Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 10236, 11 pages
doi:10.1155/2007/10236

Research Article

Content-Aware Scalability-Type Selection for Rate Adaptation of Scalable Video

Emrah Akyol,[1] A. Murat Tekalp,[2] and M. Reha Civanlar[3]

[1] *Departmet of Electrical Engineering, Henry Samuel School of Engineering and Applied Science, University of California, P.O. Box 951594, Los Angeles, CA 90095-1594, USA*
[2] *Department of Electrical and Computer Engineering, College of Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey*
[3] *DoCoMo USA Labs, Palo Alto, CA 94304-1201, USA*

Received 4 October 2006; Revised 31 December 2006; Accepted 14 February 2007

Recommended by Chia-Wen Lin

Scalable video coders provide different scaling options, such as temporal, spatial, and SNR scalabilities, where rate reduction by discarding enhancement layers of different scalability-type results in different kinds and/or levels of visual distortion depend on the content and bitrate. This dependency between scalability type, video content, and bitrate is not well investigated in the literature. To this effect, we first propose an objective function that quantifies flatness, blockiness, blurriness, and temporal jerkiness artifacts caused by rate reduction by spatial size, frame rate, and quantization parameter scaling. Next, the weights of this objective function are determined for different content (shot) types and different bitrates using a training procedure with subjective evaluation. Finally, a method is proposed for choosing the best scaling type for each temporal segment that results in minimum visual distortion according to this objective function given the content type of temporal segments. Two subjective tests have been performed to validate the proposed procedure for content-aware selection of the best scalability type on soccer videos. Soccer videos scaled from 600 kbps to 100 kbps by the proposed content-aware selection of scalability type have been found visually superior to those that are scaled using a single scalability option over the whole sequence.

1. INTRODUCTION

Scalable video coding has gained renewed interest since it has been shown [1, 2] that it can achieve compression efficiency that is close to that of H.264/AVC [3] while providing a flexible adaptation to time-varying network conditions and heterogeneous receiver capabilities. Scalable video coding methods can be clustered into two groups according to the spatial transforms they utilize, block-based and wavelet-based coders. All scalable video coders enable post-encoding flexible adaptation of video rate through signal-to-noise ratio (SNR), temporal, and/or spatial scalability [1, 2]. They employ motion-compensated temporal filtering (flexible temporal predictions, such as hierarchical B pictures in block-based scalable coders and open-loop MCTF in wavelet coders) to provide temporal scalability, followed by a spatial transform (wavelet or block transform) as shown in Figure 1. Spatial scalability can be provided by compression of low resolution with prediction among layers in block-based coders, where wavelet transform inherently provides spatial scalability in wavelet coders. All transform coefficients can then be encoded using an embedded entropy coder to obtain SNR scalability. Alternatively, SNR scalability can be achieved by requantization. The scalable video compression standard, SVC [2], is based on block-based scalable coding methods. However, the problem analyzed in this paper is common to all scalable video coding methods and the proposed solution is applicable to any scalable video coder including SVC. A survey of recent developments in scalable video coding can be found in [1] and further details on the scalable video coding standardization can be found in [2].

Rate reduction by discarding enhancement layers of different scalability types generally results in different types of visual distortion on the decoded video depending on the rate and content [4–7]. Hence, in many cases, the scalability type should be adapted to content type of different temporal segments of the video for the best visual results. There are only a limited number of works that investigate the dependency between scalability type, video content, and rate, and that present objective methods for scalability-type selection

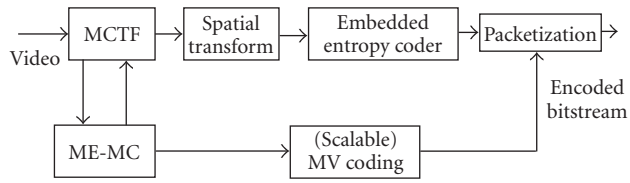


Figure 1: General structure of an MCTF-based fully scalable video coder.

