. For stored videos, UF can be generated offline at the server and sent to the adaptation engine. The engine will then select the optimal adaptation operation based on the information in the UF. For live videos, UF needs to be obtained on the fly through some estimation and update processes. In this paper, we specifically propose a content-based prediction method that estimates the UF according to the content features and statistical classification tools. Such real-time prediction methods can be implemented at either the server or the proxy.

A. Adaptation, Resource, Utility and Their Relations

UF is defined in the adaptation-resource-utility (ARU) space, where relationships among diverse types of adaptations, resources (e.g., bandwidth, power, and display) and utilities (e.g., objective or subjective quality) are modeled. We use the term "space" in a loose sense here to indicate the multiple dimensionalities involved. Fig. 2 depicts the notions of ARU involved in a video adaptation problem. The entity, e, refers to the basic unit of video data that undergoes the adaptation process. Adaptation operators are the methods to reshape the video entities, such as requantization and FD. All permissible adaptations for a given video entity constitute the adaptation space. Resources are constraints from terminals or networks, including bandwidth, display resolution, power, etc. Utility represents the quality of an entity when it is rendered on an end device after adaptation, such as PSNR, perceptual quality, or even high-level user satisfaction. The mapping relationship among ARU spaces is illustrated in Fig. 3. Typically, there exist multiple adaptation solutions that satisfy the same resource constraints, while yielding different utilities. In Fig. 3, the points in the oval shaped region in the adaptation space indicate such a constant-resource region. Likewise, different points in the adaptation space (the shaded rectangle) may lead to the same utility value. It is such a multioption situation that makes the adaptation problem interesting—we want to choose the optimal one with the highest utility or minimal resource.

We are interested in describing the relationship between rate and utility associated with each adaptation operation. We represent such relationship by using UF. The right figure in Fig. 3

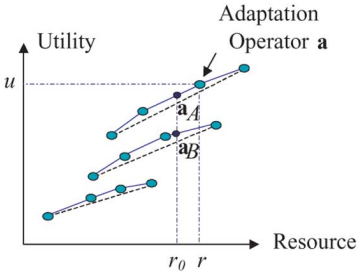


Fig. 4. Definition and representation of utility function.

shows a simple example of UF, in which only one dimension is shown in both resource and utility. This is equivalent to the known R-D curve when R is bit rate and D is related to video quality. Each point in the UF is associated with one specific adaptation operator, which may include combinations of multiple operations (such as FD and coefficient dropping). A more detailed discussion of the UF function can be found in [4].

B. FD-CD Adaptation

To illustrate our method of utility-based adaptation without losing generality, in this paper we consider a specific case involving two types of adaptations—FD and AC DCT coefficient dropping (CD) and their combinations (FD-CD). FD adapts the source stream by skipping frames, while CD transcodes the source stream by truncating some high-frequency DCT coefficients. For CD, there is more than one choice during coefficient dropping. Therefore, in order to eliminate the ambiguity and obtain the optimal CD performance, the Lagrange optimization method is employed. Typically, suitable rate-control techniques are needed to meet a specific bandwidth constraint after FD-CD adaptation. (Due to space limitations, details about the FD-CD algorithm and its implementations can be found in [16].) The advantage of FD-CD adaptation firstly lies in its simplicity, allowing real-time implementation. Also, FD can meet a coarse level of the target rate since its processing data unit is a frame. CD is able to meet the target rate with a finer granularity by adjusting the amount of dropped coefficients. The combination of FD-CD accommodates a wide range of bit-rate constraints. Furthermore, FD-CD provides adequate flexibility in balancing the trade-off between spatial and temporal quality. For simplification, we assume the entity undergoing adaptation is a group of pictures (GOP) in the MPEG-4 sequence. Namely, the same FD-CD operation parameters will be applied to the whole GOP.

C. FD-CD Representation Using Utility Function

Using the utility-based adaptation framework, a two-dimensional (2-D) adaptation space can be constituted for FD-CD, in which both FD and CD entail a finite set of adaptation operations. Specifically, an FD-CD adaptation method can be expressed as $\mathbf{a} = (f, c)$, where f and c represent a specific FD method and a coefficient dropping method respectively. For instance, $\mathbf{a} =$ (all B-frames dropped, 10%) means all of the B frames in a GOP are dropped and 10% of the bits from each remaining frame will be reduced by coefficient dropping. A typical UF is shown in Fig. 4. For a specific video clip, given an adaptation operator $\mathbf{a}$, its corresponding resource and utility value are denoted as r and u. In the case of FD-CD, we have coarse discrete values of FD (i.e., "no frame dropping", "drop all B and P frames", "drop all B frames", and "drop 1 B frame only"), and finer discrete values of CD (i.e., "drop $c\%$ of DCT coefficients"). Thus, in Fig. 4, points with the same FD are connected to a curve, and the adaptation operations between two anchor nodes are obtained through linear interpolation. The whole set of the curves define the UF, which represents the utility-resource relation associated with the given video in response to the available adaptation operations (FD-CD). Given a resource constraint r_0, all of the possible operators meeting the same resource constraint, such as $\mathbf{a}_A$ and $\mathbf{a}_B$ in Fig. 4 can be found from the UF. If an operation is selected using the actual UF, it will achieve the target resource and the utility when it is applied to the video. If the operation is selected based on predicted UF, the actual resource and utility resulting from the adaptation may be slightly different from the target values due to prediction errors. Such utility-based adaptation mechanism was also accepted as a part of MPEG-21 digital item adaptation (DIA) [10]. More information about the DIA utility-based description tool can be found in [8].

