Classification-Based Multidimensional Adaptation Prediction for Scalable Video Coding Using Subjective Quality Evaluation

Yong Wang, Mihaela van der Schaar, *Senior Member, IEEE*, Shih-Fu Chang, *Fellow, IEEE*, and Alexander C. Loui, *Senior Member, IEEE*

***Abstract*—Scalable video coding offers a flexible representation for video adaptation in multiple dimensions comprising spatial detail and temporal resolution, thus providing great benefits for universal media access (UMA) applications. However, currently most of the approaches address the multidimensional adaptation (MDA) problem in an ad hoc manner. One challenging issue affecting the systematic MDA solution is the difficulty in constructing analytical models in theoretical optimization that capture the relations between video utility and MDA operations. In this paper, we propose a general classification-based prediction framework for selecting the preferred MDA operations based on subjective quality evaluation. For this purpose, we first apply domain-specific knowledge or general unsupervised clustering to construct distinct categories within which the videos share similar preferred MDA operations. Thereafter, a machine learning based method is applied where the low level content features extracted from the compressed video streams are employed to train a framework for the problem of joint signal–to-noise ratio (SNR)-temporal adaptation selection based on the motion compensated three-dimensional subband coding (MC-3DSBC) system. We conduct extensive subjective tests involving 31 subjects, 128 video clips, and formal subjective quality metrics. Statistical analysis of the experimental results confirms the excellent accuracy in using domain knowledge and content features to predict the MDA operation.**

***Index Terms*—Classification-based video adaptation, motion compensated three-dimensional subband coding (MC-3DSBC), multidimensional adaptation (MDA), subjective perceptual quality.**

I. Introduction

A. Multidimensional Adaptation (MDA)

MULTIDIMENSIONAL adaptation (MDA) provides great benefits for universal media access (UMA) by offering flexible scalability in multiple dimensions comprising spatial detail and temporal resolution. Our research has shown that MDA often results in a considerably improved user experience for multimedia applications [19]. Early efforts in video adaptations are mainly concerned with efficient transcoding of videos generated by block-based video coding systems employing discrete cosine transform (DCT) and motion compensation (MC) [1], such as MPEG-x and H.26x codecs. Popular approaches include requantization of transform coefficients [2], DCT coefficient dropping [3], frame skipping [4], and spatial resolution reduction [5]. To address the MDA issue for these codecs, several works used the rate-distortion (R-D) framework to address the rate control issue considering quantization and frame skipping in the multidimensional space [8], [9], [14].

In addition, the MDA operation selection problem has appeared in the context of scalable coding such as MPEG-4 FGS [7] and motion compensated three-dimensional subband coding (MC-3DSBC) [6]. Most systems applied heuristic-based metrics to handle the MDA case or extended the signal-level PSNR metrics by considering the human vision system (HVS) models. Recent results in [20] modeled the quality degradation from several aspects and combined them into a single evaluation that was used to guide the MDA selection. Our previous work in [24] applied *ad hoc* rules extracted from user subjective experiments to improve the tradeoff between the signal-to-noise ratio (SNR)-temporal scalability dimensions of the MPEG-4 FGS coding.

Fig. 1. Mismatch between subjective evaluation and MSE based measurement. (a) 15 f/s, 100 kb/s, $\mathrm{PSNR} = 25.98$ dB, $\mathrm{DMOS} = 1.5/5$. (b) 8 f/s, 100 kb/s, $\mathrm{PSNR} = 24.56$ dB, $\mathrm{DMOS} = 2.2/5$.

B. Analytical Modeling versus Classification-Based Approaches

The R-D framework is very useful for finding optimal adaptation in the single-dimensional case. However, several problems arise when the R-D based methods are applied to solve the MDA case. Firstly and most importantly, a mean squared error (MSE, or equivalent) based distortion measure is known to be inadequate for MDA that involves spatial and temporal scaling. Fig. 1 provides a visual illustration on this problem,

Manuscript received May 9, 2004; revised January 3, 2005. This paper was recommended by Associate Editor J. Ostermann.

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

M. van der Schaar is with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616-5294 USA (e-mail: mvanderschaar@ece.ucdavis.edu).

A. C. Loui is with the Imaging Science And Technology Laboratory, Eastman Kodak Company, Rochester, NY 14650-1816 USA (e-mail: alexander.loui@kodak.com).

Digital Object Identifier 10.1109/TCSVT.2005.854224

where the Foreman sequence was coded at 100 kb/s with different frame rate. The PSNR value averaged over the sequence gives a misleading measurement, in contrast with the degradation mean opinion score (DMOS) obtained by subjective perceptual quality evaluation. Second, even if the distortion metrics are defined, it is difficult to come up with an efficient formulation that includes various parameters involved in MDA, and describes the relations between resources and video quality. The formulation can be easily done in conventional one-dimensional adaptation (e.g., quantization), but becomes less straightforward for the MDA case. Several works, including those mentioned above, proposed models to handle the MDA scenarios. Nevertheless, due to the lack of adequate objective metrics that capture the impact on video quality under different adaptation, systematic solutions are still missing. Because of the aforementioned challenges, the R-D optimization approaches are limited in solving the MDA selection problems.

Besides the analytical modeling, another type of methods for selecting the MDA operation is the classification-based approach. In [21], we developed a content-based classification system to accelerate the generation of the utility function that characterizes the relation between the approximated subjective quality and the bit rate. In [10], a classification paradigm was proposed to choose from different coding options of FGS according to cost functions defined based on some objective model of the perceptual quality. Our previous work [15] proposed a classification framework for selecting the optimal MPEG-4 transcoding operation based on the SNR quality metric. In [23], we have also shown preliminary results in using classification-based methods to predict MDA operation matching subjective preferences. Instead of analytical derivation of the optimal adaptation that achieves the highest utility, these methods are based on the principle that videos can be mapped into distinct categories, each of which comprises videos sharing consistent adaptation behaviors. The adaptation behaviors characterize the utility, required resources, and their relations with various types of adaptations. Given a new video, compressed-domain features are used to classify the video into a previously learned class, from which the adaptation operation is predicted. These features include domain-specific knowledge such as minimum achievable bandwidth for particular codecs and MDA operations, and the low level content features describing video characteristic such as motion intensity and texture complexity. Note such a prediction paradigm is fundamentally different from the conventional optimization approach based on analytical derivation, such as Lagrangian optimization. The predicted operation may match or differ from the actual optimal operation. The performance is measured according to the percentage of times when predictions match the ground truths.