[4–7]. In [4], authors investigate optimal frame rate selection for MPEG-4 fine granular scalability (FGS), where they conduct subjective tests to derive an empirical rule, based on the PSNR. A metric for the optimal ratio of spatial and temporal information has been defined in [5] and compared with a threshold to select between the spatial and temporal operators. Optimal tradeoff between SNR and temporal scalability is addressed in [6] using some content-based features, where a machine learning algorithm has been employed to match content features with the preferred scaling option. A similar approach is followed in [7] where content-based features have been used to select one of MPEG-4 FGS modes based on an objective distortion metric defined in [8]. Other works on adaptation of video to available bandwidth by spatial and/or temporal resolution adjustment include those using nonscalable video coders [9, 10] or transcoding [11, 12]. In [9], optimal rate adaptation is studied by varying spatial resolution, frame rate, and quantization step size using integer programming. In [10], optimum frame rate and quantization parameter selection to minimize the mean square error (MSE) are presented with rate-distortion modeling and frame skip. In [11], a content-based prediction system to automatically select the optimal frame rate for MC-DCT-coded video transcoding based on the PSNR is proposed. In [12], the MSE distortion is used for rate-distortion modeling of multidimensional transcoding.

It is well known that visual distortions cannot always be measured meaningfully in terms of MSE [13]. An example confirming this observation is shown in Figure 2, where discarding SNR enhancement layer(s) results in lower MSE (higher PSNR) value, but is visually inferior to discarding spatial enhancement layer(s) at the same base layer bitrate. Hence, although MSE may be a good measure of distortions caused by SNR scaling, visual distortions due to spatial and temporal scalings (spatial-and-temporal-frequency-sensitivity related distortions) cannot be measured accurately with the MSE [13]. Objective measures can be grouped as (i) those based on a model of low-level visual processing in the retina and (ii) those which quantify compression artifacts [14]. An early example of the latter type is [15], where visual distortion for MPEG-2 coded videos is measured considering blockiness and a perceptual model. In [16], subjective evaluation of videos coded with several coders, including scalable coders, is investigated and significant correlation is found with distortion-based objective metrics. We review examples of latter-type metrics in Section 2.

In this work, we study the relationship between scalability type, content type, and bitrate based on the assumption that a single scalability choice may not fit the entire video content well [4, 6]. We define an objective function based on specific visual distortion measures, whose weights are tuned to different shot content types at a given bitrate in order to choose the best scalability type for each temporal segment. The weights of the objective function vary according to the shot content type, since the dominant distortion may depend on the content (e.g., flatness may be more objectionable in far shots with low motion, whereas jerkiness may be more objectionable in shots with high motion). This requires video analysis to be performed for shot/segment boundary detection and shot-/segment-type classification. There is a significant amount of work reported on automatic video analysis [17–21], which is beyond the scope of this paper. Recently, specific content analysis methods have been developed for sports video [19]. Most of these methods can be implemented in real time or near real time. Content-aware video coding and streaming techniques have been proposed in [22], where different shots have been assigned different coding parameters depending on the content and user preferences.

This paper offers the following novelties compared to the state of the art.

(a) We propose an objective function for scalability-type selection, and present a procedure to adapt the coefficients of the objective function to content-type and bitrate. Previous works, such as [6], are experimental, which can determine the optimal operator but not the cost associated with choosing another operator. Hence, they cannot be used in an optimization framework (such as rate-distortion optimization or rate-distortion-complexity adaptation).

(b) We propose a procedure for *automatic* selection of the best scalability type, among all of temporal, spatial, and SNR scalabilities, for each temporal segment of a video according to content, at a *given bitrate*. Other works consider only limited scalability options, for example, [6] considers *only* SNR and temporal scaling, but not spatial scaling.

A block diagram of the proposed system is shown in Figure 3, where a fully embedded scalable video coder is employed. Bitstreams formed according to different combinations of scalability options are then extracted and decoded. Low-resolution videos are interpolated to the original resolution. Finally, the above objective cost function is evaluated for each combination, and the option that results in the minimum cost function is selected. The paper is organized as follows. We discuss distortion measures in Section 2. Section 3 presents the choice of scaling options (SNR, temporal, spatial, and their combinations) and the problem formulation. Two subjective tests and statistical analyses of the results are described in Section 4. Conclusions are presented in Section 5.

2. VIDEO QUALITY MEASURES

It is well known that different scalability options yield different types of distortions [14]. For example, at low rates,

(a) SNR scaled, PSNR = 29.19 at 100 kbps (b) Spatially scaled, PSNR = 27.79 at 100 kbps

Figure 2: Although the SNR (a) scaled video is visually poorer, its PSNR is higher than the (b) spatially scaled (and interpolated to original size) video.