To obtain a more efficient representation, we further simplify the representation of the UF by using the linear approximation of each curve as shown in Fig. 4. The approximation is defined by two end nodes of each curve. Therefore, the UF can be denoted as

$$\begin{aligned}\mathbf{F}^{\mathrm{UF}} &= (f_1^{\mathrm{UF}}, f_2^{\mathrm{UF}}, \ldots, f_N^{\mathrm{UF}}) \\ &= (r_1, r_2, \ldots, r_{2N}, u_1, u_2, \ldots, u_{2N})\end{aligned}$$

where each curve $(f_i^{\mathrm{UF}}, i = 1, 2, \ldots, N)$ in Fig. 4 is associated with two end points. Such approximation representation is very beneficial in reducing the dimensionality of the representation and improving the efficiency of the statistical-prediction method described later. Our experiment demonstrates that such linear approximation provides a very satisfactory result in UF prediction (shown in Section IV-B). The ordering of the nodes does not matter, as long as a consistent scheme is maintained.

D. Issues in Computing the Utility Function

In practice, the generation of the UF is a nontrivial process. It may be done by exhaustive computation of all of the adaptation points, each of which requires transcoding of the video, decoding the transcoded bit stream, and computing the distortion. This process is very time consuming and typically cannot be done efficiently. To avoid exhaustive computation, there are two possible solutions: analytical modeling or empirical estimation.

1) Approximate Analytical Modeling: In [7], some analytical source models were developed by extending the theoretical R-D curves derived from ideal statistical distributions to approximate models using empirical data. Certain statistical models (e.g., Gaussian, Laplacian, or variations) were assumed for video signals and parameters of the models were computed and updated from input video. Given the approximate R-D information, some recent methods have been developed to automatically adjust the frame rate and quantization step size

under low-bit-rate conditions [12]. Nevertheless, such analytical models may not be valid in general due to several reasons. First, the adopted signal models like Gaussian or Laplacian may not be valid for realistic signals. Second, the R-D relationship is greatly affected by the specific coding algorithm, which has become increasingly complex in recent video coding technologies. Simple statistical signal models may not be valid for such complex coding methods. Lastly, it is difficult to extend the analytical models in order to take into account different coding structures, utilities (e.g., subjective measures), and resources (e.g., power).

2) Empirical Estimation and Content-Based Prediction: Another approach to R-D estimation is based on empirical learning—namely, learning from the training data. Such an approach does not use explicit statistical models for the video signals to derive the R-D curves. Instead, collection of training video clips are used to generate samples of video content features and the resulting UFs, represented by some efficient schemes described in the previous section. Machine learning techniques are then applied to develop mapping functions from the content features to the UFs. We refer to the aforementioned approach as *content-based utility function prediction.*

The above prediction methods explore the potential correlation between content features and the R-D characteristics of a video. Such correlation has been observed in our experimental observations (Section IV). Here, for video content, we refer to low-level features such as motion, spatial complexity, and characteristics of the coded stream (e.g., number of inter-frame coded macroblocks, motion vector statistics, etc.). Such features can be efficiently computed from the compressed streams. In our prior work [3], we have explored such an approach in which visual features from the video objects are used to predict the subjective quality of the objects after undergoing MPEG-4 transcoding. Nevertheless, the work in [3] did not explore systematic representations of the UFs for FD-CD adaptation and did not address issues related to prediction of the optimal spatio-temporal adaptation operation.