The advantages of classification-based approach are multifold. First, it is not necessary to rely on an analytical model or empirical relational curves like those used in the R-D optimization framework to relate the impact on subjective quality of the various MDA alternatives. Alternatively, only the statistical analysis of the correlations between video features and effects of adaptation in terms of resources and quality is necessary. Second, the above statistical analysis can be performed either offline or online through classification-based prediction, which can be done efficiently by lightweight feature extraction and classification. Last, videos in the same application (e.g., videoconferencing) are likely to share consistent properties and thus make accurate classification and prediction possible.

C. Contributions of the Paper

In this paper, we apply the classification-based MDA prediction methodology to a state-of-the-art scalable video coding technique, i.e., MC-3DSBC. We investigate the problem of predicting the 2-D adaptation operation combining the SNR-temporal scalability. In contrast with the objective SNR metrics used in our prior work in [15], we explicitly adopt evaluation metrics for subjective video quality. We conduct extensive subjective experiment using a large video pool (128 video clips) and a modest group of (31) subjects. Formal statistical analysis is conducted to assess the statistical significance of the experiment. Based on the experiment data, we discover distinctive patterns of subjective preferences of different SNR-temporal resolutions when adapting video under different resource (bit rate) constraints. We also find such distinctive patterns actually depend on the video content category—validating the assumption that video content is correlated with the video adaptation behaviors and thus can be used to predict the MDA operations in a classification scheme. The experimental results confirm the excellent accuracy in using compressed-domain features to predict the MDA operation matching subjective quality evaluation. In addition, we investigate the feature selection issue to identify the optimal set of content features with good balance between computational complexity and classification accuracy.

The rest of this paper is organized as follows. Section II describes the general classification-based MDA operation prediction framework. In Section III, we describe the application of such a framework to scalable video coded in MC-3DSBC. The subjective experiment setup and result analysis are described with details in Section IV. The classification-based prediction results are presented in Section V. Conclusions and future work are given in Section VI.

II. Content-Based MDA Prediction Framework

In this section, we first formulate the MDA selection as a problem of resource-constrained utility maximization. Then, we present conceptual system architecture of predicting MDA according to classification-based classification.

Consider $\mathbf{a} = (a_1, a_2, \ldots, A_{|A|}) \in \mathbf{A}$ is an MDA operation defined in the space $\mathbf{A}$, where each element a_i stands for a constituent adaptation operation and $|A|$ is the dimension of the adaptation space. In general the MDA selection problem can be formulated as

$$\tilde{\mathbf{a}} = \arg\max_{\mathbf{a}\in\mathbf{A}} U(\mathbf{a})$$
$$R(\mathbf{a}) \le R_0$$

where $U(\mathbf{a})$ is the utility of the adapted video after operation $\mathbf{a}$ is applied. As defined in [19], the utility represents the quality or user experience of the adapted video, such as peak SNR (PSNR), perceptual quality, or even high-level comprehensibility. $R(\mathbf{a})$ is the new resource requirement of the adapted

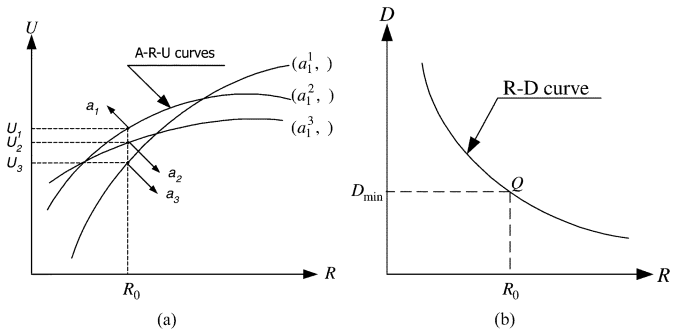

Fig. 2. (a) MDA space showing the relation between utility (U) and resource (R) when the adaptation (A) space involves two dimensions—a_1 has discrete values and a_2 has continuous values. (b) Conventional R-D curves consider the distortion as utility and often involve only one dimension of adaptation (e.g., quantization).

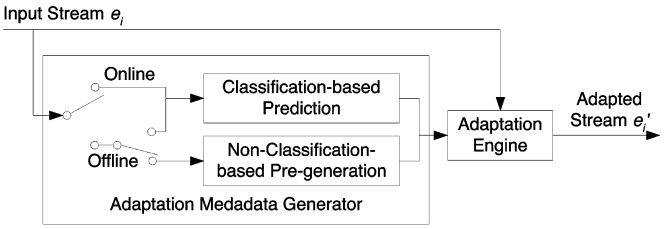

Fig. 3. General system showing the relation between the adaptation metadata generator and the actual adaptation engine.

video, and R_0 is the resource constraint implied by the user environment. Resources may include bandwidth, display resolution, power supply, or even the user's available time. One example is illustrated in Fig. 2(a), where the adaptation space consists of two dimensions—a_1 and a_2. a_1 has discrete values and a_2 has continuous values. Under the resource constraint R_0, several MDA operations are available: $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$. If analytical functions for the adaptation-resource-utility (A-R-U) exist, it is easy to choose the right MDA that maximizes the utility. As discussed above, however, usually such curves are very difficult to obtain. To solve this problem, the basic idea of our system is to "predict" such R-U relationship among different adaptation operations through statistical pattern recognition methods. This approach is based on the observation that videos with similar content characteristic share similar MDA A-R-U behaviors. Similar concepts using content-based analysis to improve rate allocation can be found in the video coding literature. For example, works in [8] use the estimation of the image frame complexity to determine the appropriate quantization step size. Nevertheless, existing works are mostly based on the R-D framework, in which an analytical model of the distortion (or utility) is needed. As a comparison, a simple R-D curve is shown in Fig. 2(b), which involves only one dimension of adaptation (e.g., quantization).

Fig. 3 shows general system architecture for video adaptation. The adaptation engine reshapes the input stream according to the resource constraints and adaptation metadata. Such metadata characterizes the relationship between feasible adaptation operations meeting the constraints and the utilities associated with each adaptation. It may also characterize how A-R-U relations vary for different types of video content. A simple representation of such metadata is a lookup table that describes all the

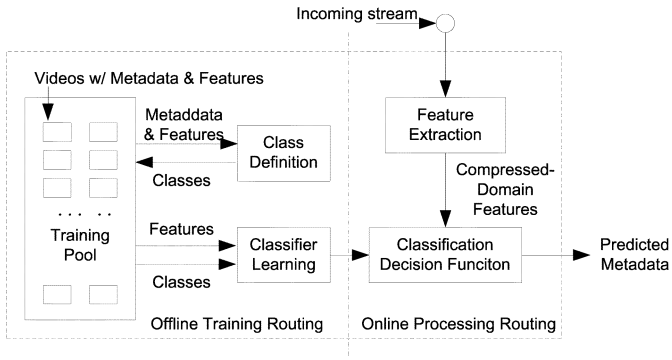

Fig. 4. Classification-based MDA prediction framework.

points of interest in Fig. 2(a). In our contributions to MPEG-21 DIA, a tool called AdaptationQoS is defined specifically for such metadata [11].