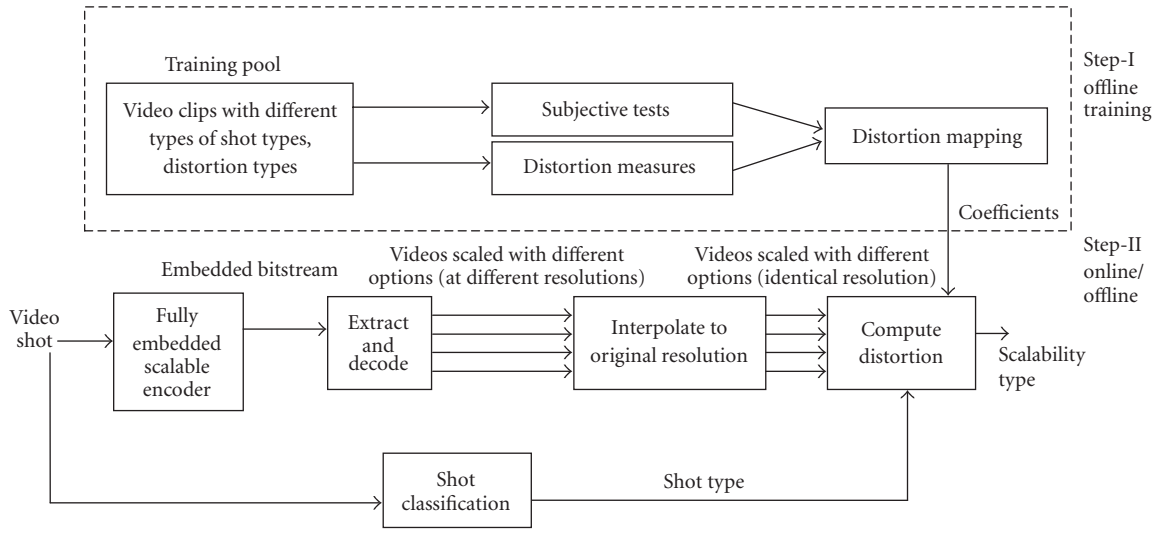


Figure 3: Overview of the proposed algorithm for scaling-type selection.

SNR scalability results in blockiness and flatness due to block motion compensation (see Figure 4) and high quantization parameter (Figure 2(a)). On the other hand, spatial scalability results in blurriness due to spatial lowpass filtering in 2D wavelet coding (Figure 2(b)), and temporal scalability results in motion jerkiness. Because the PSNR is inadequate to capture all these distortions or distinguish between them [13], we need to employ visual quality measures [23]. *It is not the objective of this research to develop new video quality metrics or verify them*. We only employ such available metrics to develop a measure for scalability-type selection; the general framework is applicable with any choice of distortion functions as long as training is performed with the same set of functions. The following recently published measures (with small modifications due to the features of the codec) have been used in this work, although the proposed framework does not rely on any specific measures.

2.1. Blurriness measure

Blurriness is defined in terms of change in the edge width [24]. Major vertical and horizontal edges are found by using the Canny operator [25], and the width of these edges is computed. The blurriness metric is then given by

$$D_{\text{blur}} = \frac{\sum_i \left(\text{Width}_d(i) - \text{Width}_{\text{org}}(i)\right)}{\sum_i \text{Width}_{\text{org}}(i)}, \tag{1}$$

where $\text{Width}_{\text{org}}(i)$ and $\text{Width}_d(i)$ denote the width of the ith edge on the original (reference) and the width of the decoded (distorted) frame, respectively. Edges in the still regions of frames are taken into consideration as done in [15].

2.2. Flatness measure

A new objective measure for flatness-based on local variance of relatively smooth regions (regions where there are no

Figure 4: An example of blockiness distortion, coded with SNR scaling at 100 kbps.

significant edges). First, major edges using the Canny edge operator [25] are found, and the local variance of 4×4 blocks that contain no significant edges is computed. The flatness measure is then defined as

$$\mathrm{D_{flat}} = \begin{cases} \dfrac{\sum_i \left[\sigma^2_{\mathrm{org}}(i) - \sigma^2_d(i)\right]}{\sum_i \sigma^2_{\mathrm{org}}(i)} & \text{if } \sigma^2_{\mathrm{org}} \le \mathrm{T}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\sigma^2_{\mathrm{org}}(i)$ and $\sigma^2_d(i)$ denote the variance of 4×4 blocks on original (reference) and decoded (distorted) frames, respectively and T is a threshold value which is experimentally determined (any value between 70 and 80 was satisfactory for the threshold in our experiments). The hard-limiting operation provides spatial masking of quantization noise in high texture areas.