III. Utility Function

A. Problem Description

The issue of UF prediction can be formalized as follows: given the content feature $\mathbf{F}^{\mathrm{CF}}$ of one video clip, develop a suitable mapping from the content feature space into the UF space, i.e.,

$$\mathbf{F}^{\mathrm{UF}} = \mathbf{G}(\mathbf{F}^{\mathrm{CF}}) \tag{1}$$

where $\mathbf{F}^{\mathrm{UF}} = \left(f_1^{\mathrm{UF}}, f_2^{\mathrm{UF}}, \ldots, f_N^{\mathrm{UF}}\right)$ is a N-dimension UF row vector and f_i^{UF} is the i^{th} component of $\mathbf{F}^{\mathrm{UF}}$, and similarly $\mathbf{F}^{\mathrm{CF}} = (f_1^{\mathrm{CF}}, f_2^{\mathrm{CF}}, \ldots, f_M^{\mathrm{CF}})$ is the M-dimension content feature row vector and f_j^{CF} is the j^{th} component of $\mathbf{F}^{\mathrm{CF}}$. Equation (1) is a typical multivariate regression problem. For each f_i^{UF} in $\mathbf{F}^{\mathrm{UF}}$, we want to find a mapping g_i, such that

$$\begin{aligned} f_i^{\mathrm{UF}} &= g_i(\mathbf{F}^{\mathrm{CF}}) = g_i\left(f_1^{\mathrm{CFD}}, f_2^{\mathrm{CF}}, \ldots, f_M^{\mathrm{CF}}\right) \\ \mathbf{G} &= (g_1, g_2, \ldots, g_N). \end{aligned} \tag{2}$$

By using Taylor expansion, this mapping can be written as

$$f_1^{\mathrm{UF}} = g_i(\mathbf{F}_0^{\mathrm{CF}}) + \nabla g_i(\mathbf{F}_0^{\mathrm{CF}}) \cdot (\mathbf{F}^{\mathrm{CF}} - \mathbf{F}_0^{\mathrm{CF}}) + O\left(\left|\mathbf{F}^{\mathrm{CF}} - \mathbf{F}_0^{\mathrm{CF}}\right|^2\right) \tag{3}$$

where $(\cdot)$ is the dot product of two vectors, and $\nabla g(F_0^{\mathrm{CF}})$ is the M-dimension partial differential row vector. By keeping the components in (3) up to first order and ignore the higher orders, this mapping can be considered as a classic linear regression problem. Based on (3), we can derive the following:

$$\begin{aligned} \mathbf{F}^{\mathrm{UF}} &= \left(f_1^{\mathrm{UF}}, f_2^{\mathrm{UF}}, \ldots, f_N^{\mathrm{UF}}\right) \\ &= \left[\mathbf{F}^{\mathrm{CF}} 1\right] \begin{bmatrix} \nabla \mathbf{g}_1^T & \nabla \mathbf{g}_2^T & \ldots & \nabla \mathbf{g}_N^T \\ c_1 & c_2 & \ldots & c_N \end{bmatrix} \\ &= \mathbb{F}^{\mathrm{CF}} \mathbb{G} \\ c_i &= g_i\left(\mathbf{F}_0^{\mathrm{CF}}\right) - \nabla \mathbf{g}_i\left(\mathbf{F}_0^{\mathrm{CF}}\right) \cdot \mathbf{F}_0^{\mathrm{CF}} \\ \nabla \mathbf{g}_i^T &= \nabla \mathbf{g}_i^T\left(\mathbf{F}_0^{\mathrm{CF}}\right), \quad i = 1, 2, \ldots, N. \end{aligned} \tag{4}$$

By applying the standard least-square error (LSE) method, the optimal estimation of $\mathbb{G}$, indicated as $\hat{\mathbb{G}}$, can be found to be

$$\hat{\mathbb{G}} = \left((\tilde{\mathbb{F}}^{\mathrm{CF}})^T \tilde{\mathbb{F}}^{\mathrm{CF}}\right)^{-1} (\tilde{\mathbb{F}}^{\mathrm{CF}})^T \tilde{\mathbf{F}}^{\mathrm{UF}} \tag{5}$$

where $\tilde{\mathbb{F}}^{\mathrm{CF}}$ is the set of observed content feature vectors in the training data with each row corresponding to a training sample, and $\tilde{\mathbf{F}}^{\mathrm{UF}}$ is the observed UF. Moreover, the Taylor expansion works only in a small neighbor of the center $\mathbf{F}_0^{\mathrm{CF}}$. Thus, the first-order approximation is effective if the content feature space can be divided into some small areas, and the regression procedure is applied for each area separately. Specifically, this can be done by forming K such subareas S_k, $k = 1, 2, \ldots, K$, and conducting the above approximate estimation method for points within each subarea. Therefore, the problem can be modeled as a K -segment piecewise linear regression problem and the parameters (c_i and $\mathbf{g}_i$) can be obtained for each subset. In forming the partitions of the space, we can consider clustering in the CF space, clustering in the UF space, or a hybrid one that partition the CF space subject to some compactness constraints of corresponding UF values. In this work, in order to apply the local regression method discussed above we chose clustering in the CF space, plus combinations of classification techniques mapping content features to the CF clusters. Our experiment results presented later indeed confirm the superiority of this choice.