Note that in Fig. 3 several methods are shown for generating the adaptation metadata according to different scenarios. For an online application (e.g., broadcasting of live events), classification-based prediction offers a light solution suitable for real-time implementation. An alternative is to compute the utility, resource, and their relations with various adaptations for each video online, which is infeasible as subjective evaluation is needed. Another alternative is to generate the adaptation behavior metadata based on some analytical models, such as those used for modeling the R-D curves. However, as discussed earlier, adequate metrics matching the subjective perception quality for the various MDA are still missing. For offline applications (e.g., delivery of archived videos), both classification-based prediction and nonclassification-based methods are possible, although classification-based prediction methods still offer great benefits in terms of implementation efficiency.

The detailed mechanism of the proposed classification-based prediction framework is shown in Fig. 4. It can be roughly categorized to an offline training route and an online processing route. The former includes the modules for the class definition and classifier learning, and the latter mainly involves classification and prediction. For the class definition part, firstly a pool of video clips is used as the training data. For each video clip, the compressed-domain features (including domain-specific knowledge, if available, and low level content features) are extracted, so are its adaptation metadata describing the relations between MDA operations and associated utility/resource. Then, the videos are grouped into distinct categories based on domain-specific rules or domain-independent unsupervised clustering. Videos in the same class are represented by a unique class label and associated with distinctive adaptation behavior metadata. More details about video categorizing will be given in Section IV. Given the class definition and labeled training data, machine learning techniques are used to learn statistical classifiers for mapping video features to corresponding classes. The learning procedure can be as simple as applying domain knowledge directly, or through standard pattern recognition methods such as support vector machine (SVM). Such classifiers are then used in the online processing routine to classify the incoming video according to its content features, and predict the corresponding adaptation metadata for the specific video,

which will be sent to the adaptation engine for selecting the MDA operation.

The effectiveness of the above classification-based prediction approach greatly depends on whether consistent, distinctive video classes can be defined, and whether accurate classifiers can be realized. To achieve these goals, it is important to investigate what video features among a vast myriad of possible choices contribute most to the class definition and classification. We will specify the feature investigation issue including some techniques using mutual information analysis in Sections IV and V.

It is worth noting that the proposed framework is general, flexible, and not tied to any specific video codecs, adaptation techniques, or statistical learning methods. Specifically, the design and implementation of the function components used in classification-based prediction are flexible and extensible. In addition, from Fig. 4 it is clear that the online processing route involves only feature extraction and classification, both of which can be implemented efficiently. For compressed input video, some features (like coding parameters and bit rates) are readily available in the headers of the encoded bit streams, while others (like motion intensity and frame complexity) can be efficiently extracted from the compressed domain without full decoding. The offline training process may require some intensive computation, but it is outside the online prediction route and needs to be performed once only.

III. MDA Illustration for MC-3DSBC Codecs

MC-3DSBC codecs [6] perform wavelet filtering along the temporal axis. This process is known as motion-compensated temporal filtering (MCTF). Specifically, MC-3DSBC provides three primary freedoms to support MDA: *SNR scalability, spatial scalability*, and *Temporal scalability*. In this paper, we illustrate the proposed MDA mechanism to enable only SNR-temporal adaptation. This is due to the inability of the existing state-of-the-art MC-3DSBC coding schemes such as motion compensated embedded zero block coding (MC-EZBC) [12] to provide an efficient implementation of spatial scalability. The SNR and temporal adaptations can be denoted as $\mathbf{a}_m = (a_{\mathrm{SNR}}, a_{\mathrm{Temp}})$. Usually there are a set of $\mathbf{a}_m$ satisfying a given target bit rate R_0, while yielding different perceptual quality. Specifically, given a target adaptation bit rate, our system allows two dimensions of adaptation to meet the bit rate—truncation of the bit planes of the spatial subbands or change of the temporal resolution. The choices of the adaptation in the temporal dimension are discrete, and the choices of the bit plane truncation are more fine-grained. This case is exactly an instance of the M-D scenario depicted in Fig. 2(a). Given the bit rate constraint, candidate points satisfying the constraint are reduced to a subset in the 2-D SNR-temporal adaptation space. The subset of points can be conveniently indexed by discrete labels in one of the adaptation dimensions. For example, in the subsequent parts of the paper, we will index each SNR-temporal adaptation point by referring to its corresponding frame rate (i.e., full, half, quarter frame rate). To choose the best combination of SNR-temporal point from the subset, we adopt the subjective evaluation as the utility metric. Note such formulation is fundamentally different from one that uses one-dimensional adaptation only. One-dimensional adaptation refers to the case when one type of adaptation is available to meet the reduced resources. For example, in the case of rate-constrained quantization, the only dimension available for meeting a reduced target bit rate is to change the amplitude fidelity through quantization or bit plane truncation.

By using our classification-based MDA operation prediction framework, the selection of SNR-temporal adaptation operation can be transformed into a pattern classification problem. First, videos are clustered into specific categories according to their SNR-temporal adaptation behaviors. Such behaviors characterize the subjective quality and user preferences of adaptation operation under different bit rate constraints. Each category is then assigned some descriptors indicating the preferred adaptation operation under each bit rate. For any incoming video clip, its compressed-domain features are used by a trained decision function to determine its class label. Afterwards, the preferred SNR-temporal operation for a given bit rate is predicted from the descriptors of the specific category. Hence, two issues need to be addressed. First, in order to train the decision function, an observation video pool with perceptual quality evaluation for different levels of SNR-temporal resolution is required. Second, the selection process of the video features needs to be considered based on their prediction performance. In the next section, we will specify the subjective experiment that handles the first issue, and the results of the feature selection and prediction will be detailed in Section V.