2.3. Blockiness measure

Several blockiness measures exist to assist PSNR in the evaluation of compression artifacts under the assumption that the block boundaries are known a priori [15, 16, 26]. For example, the blockiness metric proposed in [26] is defined as the sum of the differences along predefined edges scaled by the texture near that area. When using overlapped block motion compensation and/or variable-size blocks, location and size of the blocky edges are no longer fixed. To this effect, first the locations of the blockiness artifacts should be found. Horizontal and vertical edges detected in the decoded frame, which do not exist in the original frame, are treated as blockiness artifacts. Canny edge operator [25] is used to find such edges. Any edge pixels that do not form vertical or horizontal lines are eliminated. Alternatively, block locations can be determined after decoding the bitstream. A measure of texture near the edge location, which is included to consider spatial masking, is defined as

$$\begin{aligned} \mathrm{TM_{hor}}(i) = &\sum_{m=1}^{3}\sum_{k=1}^{L} \left| f(i-m,k) - f(i-m+1,k) \right| \\ &+ \sum_{m=1}^{3}\sum_{k=1}^{L} \left| f(i+m,k) - f(i+m+1,k) \right|, \end{aligned} \tag{3}$$

where, f denotes the frame of interest, and L is length of the straight edge, where we set $L = 16$. The blockiness of the ith horizontal edge can be defined as

$$\begin{aligned} &\mathrm{Block_{hor}}(i) \\ &= \frac{\sum_{k=1}^{k=L} \left| f(i,k) - f(i-1,k) \right|}{1.5 \cdot \mathrm{TM_{hor}}(i) + \sum_{k=1}^{k=L} \left| f(i,k) - f(i-1,k) \right|}. \end{aligned} \tag{4}$$

The blockiness measure for that frame containing M edges, $\mathrm{BM_{hor}}$, is defined as $\mathrm{BM_{hor}} = \sum_{i=1}^{M} \mathrm{Block_{hor}}(i)$.

Blockiness measure for vertical straight edges $\mathrm{BM_{vert}}$ can be defined similarly. Finally, total blockiness metric $\mathrm{D_{block}}$ is defined as

$$\mathrm{D_{block}} = \mathrm{BM_{hor}} + \mathrm{BM_{vert}}. \tag{5}$$

2.4. Jerkiness measure

In order to evaluate the difference between temporal jerkiness of the decoded and original videos with full frame rate, we compute the sum of magnitudes of differences of motion vectors over all 16×16 blocks at each frame (without considering the replicated (interpolated) frames),

$$\mathrm{D_{jerk}} = \frac{\sum_i \left| \mathrm{MV}_d(i) - \mathrm{MV_{org}}(i) \right|}{N}, \tag{6}$$

where $\mathrm{MV_{org}}(i)$, $\mathrm{MV}_d(i)$, and N denote the ith element of the motion vector of the original 16×16 block, motion vector of the 16×16 block i, and the number of 16×16 blocks in one frame, respectively. Specifically, we perform motion estimation on the original video and denote the motion vectors as $\mathrm{MV_{org}}(i)$ for block i. We then calculate the MV on the distorted video (temporally sampled frames if temporal scaling is used) and estimate the MV for the frame of interest (i.e., we scale the MV accordingly) and denote as $\mathrm{MV}_d(i)$ for the ith block.

2.5. Dependence on the interpolation filter

In cases where bitrate reduction is achieved by spatial and temporal scalabilities, the resulting video must be subject to spatial and/or temporal interpolation before computation of distortion and for proper display. Then, the distortion between the original and decoded videos depends on the choice of the interpolation filter. For spatial interpolation, we use the 7-tap synthesis filter, which is reported as the best interpolating filter for signals downsampled using the 9-tap (9–7) Daubechies wavelet [27]. We verified that this inverse wavelet filter performed, on the average, 0.2 dB better than

the 6-tap filter of the H.264 standard [2]. Temporal interpolation should ideally be performed by MC filters [28]. However, when the low frame rate video suffers from compression artifacts such as flatness and blockiness, MC filtering is not successful. On the other hand, simple temporal filtering, without MC, results in ghost artifacts. Hence, we employ a zero-order hold (frame replication) for temporal interpolation, which results in temporal jerkiness distortion.