Fig. 5 shows the overall architecture of the proposed framework. The top diagram shows the procedures of extracting content features, classifying the video to one of the UF classes, and local regression for predicting the UF. In the bottom diagram, an offline mechanism is shown to illustrate the use of a training pool in developing the UF clusters, the classifier for mapping future clips, and the local regression method for each cluster. Each training clip is associated with the content features and the actual UFs that are obtained in advance through exhaustive computation of all the adaptation operations. Details of each component mentioned above will be described in the following subsections.

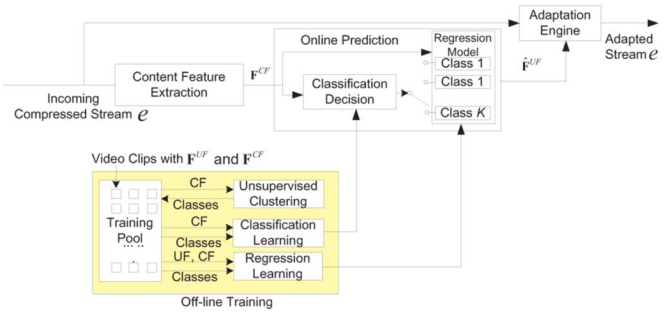


Fig. 5. Overall architecture of the proposed framework.

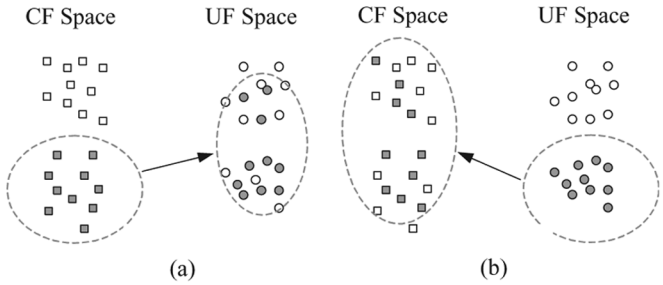


Fig. 6. Difference between CF clustering and UF clustering. Shaded points show a cluster formed in one space and the corresponding values in the other space.

B. Content Feature Extraction

We adopt the content features based on the set adopted in our prior work [3] with minor modification. Three groups of features are considered: motion intensity, AC DCT energy, and quantization parameters. The first two groups of features embody the spatial texture complexity and temporal motion intensity information. The third group also indirectly reflects the scene complexity subject to the specific rate control algorithm used. They are extracted directly from the encoded stream or the stream metadata without decoding the video to the pixel domain. Our experimental results show that the performance of prediction can be improved if we also include the peak SNR (PSNR) information from the metadata associated with the original encoded stream.

Content features are extracted from each local video segment that is one second long. The length of the local segment is currently empirically determined, to keep an adequate balance between efficiency and accuracy. Note to ensure the video content in each segment is more or less consistent, we avoid shot boundaries within a segment by running automatic shot boundary detection and keeping the shot boundaries aligned with the segment boundaries. Although the shot boundary detection tool is not perfect, performance of the existing detection tools is quite high (precision up to 97% and recall up to 98% in [22]). Specifically, the following features are used in our system:

1) average motion intensity approximated by computing motion vector magnitude;
2) motion variance within the adaptation unit;
3) average percentage of macroblocks which have non-zero motion vector;
4) average I frame AC DCT coefficient energy;
5) average P frame AC DCT coefficient energy;
6) average quantization step size;
7) average PSNR if available in the stream metadata.

C. Unsupervised Clustering

The average values are computed over the frames in the one-second segment. To further improve efficiency, we only process the I and P frames. The AC DCT energy of I and P frames are kept separate because our statistical feature analysis (principal component analysis) shows they have distinctive contributions to the final performance. This is reasonable considering the DCT energy in the I frame is more related to the texture complexity due to the use of intra-frame coding, while the DCT energy in the P frames is mainly related to motion compensation residues because of the use of inter-frame coding.

The purpose of unsupervised clustering is to partition the content feature space into separate subspaces so that the local regression technique described in Section III can be applied in each local area. We adopt the K-harmonic mean (KHM) [21] clustering method, which in principle is related to the popular K-mean method. The main improvement of KHM over K-mean is by using the p^{th} -order harmonic distance, rather than the Euclidian distance. It was shown in [13] that KHM outperform K-mean in reducing the sensitivity to initialization and avoiding local optimal points. Note the above clustering process is performed in the CF space, instead of the UF space. As shown in Fig. 6, clusters formed in the CF space will ensure points in the same cluster have similar CF values. This is important for keeping a subarea of small variation of CF values and thus the first-order approximation by Taylor expansion described in Section III-A remains valid. Although the alternative of doing clustering in the UF space can achieve compact data sets with similar UF values. The corresponding values in the CF space may be spread over a large range, and thus violate the assumption of proximity of the local regression method mentioned above. We will present performance comparison of these competing options later.