IV. Subjective Quality Evaluation of Spatiotemporal Adapted Videos

As discussed in Section I, conventional quality measures such as MSE or SNR are not adequate for evaluating the video quality in comparing different MDA operations that incur videos at different spatiotemporal rates. To estimate the overall perceptual quality, subjective evaluation is needed. Lots of work in the literature [13] collected subjective evaluation results for source coding process, instead of adaptation of existing coded videos. Therefore, we launch an extensive subjective experiment aiming at an indepth understanding of the interdependence of adaptation operation and user, bandwidth, and video content characteristics.

A video pool comprising 128 clips were constructed in our experiment, including standard sequences (such as *Foreman*), test sequences used by Video Quality Experts Group (VQEG) [13], and some movies clips. All clips were about 10 s, with CIF resolution and frame rate of 30 f/s. They covered a wide range of content characteristic and thus were suitable for our goal in studying the effect of content on adaptation behavior. Also, to ensure that each clip has consistent content features, we made certain that no shot boundary existed within each clip. Although the video content may still vary within a clip, most content features were consistent in such short clips. All of the clips were coded using the MC-EZBC codec [12] with the group of pictures (GOP) size of 16 frames. The adaptation bit rates used in the experiment were $R = \{50, 100, 200, 400, 600, 1000\}$ kb/s,

covering a wide range of bandwidth frequently seen in practical applications, with an emphasis on the low bandwidth end. Each clip was adapted into versions of all different bandwidths whenever the bit rate was achievable by the codec.

A. Subjective Experiment

The subjective experiment was carried out in a quiet, separated conference room. The video clips were displayed in a 19-in Dell P991 Trinitron monitor at a resolution of 1280×960. The viewing distance was fixed at five times of the picture width. Totally 31 subjects participated in the experiment. They were undergraduate and gradate students at Columbia University from different departments. Due to the volume of the evaluation, the video pool was divided into eight groups, each with 16 distinct clips. Each video group was assessed by five subjects. Some subjects enrolled in more than one group.

We adopted the double stimulus impairment scale (DSIS) experiment recommended by the ITU-R standard [16] with minor revision. Four display windows were aligned in two rows and two columns. The left-top window displayed the unadapted reference sequence. The other three windows displayed the adapted clips. According to the MC-EZBC architecture, these three clips were adapted into the same bandwidth with full frame rate (30 f/s, without temporal adaptation), half frame rate (15 f/s) and quarter frame rate (7.5 f/s), respectively. Their display windows were randomized to avoid opinion bias. During the experiment, the reference clip was firstly played. When it was finished, the user could choose to see the adapted clips one by one and give a degradation mean opinion score (DMOS) ranging from 1 to 5, corresponding to the worst quality to the best quality based on the perceived impairment when comparing the adapted video with the reference sequence. For each clip, all of the adapted versions at different bandwidths were evaluated resulting in a score for each (clip, bandwidth) combination. The temporal order of evaluating different (clip, bandwidth) pairs was randomized to avoid any potential bias.

B. User Behavior Consistency

In addition to the 128 test clips, we included three baseline clips that were seen by all 31 subjects. We used these three clips to assess the consistency of preferences among users. The three common clips were of diverse content characteristics. Each clip was tested at six different bandwidths. For each video-bandwidth pair, each user assigned subjective scores of different temporal rates—resulting in an 18-D score vector for each user over the baseline video set. The correlation matrix of the score vector for all users was calculated. The analysis of such correlation data revealed that most of users behaved similarly with high or medium correlation, with a small number (about five users) behaving in a relatively dissimilar way. In other words, this indicates that there is a high degree of agreement in adaptation operation preferences among a great majority of users. In the subsequent analysis, we included all the scores from all 31 subjects over the 128 test clips, without attempting to filter out the five relatively dissimilar users.

C. Statistical Data Analysis

For each clip-bandwidth pair, the experiment produced three different adapted versions (full frame rate, half frame rate, and quarter frame rate), each of which received DMOS scores from five different subjects. The next step was to apply statistical analysis to assess the ranking among different adaptations and the statistical significance of such rankings.

First, the mean scores for three adaptation methods were ranked for the given (clip-bandwidth) pair, resulting in a descending ranked list of adaptations a_1, a_2, a_3, where a_i is the adaptation that has the ith rank subjective score. Then, we used a paired t -test technique to calculate the confidence score $P_{i,j}$ for the claim that adaptation a_i is preferred to a_j. Given a confidence threshold P_η, the following rule was employed to resolve any inconsistence among the pair-wise preference relations among all three adaptations

if $P_{1,2} \geq P_\eta$ and ($P_{2,3} \geq P_\eta$ or $P_{1,3} \geq P_\eta$)
  a_1 is the optimal operation;
else if $P_{1,2} < P_\eta$ and $P_{2,3} \geq P_\eta$
  a_1, a_2 tie and are optimal operations;
else all other cases
  three of the operations tie;

The threshold P_η plays an important role in determining the significance of claims about adaptation preferences. The higher threshold value we set, the more confident we can be in claiming about preferences of specific adaptations. However, since the experiment data is not unlimited, a higher confidence threshold also results in more cases of ambiguity, namely, "tie" as defined in the above procedure. Therefore, we need to find a balance between a high confidence and a low number of ties. Our experiment results indicated that $P_\eta = 0.75$ is a good tradeoff, leading to a moderate amount of ties (around 25%). Although the selection is ad hoc and the data size (five subjects per clip-bandwidth pair) may appear to be limited, we will confirm the effectiveness of such experimental approaches through a satisfactory accuracy in adaptation prediction later.

D. Prediction of Adaptation Operation

We generated the histogram of subjective preference for different frame rates at different bandwidths by counting the number of subjects preferring each different adaptation operation at specific bandwidths. In the case of tie, the counts were split equally to tied adaptations. Some examples of such histograms are shown in Fig. 5.

From the histogram, it was observed from high, medium to low bandwidth, the preferred adaptation operation shifts from full frame rate, half frame rate to quarter frame rate gradually. Such a trend is intuitive and re-confirms earlier findings.