3. CONTENT-AWARE SELECTION OF SCALABILITY TYPE

In this section, we first present a list of scalability options for each video segment, assuming that the input video is parsed (divided) into temporal segments and each segment is classified into one of K classes according to content type using a content analysis algorithm. Shot boundary determination and shot-type classification, which are beyond the scope of this paper, can be done automatically for certain content domains using existing techniques, for example, for soccer videos [19]. Next, we formulate the problem of selecting the best scalability option for each temporal video segment (according to its content type) among the list of available scalability options, such that the optimal option yields minimum total distortion, which is quantified as a function of the individual distortion measures presented in Section 2. Finally, the training procedure for determination of the coefficients of the linear combination, which quantify the total distortion, as a function of the content type of the video segment is presented.

3.1. Scalability-type choices

There are three basic scalability options: temporal, spatial, and SNR scalabilities. Temporal scalability can be achieved by skipping high frequency frames and their motion vectors following MCTF. Jerkiness may be observed at the low frame rate. Spatial scaling introduces blur (due to interpolation back to original size for display) and ringing. We observe that spatially scaled videos have lower PSNR (after interpolating back to original size) than their visual quality suggests (see Figure 2). SNR scalability is provided by the embedded entropy coding of subbands after temporal and spatial decompositions. We also consider combinations of scalability types to allow for hybrid scalability modes. In this work, we allow six combinations of scaling operators, shown in Table 1, that constitute a reasonable subset of scalability options for the target bitrates (100–300 kbps), where the original resolution has been CIF-30 fps.

3.2. An objective function for scalability-type selection

Most existing methods for adaptation of the video coding rate are based on adaptation of the SNR (quantization parameter) only, because (i) it is not straightforward to employ the conventional rate-distortion framework for adaptation of temporal, spatial, and SNR resolutions simultaneously, which requires multidimensional optimization; (ii) PSNR is not an appropriate cost function for considering tradeoffs between temporal, spatial, and SNR resolutions.

Table 1: Scaling options, included scalability types, and resulting resolutions used.

Options	Included scalabilty types	Resolution
Option 1	SNR only	CIF, 30 fps
Option 2	Temporal + SNR	CIF, 15 fps
Option 3	Spatial + SNR	QCIF, 30 fps
Option 4	Spatial + temporal + SNR	QCIF, 15 fps
Option 5	2-level temporal + SNR	CIF, 7.5 fps
Option 6	2-level temporal + spatial + SNR	QCIF, 7.5 fps

Considering the above limitations, we propose a quantitative method to select the best scalability option for each temporal segment by minimizing a visual distortion measure (or cost function). In [29], a distortion metric which is a linear combination of distinct distortion metrics such as edgeness and temporal decorrelation has been proposed. Following a similar approach, we define an objective function of the form

$$\begin{aligned} \mathrm{D}(m) = {} & \alpha_{\mathrm{block}}(i)\mathrm{D}_{\mathrm{block}}(m) + \alpha_{\mathrm{flat}}(i)\mathrm{D}_{\mathrm{flat}}(m) \\ & + \alpha_{\mathrm{blur}}(i)\mathrm{D}_{\mathrm{blur}}(m) + \alpha_{\mathrm{jerk}}(i)\mathrm{D}_{\mathrm{jerk}}(m), \end{aligned} \tag{7}$$

where, $\alpha_{\mathrm{block}}(i)$, $\alpha_{\mathrm{flat}}(i)$, $\alpha_{blur}(i)$, and $\alpha_{\mathrm{jerk}}(i)$ are the weighting coefficients for blockiness, flatness, blurriness, and jerkiness measures for shot type $i(1 \le i \le K)$, and $\mathrm{D}_{\mathrm{block}}(m)$, $\mathrm{D}_{\mathrm{flat}}(m)$, $\mathrm{D}_{\mathrm{blur}}(m)$, $\mathrm{D}_{\mathrm{jerk}}(m)$, $\mathrm{D}(m)$, respectively, denote the blockiness, flatness, blurriness, jerkiness, and total distortions of video m with shot type i. A procedure for determination of the coefficients of the cost function according to content type is presented in the following section. The weights depend on the content type because different distortions appear to be dominant for different content types.