Another general problem of unsupervised clustering is determining the number of cluster, K. A K value that is too large will lose generality and result in overfitting, while a K value that is too small will result in significant bias. In our experiment, we determine the number of cluster through empirical trials and find $K = 16$ yields satisfactory performance. We expect the adequate choice of K depends on the characteristics of the video content and dynamic variations of the video over time. It is conceivable to propose some prediction schemes to determine the cluster number based on computable content features. Study of such methods and analysis of the effect on the UF-prediction performance is beyond the scope of the current work.

D. Supervised Classification by SVM

The purpose of classification is to categorize an incoming video clip into one of the classes and then apply the corresponding regression model to predict the UF. Note if the classes are formed by clustering in the CF space, the same clustering method can be used for such a classification purpose. But if the clusters are formed in the UF space, the corresponding points in

TABLE I
SUMMARY OF DATA SET

Video source	*A Beautiful Mind* (736 clips), *Crouch Tiger Hidden Dragon* (589 clips) and *Taxi II* (741 clips)
Clip length	1 second (2 GOPs)
Image format	352 x 240 pixels
Videocompression	MPEG-4 with 30 fps and TM5 rate control
GOP structure	GOP size N=15, sub-GOP size M=3

the CF space may not be compact and therefore we need a separate process for classification. We employ support vector machine (SVM) for the classification task. Basic SVM classifiers are for two-class discrimination. There are several ways to extend a binary classifier to support multiple-class separation, such as classifiers for one against others [14], or ones that fuse a set of two-class classifiers by methods like the Max Wins algorithm in [1]. We adopt the directed acyclic graph SVM (DAGSVM) algorithm presented in [11] with minor modification to resolve the ambiguous region issue. In DAGSVM, the multiple-class classifier is constructed by using a decision directed acyclic graph. The classifier starts with separation between two most distinguishable classes using a regular two-class SVM. The negative class is excluded and the same two-class discrimination procedure is repeated for the remaining classes. It has been known to be a fast multiclass classifier with satisfactory performance [11].

IV. EXPERIMENT RESULTS

A. Experiment Setup

In our experiment, we selected video from three movies to form the training and testing pool. The details of the video pool are summarized in Table I. There were 2066 clips in total, each of which was 1-s long. The clips were carefully selected to cover a wide range of content features. Every clip was extracted from within a shot and thus no abrupt transitions like shot changes occurred within a clip. The proposed algorithm was tested using a standard cross validation procedure in which training and testing was done with random partitions of the pool (70% for training and 30% for testing) over multiple runs. First, we need to compute UFs and extract content features for each set of training clips. In computing UFs, we defined an adaptation space of FD-CD similar to that described in Section II-C (see Fig. 4). Based on the given GOP structure ($N = 15$, $M = 3$, $IBBPBBP\ldots$), we adopt four FD operators: "no frame dropped", "the first B frame dropped in each sub GOP", "all B frames dropped", and "all B and P frames dropped\rdquo". In the CD dimension, we adopted six CD levels: from 0% to 50% with 10% increment. As a result, there were totally 24 anchor nodes and four operation curves in each UF. Further details of the implementation are described in [16]

Evaluation of the proposed prediction method can be based on various performance metrics. For example, errors in predicting the UF can be defined based on the L_2 metric as follows:

$$D = \frac{1}{L}\sum_{t=1}^{L}\left\|\mathbf{F}_l^{\mathrm{UF}} - \hat{\mathbf{F}}_l^{\mathrm{UF}}\right\|^2 \tag{6}$$

TABLE II
ALGORITHM SPECIFICATION

Unsupervised Clustering	K-Harmonic Mean (KHM) using p-th order harmonic distance. $p = 0.5$, Number of clusters $K = 16$
Classification	DAGSVM multiple-class classification: $C = 100$, kernel=RBF with $\gamma = 0.5$
Linear Regression	Trained by LSE algorithm. See Equation (5).

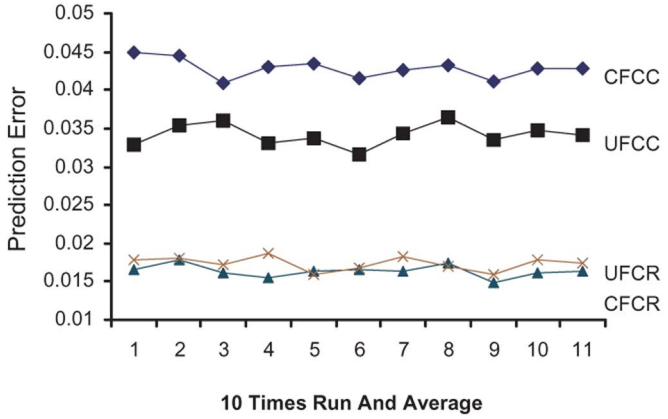


Fig. 7. Comparison of the prediction performance in terms of prediction error.

where $\mathbf{F}_l^{\mathrm{UF}}$ is the actual UF and $\hat{\mathbf{F}}_l^{\mathrm{UF}}$ is the predicted one. L is the number of the test clips. Alternatively, the utility ranking of permissible operators at fixed bitrates can be evaluated, comparing results using the predicted UF versus the ground-truth UF.