The statistics shown in the histograms also reveals a very important piece of information—there exist distinct switching bandwidths r_{s1}, s_{s2} at which the preferred frame rate changes. This is very useful for practical applications—it provides a coarse prediction of the adaptation operation (with different

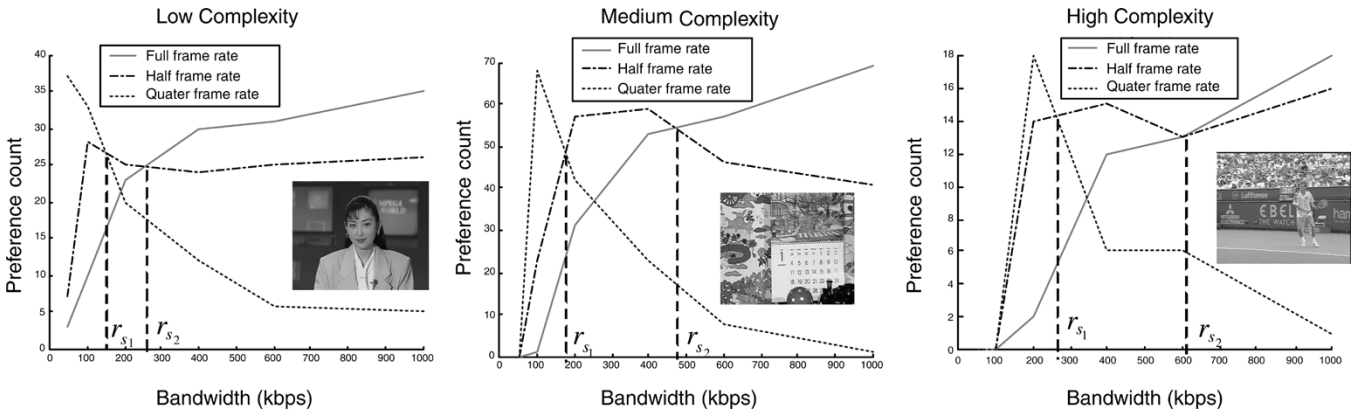

Fig. 5. Histogram of preferred frame rate for videos with different content complexity.

frame rates in this case) at any given bandwidth. Such information can be used as the adaptation metadata defined in the system architecture in Section II earlier.

E. Video Clustering

Video clustering, the process of mapping videos into distinct classes, is a fundamental issue for our framework. And there are two basic approaches: manual construction based on domain-specific knowledge, and full automatic discover by unsupervised clustering. The first approach utilizes prior knowledge (if available) about the domain to help define the classes. For example, for the MC-EZBC codec, we observe that the Minimal Achievable Bandwidth r_{MAB} of each video stream is a good indicator of the video content complexity and thus can be used as good criteria for defining video categories. r_{MAB} is defined as the minimal bandwidth achievable by any possible adaptation. It corresponds naturally to the intrinsic spatio complexity and motion activity of the clip. In our experiment, r_{MAB} had three distinct values: 50, 100, and 200 kb/s. Therefore, the clips in the video pool were labeled using three corresponding categories with low, medium and high content complexity, respectively. Each category of videos has its unique patterns of adaptation preference, as illustrated by the preference histograms in Fig. 5.

We can see a clear trend that when the video content complexity increases, the switching bandwidths also shift to the higher end. This is quite intuitive: more complex videos need more bits for spatial details before a higher frame rate is needed. The above finding is significant, laying the foundation for predicting the MDA operation based on the content features.

r_{MAB} is a good domain-specific knowledge that can be used to categorize the videos, and it can be easily obtained from parsing the MC-EZBC bitstream. However, such knowledge depends on the coding mechanisms and is not always reliable for generic codec such as MPEG-4 and MC scalable system. Our previous work in [15] reported some results where such simple rule is not applicable. In this situation our prediction framework utilizes automatic unsupervised clustering to discover video categories without any user supervision. To evaluate the usability of such method, herein we also apply an entropy-based unsupervised clustering method, called COOLCAT [22], to discover the video clusters. The cluster number was set to 6 empirically based on the cross-validation criterion. The clustering process was carried out in a feature space, in which the adaptation behavior of each video is represented by a set of features including the preferred frame rates at different bandwidths. Due to the space limit, readers are referred to our prior work in video category discovery [15] and the statistical clustering tool [22] for the detailed processes used in unsupervised clustering. The performance of such unsupervised clustering techniques, however, are compared against other alternatives in Section V.

V. Classification-Based Prediction of MDA Operation

Using the subjective quality evaluation data obtained through the extensive experiments, we next apply pattern classification techniques to develop classifiers and predict preferred MDA operation in a real time scenario as introduced in Section II. The classification problem is formulated as the following. Given the features extracted from each input video, classify the video to one of the classes defined in the previous subsections. Following the discussion in Section IV-E, although classes defined upon r_{MAB} can be directly achieved, such direct class information may not always be available during real time processing. Therefore, we also demonstrate the feasibility of using content feature to achieve the classification.

A. Content Feature Selection

In the same way that subjective video quality is influenced by the HVS, the criterion of content feature selection should be associated with the characteristic of HVS. Though still not fully understood, HVS can be roughly categorized into the spatial mechanism and temporal mechanism[1] [18], where the former accounts for HSV sensitivity to the characteristics of spatial variations, especially different spatial frequency signals, and the latter models the masking and response to temporal phenomena, especially different temporal frequency signals. Therefore, our feature set consists of attributes related to the spatial and temporal characteristics. Furthermore, since the video adaptation scenario considers only pre-encoded video as input, we focused on the features that can be readily extracted from the MC-EZBC encoded format.

[1]In the temporal dimension, the HVS also involves the component of motion tracking. But it is difficult to extract content feature approximating this component.

Given input videos coded in the MC-EZBC format, our raw content feature candidates consisted of about 1200 variables, including the block size, block type, motion magnitude, motion phase, residual energies, etc., each computed from multiple frames within the clip due to the spatial-temporal interleaving used in MC-EZBC. It is not surprising to confirm through simulations that adoption of the whole feature set without selection indeed results in poor prediction accuracy because adding noisy irrelevant features usually hurts the classification performance.

Based on HVS mechanisms discussed above, we include the features related to the spatial texture complexity and the temporal motion intensity. These two attributes reflect the spatial and temporal frequency details, which shape the contrast sensitivity functions (CSF) of HVS. For MC-EZBC, the most important set of features are related to three components: 1) the variable block size distribution. MC-EZBC employs hierarchical variable block size matching during the motion estimation, and therefore the variable block size is well associated with spatiotemporal details; 2) motion magnitude, indicating the motion intensity between two temporal frames; and 3) residual energies, indicating the squared coefficient magnitude after the motion-compensated spatiotemporal decomposition.

Therefore, the following two steps were applied during feature selection. First, only the features belonging to the three categories above were kept and the remaining excluded. Furthermore, the kept content features were merged by summing up the ones within the same category and the same spatiotemporal subbands. After this step, 86 combined features were kept. Specifically, these features were: 1) 20 variables describing the block size histogram; 2) 36 variables describing the motion magnitude histogram; and 3) 30 variables describing the residual energies located in different spatial subbands.