3.3. Distortion mapping procedure

In this section, we present a training procedure, including a subjective test (Subjective Test-I), in order to determine the coefficients $\alpha_{\mathrm{block}}(i)$, $\alpha_{\mathrm{flat}}(i)$,$\alpha_{\mathrm{blur}}(i)$, and $\alpha_{\mathrm{jerk}}(i)$ $(1 \le i \le K)$ of the cost function for each content type. This procedure is summarized in Table 2. The basic idea is to select the coefficients such that the objective measure (7) is in agreement with the results of the Subjective Test-I as closely as possible. To this effect, a subjective distortion score (8) is defined in Section 4.3 based on the results of Subjective Test-I conducted on a training set of shots representing each content-type class. The coefficients are computed for each content-class type separately by linear regression, that is, least-squares fitting of the objective cost function (7) to subjective distortion scores for that class type. Specifically, let y^i be $M \times 1$ vector consisting of the subjective distortion scores of M training videos belonging to the shot type i, $1 \le i \le K$. Also, let w^i be the $N \times 1$ vector of coefficients of shot type i, where N is the cardinality of the distortion function set, where $N = 4$ in our case, that is, $w^i = [\alpha_{\mathrm{block}}(i), \alpha_{\mathrm{flat}}(i), \alpha_{\mathrm{blur}}(i), \alpha_{\mathrm{jerk}}(i)]^T$. Let distortion measures of M training videos form the $M \times N$ H

Table 2: Coefficient determination procedure.

(1)	Divide video into shots and identify shot content type using the method in [17]
(2)	For each shot type i, $1 \le i \le K$
(3)	Generate a pool of training videos that contain all distortion types
(4)	Calculate distortion measures for each video m, $1 \le m \le M$
(5)	Obtain subjective distortion measures, that is, y from subjective tests
(6)	Find optimal coefficient set for shot type i, as $w^i_{\text{opt}} = (H^T H)^{-1} H^T y$, from (9)

matrix, where mth ($1 \le m \le M$) row of the H matrix is $[\mathrm{D}_{\text{block}}(m), \mathrm{D}_{\text{flat}}(m), \mathrm{D}_{\text{blur}}(m), \mathrm{D}_{\text{jerk}}(m)]$, corresponding to the distortion measures for video m. Then, optimal coefficients can be found by minimizing the mean square error:

$$w^i = \arg\min \|y - Hw\|. \tag{8}$$

The solution of this problem is well known when $H^T H$ is invertible,

$$w^i_{\text{opt}} = \left(H^T H\right)^{-1} H^T y. \tag{9}$$

If $H^T H$ is near singular (which is not observed in our experiments), a regularized solution (in the Tikhanov-Miller sense [28]) given by $w^i_{\text{opt}} = (H^T H + \alpha I)^{-1} H^T y$, where α is the regularization coefficient, should be computed.

3.4. Potential applications and methods for complexity reduction

Potential applications of the proposed method include (1) Content repurposing: video stored at a server using embedded coding at a high enough bitrate can be downscaled to the target bitrate (CBR). Both steps in Figure 3 can be performed offline for this application. (2) Video streaming over time-varying channels: if the throughput of the user is time-varying, then a different target bitrate can be determined for each group of pictures (GoP), and the process becomes GoP-based rate adaptation by scaling option selection. The scaling option selected at the server side can be sent as side information so that the receiver (client) performs appropriate spatial/temporal interpolation, when necessary, for display. In the latter application, some additional steps may be taken to reduce the complexity of the proposed method for real-time rate adaptation.

(i) Distortion functions can be replaced with less complex ones. For example, the current jerkiness measure requires performing another motion search between downsampled frames. An alternative metric can be employed, which is based on only motion vectors between frames at the original temporal resolution computed at the time of encoding. Also, calculations that are common to different scaling options may be estimated from previously calculated values.