Table II is the specification of the algorithms employed in the experiment.

B. Performance

Fig. 7 shows the prediction errors from four methods: our proposed method, content-feature clustering-based regression (CFCR); our proposed method but without local regression, content-feature clustering-based classification (CFCC); an alternative approach using clustering in the UF space instead of the CF space, UF-clustering-based classification (UFCC), which is adopted in [3]; and UF-clustering-based regression (UFCR). The prediction error is measured by the L_2 distance between the true UF and the predicted UF [see (6)]. The experiments were run ten times and the average performance was computed. The proposed method (CFCR) achieves the best result. That is to say, when classification is combined with regression, clustering in the CF space is the best. This validates our decision in adopting the CF-space clustering method. Nevertheless, it is interesting to note that without regression, techniques using clustering alone (UFCC and CFCC) favors clustering in the UF space. This is consistent with the UF-space clustering techniques used in our prior work [3]. Note in the pure clustering approach, the representative UF of each cluster is used as the predicted UF for all the points mapped to the same cluster.

Fig. 8 shows comparison between some predicted UFs and the corresponding ground truth. The predicted UFs indeed match the true values very well. Typically, the prediction of

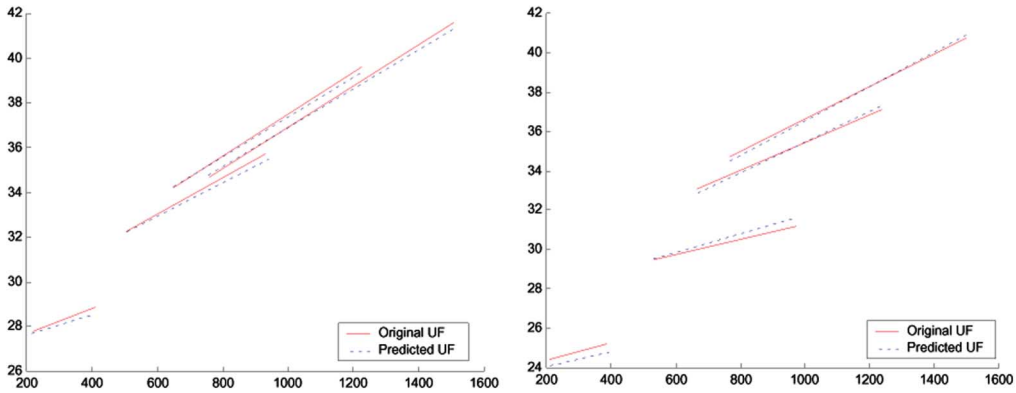


Fig. 8. Matching predicted UF to the ground truths.

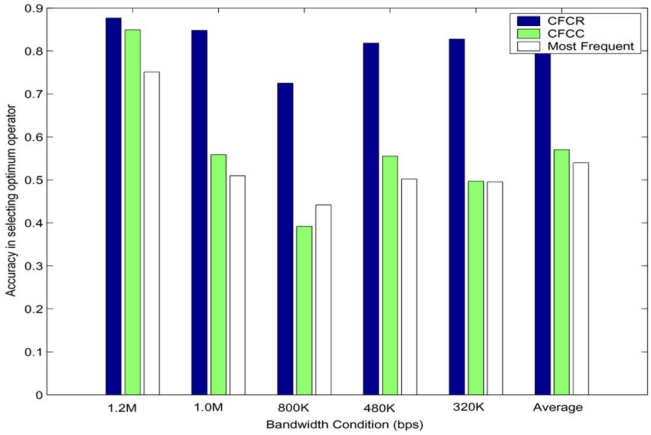


Fig. 9. Performance of prediction accuracy in choosing the optimal operator.

the utility value (y axis) is not as good as the prediction of the resource value (x axis). However, the ranking of utility values among different transcoding options are quite consistent. Such ranking information provides the most important input to our adaptation system for selecting the optimal transcoding option meeting a given target resource constraint.

In addition, we also measured the accuracy in selecting the optimal operator given various target bit rates. Five typical bandwidths were used as the test target rates: 1.2 M, 1.0 M, 800 K, 480 K, and 320 kbps. The original input video rate before transcoding was 1.5 Mbps. Our proposed method (CFCR) was compared with two alternatives: CFCC and the most frequent adaptation method. The latter did not take into account content features in each video, and simply selected the operation that achieves the highest quality for the most number of video clips in the training pool. Fig. 9 shows our method outperforms the other two and exhibits significantly higher accuracy (up to 89%).