In order to further simplify the classification process, we applied mutual information feature selection (MIFS) [17] method to select a subset of features from the above 86 features. Specifically, if we want to select K features from the original feature space F_0, the following MIFS procedure was conducted to select one feature f_k in each iteration (k)

$$f_k = \arg\max_{f \in F_k} \left\{ I(C, f) - \beta \sum_{s \in S_k} I(f, s) \right\}$$

where C is the video class labels, F_k and S_k are the remaining feature set and selected feature set at iteration k, respectively, (note $F_k + S_k = F_0$), I is the mutual information (MI) between two random variables, and β is a weight to regulate the relative importance of the MI between the candidate feature and the selected features with respect to the MI between the candidate feature and the output class. The higher β is, the more the algorithm penalizes the use of correlated features. If β is set to 0, the algorithm selects individual features that have maximal mutual information with the class labels, without considering the redundancy among the features. Essentially, the above process selects the feature that adds most new information about the class given the features that have been chosen already.

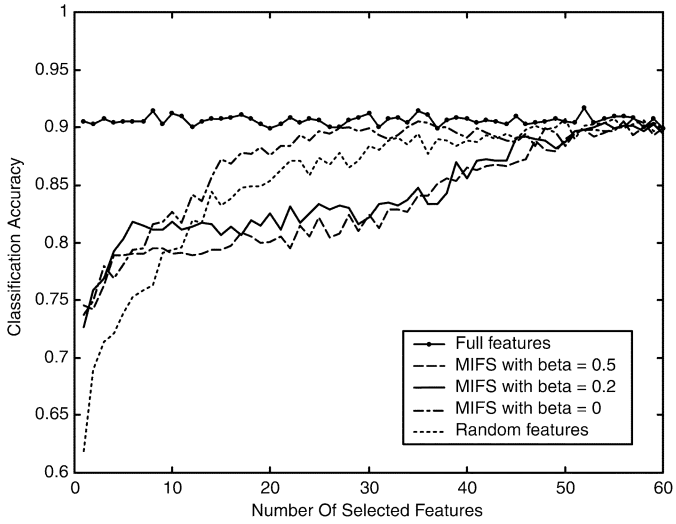

Fig. 6. Classification performance.

B. Classification—Category Prediction

We analyze the effectiveness of video classification performance and preferred adaptation operation prediction in this section. We apply supervised machine learning techniques to develop classifiers. The training data consist of samples of video clips, each represented by the features extracted from the clip. The learned classifier takes the extracted features for each new video as input, and then predicts the class that the input video most likely belongs to. Once the class is predicted, adaptation preference information of each class (as shown in Fig. 5) is used to predict the adaptation operation under different bandwidth conditions.

In our experiment, each observation was extracted from a subclip with one-GOP length, resulting in a total of 2275 observations. Each subclip carried the same category label as that of the source video clip. The results below were based on the average over ten runs. In each run, a cross-validation scheme was employed, where 80% of the observations were randomly chosen as the training pool and the remaining the testing pool. We adopted the support vector machine (SVM) as our classification technique. We set the SVM parameters radial basis function (RBF) kernel $\gamma = 2$ and nonseparable penalty $C = 100$.

Fig. 6 shows the performance of the classification based on the video categories defined using r_{MAB} (the partition based on the unsupervised clustering gave similar results). The MIFS with different β values are compared with the results using the full set (i.e., all the 86 features) and a subset randomly chosen from the 86 features. Several interesting observations are found. First, the full set of 86 features yields satisfactory accuracy (above 90%). As a comparison (results not listed here), using the set of about 1200 raw features resulted in accuracy below 70%. Second, it is clear to see the impact of β on the performance. When the number of selected features is small, MIFS works better than random and the results using nonzero β values are slightly better than that with $\beta = 0$. When more features are selected, the superiority of MIFS over random degrades. Moreover, except for $\beta = 0$, random feature selection beats MIFS after certain point (e.g., for $\beta = 0.5$ after

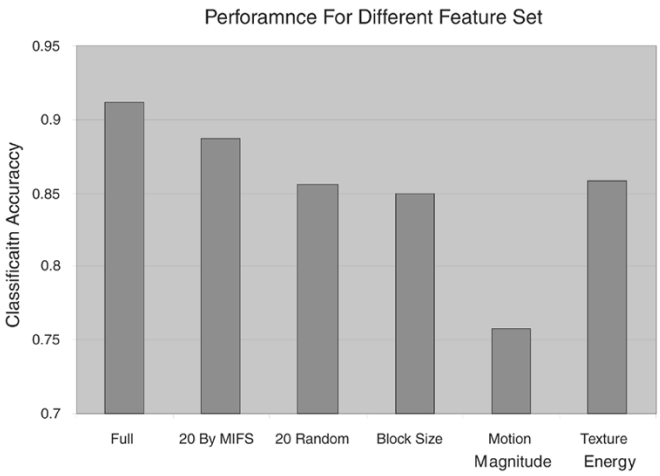

Fig. 7. Performance comparison among different feature sets.

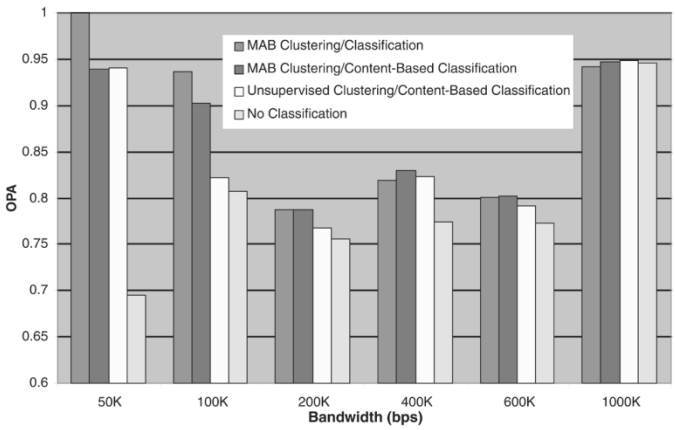

Fig. 8. Operation prediction accuracy.

about ten features). This is because β penalizes the case when selected features have high correlations. It remains to be an interesting research issue how to automatically select a suitable β value. Last, by selecting up to 20 features through MIFS, the performances reach a saturation platform ($\sim 88\%$). A reduced set of features will improve the efficiency of the algorithms.