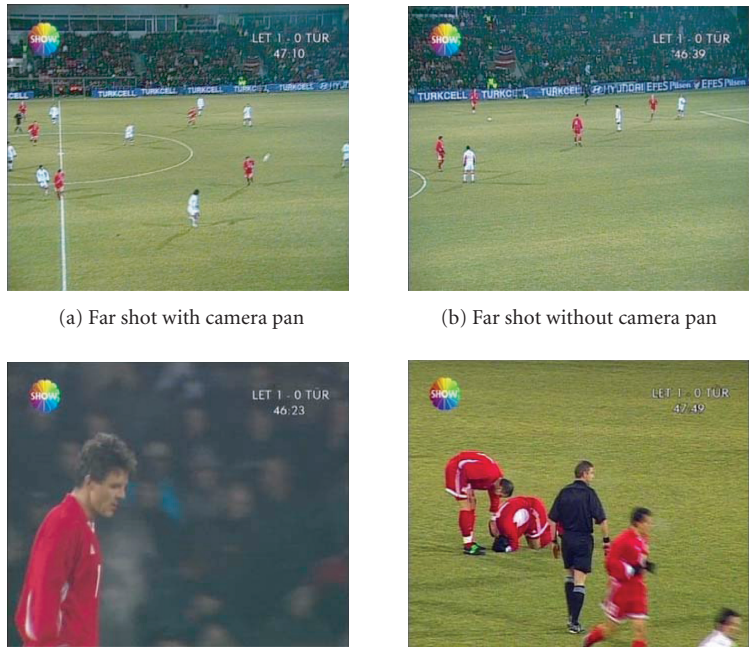
(ii) A smaller set of scaling options can be tested depending on the shot type. For example, according to our experiments, spatial scalability was not preferred for most shot types. Hence, the option of spatial scalability can be excluded depending on the shot type.

4. RESULTS

We present two subjective tests, Test-I for training and Test-II for validation of the proposed scalability-type selection method. The goal of Test-I is the determination of the coefficients of the overall cost function for individual shot types using a training process. Test-II aims to evaluate of the performance of the proposed content-adaptive bitrate scaling system for an entire video clip which consists of several temporal segments to demonstrate that video scaled according to the proposed adaptive segment-based variation of the scalability type is visually preferred to videos scaled by using a single scalability type for the whole duration. The data set obtained from Test-I is also statistically analyzed to verify that the best scaling type depends on the bitrate, shot type, and user preferences. In our tests, a wavelet coder [30] is employed with four-level temporal and three-level spatial decomposition and GoP size of 32 frames, using advanced motion compensation (MC) techniques, such as variable block sizes, 1/4 pixel accuracy motion vectors, several MC modes as those used in the H.264 standard [31], and overlapped block MC. For entropy coding, it uses the 3D embedded subband coder with optimized truncation (3D-ESCOT) [32], which provides rate-distortion-optimized multiplexing of subbands that are independently coded by bitplane coding. Any other video coder can be utilized within the proposed scheme, with minor modifications to the distortion functions. Also, the subjective test to find the coefficient sets should be performed again with the new coder. For practical deployment of the proposed scalability-type selection method, video encoded at the highest resolution (rate) is taken as the original video at the server for the computation of distortion functions. Examples provided in the tests have been selected from the sports domain. In order to apply the proposed procedure to other content domains, the training step (presented in Section 3.3) and hence the subjective tests need to be reperformed.

4.1. Subjective Test-I

The goal of Test-I is to determine the coefficients of the objective cost function (6) for individual shot types using a training process (presented in Section 3.3). This test is set up with 20 subjects according to ITU-R Recommendation BT.500-10 [33], using a three-level evaluation scale instead of ten levels. A *single-stimulus comparison scale* is used in the test, that is, assessors viewed six videos generated by the scaling



(a) Far shot with camera pan

(b) Far shot without camera pan

(c) Close shot with camera pan

(d) Close shot with camera pan

Figure 5: Four shot types with respect to distance of shots and type of motion.

options listed in Section 2.2 in random order without seeing the originals. For each "rate"-"shot-type" combination, each assessor was asked to rank the six videos using the three levels: good, fair and poor; with ties allowed. The video clips used are of 3–5-second duration at CIF resolution and contain typical shots from a soccer game. For the soccer video domain, we define 4 shot types according to camera motion and distance Type-1, far shot with camera pan; Type-2, far shot without camera pan; Type-3, close shot with camera pan; Type-4, close shot without camera pan. Examples of these shot types are shown in Figure 5. We tested three different rates: 100 kbps, 200 kbps and 300 kbps. At these rates, all shot types other than Shot-3 (close shot with camera pan) are affected by flatness, blurriness, and jerkiness distortions; Shot-3 has blockiness instead of flatness as the significant artifact. Each subject evaluated four shot types decoded at three different bitrates with 6 different scaling options. For each subject, the evaluation is organized into 12 sessions, where in a single session a subject evaluated one shot type decoded at the same bitrate for six different scaling options. Calculation of coefficients given the results of Test-I is explained in Section 4.3.