From both the above evaluation criteria, the results are quite encouraging—the proposed content-based prediction method achieves very good accuracy in predicting the UF values as well as the ranking among competing adaptation operations.

Besides prediction performance, computational complexity is another important factor for a real-time application. Because the MPEG-4 codec we used was not a real-time implementation, we did not provide the real-time benchmark data. However, all of the computation processes in our system are light-weight. As shown in Fig. 5 the main costs in our system include feature extraction and online prediction. The online-prediction process, including classification and regression, can be implemented efficiently. Specifically, SVM classification only needs to calculate the kernel function and dot product between the content features and a sparse set of support vectors; linear regression involves only a multiplication between the model matrix and content feature vector. For feature extraction, partial bit-stream decoding is necessary in order to obtain the content features, plus some minor extra calculation such as computing averages. The combination of all these computation steps is still much lighter than the complexity of a regular decoder (because the most complex component, motion compensation, is not needed). Considering video decoders can be implemented on most platforms with a real-time performance, it is reasonable to conjecture that our system can be implemented in a real-time fashion as well.

V. Conclusions and Future Work

In this paper we present a utility-based video adaptation framework as a systematic methodology to meet diverse resource conditions and user preferences in UMA. The framework explicitly models the major concepts involved in adaptation processes—adaptation, resource, and utility using a UF. In order to address the computational complexity issue in UF generation and support the real-time adaptation scenario, we further propose a general content-based UF-prediction approach using automatic content feature extraction, and regression over clustering and classification. Our experiment results using MPEG-4 FD-CD transcoding demonstrate very promising prediction accuracy over diverse types of video content. The proposed content-based utility-prediction framework is general and can be expanded to handle heterogeneous scenarios where various resources, adaptations, or utilities are employed. Recently, we have successfully expanded our framework to scalable video coding and subjective evaluation utility case with satisfactory results [18]. Future work will include extensions that consider multiple utilities and resources at the same time. An example scenario is to find the balance between bandwidth demand and power consumption in selecting appropriate adaptation for handheld devices.

ACKNOWLEDGMENT

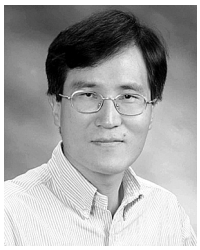
This work has been partly supported by Electronics and Telecommunications Research Institute of Korea. We thank Dr. K. Kang and Dr. J. Kim of ETRI for their discussions and suggestions. We also thank the anonymous reviewers for careful reviews and valuable comments.

REFERENCES

[1] T. Berger, *Rate Distortion Theory*. Upper Saddle River, NJ: Prentice Hall, 1971.
[2] N. Björk and Y. Christopoulos, "Video transcoding for universal multimedia access," in *Proc. ACM Multimedia Workshops*, 2000, pp. 75–79.
[3] P. Bocheck, Y. Nakajima, and S.-F. Chang, "Realtime estimation of subjective utility functions for MPEG-4 video objects," in *Proc. IEEE Packet Video Workshop (PV'99)*, Apr. 1999.
[4] S.-F. Chang, "Optimal video adaptation and skimming using a utility-based framework," in *Proc. Int. Workshop Digital Communications*, Capri Island, Italy, Sep. 2002.
[5] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 10, pp. 1183–1194, Oct. 2004.
[6] A. Eleftheriadis, "Dynamic Rate Shaping of Compressed Digital Video," Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia Univ., New York, 1995.
[7] H.-M. Hang and J.-J. Chen, "Source model for transform video code and its application, part I and II," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287–311, Apr. 1997.
[8] J.-G. Kim, Y. Wang, and S.-F. Chang, "Content-adaptive utility-based video adaptation," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, Jul. 2003.
[9] R. Mohan, J. R. Smith, and C.-S. Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 104–114, Mar. 1999.
[10] D. Mukherjee, E. Delfosse, J.-G. Kim, and Y. Wang, "Optimal adaptation decision-taking for terminal and network quality-of-service," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 454–462, Jun. 2005.
[11] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," *Adv. Neur. Inform. Process. Syst.*, vol. 12, pp. 547–553, 2000.
[12] E. C. Reed and J. S. Lim, "Optimal multidimensional bit-rate control for video communications," *IEEE Trans. Image Process.*, vol. 11, no. 8, pp. 873–885, Aug. 2002.
[13] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats," *IEEE Trans. Multimedia*, vol. 2, no. 2, pp. 101–110, Jun. 2000.
[14] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[15] A. Vetro, Y. Wang, and H. Sun, "Rate-distortion optimized video coding considering frameskip," in *Proc. Int. Conf. Image Processing (ICIP)*, Vancouver, BC, Canada, Oct. 7–10, 2001, pp. 534–537.
[16] Y. Wang, J.-G. Kim, and S.-F Chang, MPEG-4 Real Time FD-CD Transcoding Columbia Univ. DVMM Group, New York, 2003, Tech. Rep. 208-2005-2.
[17] ——, "Content-based utility function prediction for real-time MPEG-4 video transcoding," in *Proc. Int. Conf. Image Processing (ICIP)*, Sep. 14–17, 2003, vol. 1, pp. 189–192.
[18] Y. Wang, T.-T. Ng, M. van der Schaar, and S.-F. Chang, "Predicting optimal operation of MC-3DSBC multi-dimensional scalable video coding using subjective quality measurement," in *Proc. SPIE Video Communication and Image Processing (VCIP)*, 2004.
[19] O. Werner, "Requantization for transcoding of MPEG-2 intraframes," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 179–191, Feb. 1999.
[20] P. Yin, M. Wu, and B. Liu, "Video transcoding by reducing spatial resolution," in *Proc. Int. Conf. Image Processing (ICIP)*, Sep. 10–13, 2000, pp. 972–975.
[21] B. Zhang, Generalized K-harmonic Means-Boosting in Unsupervised Learning Hewlett-Packard Lab, 2000, Tech. Rep. HPL-2000-137.
[22] D. Zhong, "Segmentation, Index and Summarization of Digital Video Content," Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia Univ., New York, 2001.