Fig. 7 further summarizes the classification performance using different sets of features. Six results are compared: the full set of 86 features, 20 features selected by MIFS with $\beta = 0$, 20 random features (randomly selected in each run from the whole set of 86 features), 20 block size features, 36 motion magnitude features, and 30 residual energy features. This comparison helps us better understand the behavior of different content characteristic. Each component has its own contribution—among the reduced feature sets, the 20 features by MIFS work best, including 10 for the block size, 2 for motion magnitude, and 8 for residual energies. Block size and residual energy reflect both motion and texture information and therefore outperform the motion magnitude features.

C. Classification—Adaptation Operation Prediction

The category classification is not our ultimate goal. Our aim is to use the prediction result to guide selection of the MDA operation. Within each category, we can obtain the category-specific preference histogram (see Fig. 5). Given the histogram, a straightforward prediction method is to choose the operation that is ranked best most frequently based on the preference scores given by human subjects. We measure the operation prediction accuracy (OPA) in the following way:

$$\mathrm{OPA}_r = \frac{\text{Number of correct prediction}}{\text{Total Number of observations}}.$$

A prediction is considered correct when the predicted operation matches the actual preferred operation or one of the preferred operations in a tie. Fig. 8 is the result of OPA over different bandwidths. As a comparison, four different approaches are analyzed: MAB Clustering/Classification used r_{MAB} in both clustering and classification routine (i.e., a pure domain knowledge based approach); MAB Clustering/Content-based Classification used categories defined through r_{MAB} and applied content features for classification; Unsupervised Clustering/Content-based Classification used unsupervised clustering for class definition and content-based classification; and No Classification predicted the operations using the preference histogram over the entire video pool. A notable improvement gain (up to 30%) can be observed by applying classification-based prediction, especially at the low bandwidth end where practical UMA applications focus most. Among different classification approaches, MAB Clustering/Classification outperform other methods. This is reasonable as the MAB categories are defined based on domain-specific knowledge and each new video clip can be classified using the r_{MAB} value without errors. In comparison, two content-based classification methods perform almost as well as the MAB classification method at the high-bandwidth region (200 kb/s and above), but not at the low-bandwidth region (50 and 100 kb/s). Between the two content-based methods, the performances are quite close at different bandwidths, except 100 kb/s where the approach using domain-specific knowledge is better. It is also clear that for different bandwidths, the prediction performance varies, reaching the lowest in the medium bandwidth range (200–600 kb/s). This phenomenon comes from the fact that at mid bandwidths human subjects do not show consistent preferences to specific dimensions among different spatiotemporal scales, resulting in a large variance in the subjective scores. Fig. 9 shows the entropy estimation of the preferred operation rankings over different bandwidths. Such estimates in some degree can be considered as the measurement of the prediction difficulty. Actually, the performance of our proposed method matches the entropy measure very well, while the approach without classification cannot. To some extent, this also validates the approach of the content-based prediction.

D. Computational Complexity Analysis

Computational complexity of the proposed system is very important for a real time application scenario. Because the MC-EZBC codec we used was not a real-time implementation, we were not able to provide the real time benchmark data. However, all of the computation processes in our system can be easily verified to be lightweight. As shown in Fig. 4 the main costs in our system include feature extraction and classification. The classification process is very efficient. For example, SVM classification only needs to calculate the kernel function and dot

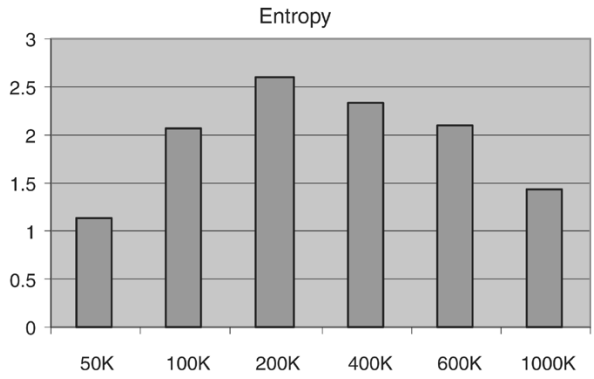

Fig. 9. Entropy of the distributions of preferred operation.

product between the content features and a sparse set of support vectors. For feature extraction, r_{MAB} can be easily retrieved by parsing the encoded bit stream. In the methods using low-level content features, partial bit stream decoding is needed to obtain motion vectors, block size and wavelet coefficients, plus some minor extra calculation such as histogram counting (for motion magnitude and block size) and subband energy calculation (for residual energy). The combination of all these computations is still much lighter than the complexity of a regular decoder (because a much more complex process, MC, is not needed). Considering video decoders can be implemented on most platforms with a real-time performance, it is reasonable to claim that our system can be implemented in a real-time fashion. In some special cases, the content features can even be computed at the encoder and transmitted to the adaptation decision module as side information, thus at no additional cost.

VI. Conclusion

In this paper, we address the issue of MDA operation selection matching subjective quality evaluation by generalizing a classification-based prediction framework. In this framework, instead of using analytical modeling or exhaustive computation to characterize the relations among adaptation, quality, and resource, a machine learning based method is applied where the compressed-domain features extracted from the video streams are used to automatically predict the MDA operation through a statistical classifier. Contrary to most of prior works, we conducted large-scale subjective studies to evaluate the perceptual quality of videos adapted at different spatiotemporal scales. Rigorous methods were applied to assess the statistical significance of the experimental data, select the optimal subset of features, and understand the contributions of individual components of features. To explore the merits of the latest multidimensional scalable video coding techniques, we tested the usability of the framework to an MC-3DSBC system.

The experiment results indicate that our proposed method can effectively reveal the relationship between the content characteristic and the MDA behavior, and therefore accurately predict the adaptation operation with accuracy from 77% to 95% over different bandwidth. To the best of our knowledge, this is the first work investigating the relations between the preferred spatiotemporal adaptation operation and the content characteristics, using the subjective quality evaluation metric.

Acknowledgment

The authors would like to thank P. Chen for his kind help in providing the codes of MC-EZBC; T.-T. Ng, G. Kim, and D. S. Turaga for their valuable discussion and contribution during early part of the work; J. Kender and Y. Liu for their discussion and providing some results in feature selection. They also thank the anonymous reviewers for their careful reviews and valuable comments.

References

[1] A. Vetro, C. Christopoulos, and H. Sun, "Video transcoding architectures and techniques: An overview," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 18–29, Mar. 2003.

[2] O. Werner, "Requantization for transcoding of MPEG-2 intraframes," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 179–191, Feb. 1999.

[3] A. Eleftheriadis, "Dynamic rate shaping of compressed digital video," Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia Univ., New York, 1995.

[4] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatiotemporal resolutions and different encoding formats," *IEEE Trans. Multimedia*, vol. 2, no. 2, pp. 101–110, Jun. 2000.

[5] P. Yin, M. Wu, and B. Liu, "Video transcoding by reducing spatial resolution," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, 2000, pp. 972–975.

[6] S.-J. Choi and J. W. Woods, "Motion-Compensated 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[7] H. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 53–68, Mar. 2001.

[8] A. Vetro, Y. Wang, and H. Sun, "Rate-distortion optimized video coding considering frameskip," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Rochester, NY, Oct. 2002, pp. 534–537.

[9] 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.

[10] B.-F. Hung and C.-L. Huang, "Content-based FGS coding mode determination for video streaming over wireless networks, selected areas in communications," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1595–1603, Dec. 2003.

[11] 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/ 2004.

[12] P. Chen and J. W. Woods, "Bi-directional MC-EZBC with lifting implementation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 10, pp. 1183–1194, Oct. 2004.

[13] A. Rohaly *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," ITU-T Standards Contribution COM 9-80-E, 2000.

[14] J.-J. Chen and H.-M. Hang, "Source model of transform video coder and its application—Part II: Variable frame rate coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 299–311, Apr. 1997.

[15] Y. Wang, J.-G. Kim, and S.-F. Chang, "Content-based utility function prediction for real-time MPEG-4 transcoding," in *Proc. ICIP*, Barcelona, Spain, Sep. 2003, pp. 189–192.

[16] *Methodology for the subjective assessment of the quality of television pictures*, Recommendation ITU-R BT.500-10, 2000.

[17] R. Battiti, "Using mutual information for selecting features in supervised neural-net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[18] C. J. van den Branden Lambrecht, "Perceptual models and architectures for video coding applications," Ph.D. dissertation, Ecole Polytechnique Federale de Lousanne EPFL, Lausanne, Switzerland, 1996.

[19] S.-F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proc. IEEE*, vol. 93, pp. 148–158, 2005.

[20] E. Akyol, A. M. Tekalp, and M. R. Civanlar, "Optimum scaling operator selection in scalable video coding," presented at the Picture Coding Symp., San Francisco, CA, Dec. 2004.

[21] P. Bocheck, A. T. Campbell, S.-F. Chang, and R.-F. Liao, "Content-Aware network adaptation for MPEG-4," presented at the ACM Int. Workshop n Network and Operating Systems Support for Digital Audio and Video, Jun. 1999.

[22] D. Barbar'a, J. Couto, and Y. Li, "COOLCAT: An entropy-based algorithm for categorical clustering," in *Proc. Conf. Information and Knowledge Management (CIKM)*, McLean, VA, 2002, pp. 582–589.

[23] 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 Communications and Image Processing (VCIP)*, Jan., pp. 529–542.

[24] R. K. Rajendran, M. van der Schaar, and S.-F. Chang, " FGS+: Optimizing the joint SNR-temporal video quality in MPEG-4 fine grained salable coding," in *Proc. ISCAS*, May 2002, pp. I-445–I-448.

Yong Wang received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1999 and 2001, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering at Columbia University, New York, supervised by Professor S.-F. Chang.

From 2001, he has been as a research assistant in the digital video and multimedia (DVMM) group at Columbia University. During the summer of 2003, he was an intern working at HP 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.

Mihaela van der Schaar (SM'04) received the M.Sc. and Ph.D. degrees in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands.

She is currently an Assistant Professor in the Electrical and Computer Engineering Department, University of California, Davis. Between 1996 and June 2003, she was a Senior Member of Research Staff at Philips Research, both in The Netherlands and the U.S., where she led a team of researchers working on scalable video coding, networking, and streaming algorithms and architectures. From January to September 2003, she was also an Adjunct Assistant Professor at Columbia University, New York. Since 1999, she has been an active participant to the MPEG-4 standard, for which she received an ISO recognition award. She is currently chairing the MPEG Ad-Hoc group on Scalable Video Coding, and is also Co-Chairing the Ad-Hoc group on Multimedia Test-bed. She has coauthored more than 90 book chapters, conference and journal papers in this field and holds 11 patents and several more pending.

Dr. van der Schaar is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, was anAssociate Editor of the *SPIE Electronic Imaging Journal*, and a Guest Editor of the *EURASIP* Special Issue on Multimedia over IP and Wireless Networks. She was elected Member of the Technical Committee on Multimedia Signal Processing of the IEEE Signal Processing Society. She has also chaired and organized many conference sessions in this area and was the General Chair of the Picture Coding Symposium 2004. In 2004, she received the NSF Career Award.

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 Laboratory, 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. 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. 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.

Dr. Chang's group has received best paper or student paper awards from the IEEE, ACM, and SPIE. He is 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, and NSF CAREER Award. He served as a General Co-Chair for ACM Multimedia Conference 2000 and IEEE ICME 2004.

Alexander C. Loui (SM'97) received the B.A.Sc. (honors), M.A.Sc., and Ph.D. degrees, all in electrical engineering, from the University of Toronto, Toronto, ON, Canada.

After receiving the Ph.D. degree in 1990, he joined Bellcore (now Telcordia Technologies), Red Bank, NJ, as Member of Technical Staff working on audiovisual compression technologies including MPEG and H.261, multipoint video conferencing systems, and the transport of multimedia over IP-based networks. He joined the Network Imaging Technology Center of Eastman Kodak Company, Rochester, NY, in 1996. Since 1998, be has been a Technical Lead of Multimedia Imaging at Kodak Imaging Research Labs. He has been an Adjunct Professor with the ECE Department, University of Toronto, since 1999. His current research interests include algorithm development and system research for multimedia content analysis, organization, authoring, and delivery. He has published many refereed papers and authored a number of U.S. patents in image/video processing, auto-albuming, and multimedia authoring technology.

Dr. Loui has been a member of ISO/IEC MPEG, ITU-T Study Group 15, Experts Group for Audiovisual Coding and Systems, and NCITS/L3.1. He is an associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He is a member of the IEEE Circuits and Systems Society Technical Committee on Multimedia Systems and Applications, and a member of ACM. He has been a session organizer, and served as session chair and member of Program Committee in many technical conferences including IEEE ICME, ICIP, ISCAS, and the IEEE Pacific-Rim Conference on Multimedia. He is also an associate editor of the SPIE/IS&T *Journal of Electronic Imaging*. He received an award for leadership and innovation in research and development of digital storytelling and auto-albuming technology from Kodak in 1998.