Yong Wang (M'06) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1999 and 2001, respectively, and the Ph.D. degree from Columbia University, New York, all in electrical engineering.

He is currently a Senior Member Research Engineer at Motorola Multimedia Lab, Schaumburg, IL. During the summer of 2003, he was an intern at Hewlett Packard Labs, Palo Alto, CA. From 1999 to 2001, he was a visiting student at Microsoft Research Asia, Beijing, China. His research interests include video coding, adaptation, communication, and content-based analysis.

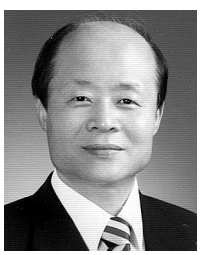


Jae-Gon Kim (M'04) received the B.S. degree in electronics engineering from Kyungpook National University, Kyungpook, Korea, in 1990, the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, in 1992 and 2005, respectively.

Since 1992, he has been a Senior Member of Research Staff in the Broadcasting Media Research Group, Electronics and Telecommunications Research Institute (ETRI), Daejon, Korea. He is currently the Team Leader of the Convergence Media Research Team. From 2001 to 2002, he was a Staff Associate at the Department of Electrical Engineering, Columbia University, New York. His research interests include scalable video coding, video adaptation, networked video, multimedia applications, and MPEG-4/7/21.

Shih-Fu Chang (M'93–SM'02–F'04) is a Professor in the Department of Electrical Engineering, Columbia University, New York. He leads Columbia University's Digital Video and Multimedia Lab (http://www.ee.columbia.edu/dvmm), conducting research in multimedia content analysis, video retrieval, multimedia authentication, and video adaptation. Systems developed by his group have been widely used, including VisualSEEk, VideoQ, WebSEEk for image/video searching, WebClip for networked video editing, and Sari for online image authentication. His group has made significant contributions to the development of MPEG-7 multimedia description schemes, and MPEG-21 Digital Item Adaptation schemes. He has been a consultant of several media technology companies. He has initiated major projects in several domains, including a digital video library in echocardiogram, a content-adaptive streaming system for sports, and a topic tracking system for multisource broadcast news video.

Dr. Chang was a Distinguished Lecturer of the IEEE Circuits and Systems Society (2001–2002) and General Co-Chair for the ACM Multimedia Conference 2000 and IEEE ICME 2004. He received a Navy Office of Naval Research (ONR) Young Investigator Award, an IBM Faculty Development Award, and a National Science Foundation (NSF) CAREER Award. His group has received the best paper or student paper awards from the IEEE, ACM, and SPIE.



Hyung-Myung Kim (S'86–M'86–SM'99) received the B.S. degree in electronics engineering form Seoul National University, Seoul, Korea, in 1974 and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1982 and 1985, respectively.

During the summer of 1997, he was on sabbatical leave as a Visiting Researcher with the Department of Electrical Engineering, The Pennsylvania State University, University Park. Currently, he is a Professor with the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejon. His research interests include digital signal/image processing, digital transmission of voice, data and image, and multidimensional system theory.

Dr. Kim was the Treasurer of the IEEE Daejon Section in 1992. He has been an Editorial Board member of *Multidimensional Systems and Signal Processing* since 1990 and an Editor of the EURASIP *Journal of Wireless Communication Networks* since 2003.