4.2. *Statistical analysis of Test-I results*

We performed statistical analysis of the results of these subjective tests to answer the following questions.

(i) Is there a statistically significant difference in the assessors choices created by the scalability type selection? In other words, does scalability-type matter?

(ii) Is the shot-content type a statistically significant factor in the assessor's choices of scalability type?

(iii) Is the bitrate a statistically significant factor in the assessor's choices in addition to the shot-content type?

(iv) Are there significant clusters in the choices of assessors, that is, is the scalability-type preference user-dependent?

To answer the first three questions, we applied the Friedman test [34], which evaluates whether a selected test variable, for example, rate, shot type, and so forth, can be used to form test result clusters that contain significantly different results as compared to a random clustering. The Friedman test is especially a good fit for this evaluation since it does not have any distribution assumption on the data. The output of this test, ρ, is the significance level, which represents the probability that a random clustering would yield the same or better groups. A result with ρ less than 0.05 or 0.01 is assumed to be significant in general. We found that

(i) clustering with respect to the scaling option is significant with ρ almost equal to zero, that is, scaling-type selection is indeed significant;

(ii) clustering with respect to shot type is also found to be significant with $\rho = 0.004$;

(iii) in addition to scaling type and shot type, rate is a significant factor in clustering with significance $\rho = 0.001$.

In order to analyze dependence of the results on user preferences, we first calculated the correlation of user scores. The correlations shown in Figure 6 indicate that there are two types of users: one group prefers higher picture quality over higher frame rate (type-A) and the other group prefers higher frame rate (type-B). Based on this observation,

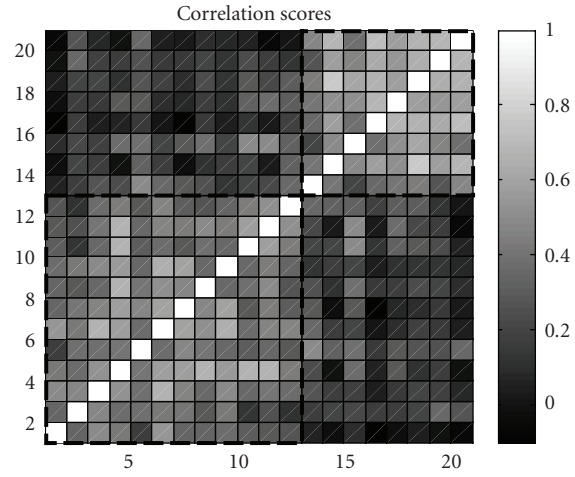


Figure 6: The autocorrelation of subjective scores shows a noticeable clustering of two groups of subject.

we clustered subjects into two groups using 2-mean clustering. We also determined the significance of the clustering by rank-sum test for each video. The separation of users into two groups is found to be significant at 5% level for 30 videos out of 72 videos coded with different scaling option, rate, and shot-type combinations. Most of these 30 videos that users' preferences differ are coded at low rates, which leads us to conclude that the difference in the users frame rate preferences increases as the overall video quality decreases. This observation is also confirmed by Subjective Test-II.

4.3. Distortion mapping

To map the subjective scores to objective scores, we define the subjective distortion score (SDS) of a video shot (segment) as

$$\mathrm{SDS} = \frac{5}{1 + (2S_1 + S_2)/(2S_{\max})}, \tag{10}$$

where S_1 and S_2 are the numbers of "good" and "fair" grades, respectively, and $S_{\max}$ is the number of subjects. This is an empirical function that matches the visual quality (i.e., good, bad, fair) to objective measure in the range of 1–5. Alternatively, distortions also can directly be asked to the subjects and average can be used as measure, as done in [6]; however, this requires a larger distortion measure set that may decrease the performance of subjective test, for example, subjects may be inconsistent to decide between distortion levels, such as between distortion levels 4 and 5, but are likely to make a more reliable decision among bad, fair, and good quality. Nevertheless, any of the methods will not affect the results significantly, as long as identical methods are used in both training and testing.

We determine the coefficients of the objective cost function (7) for each shot type by least-squares fitting to corresponding SDS (10), as explained in Section 3.3. The coefficient sets computed for all users together, and type-A users and type-B users separately, are shown in Table 3, showing the variation of coefficient with respect to shot type. Also note that flatness and blockiness are not present in every shot type, which results zero coefficients.

To show that the coefficients computed at a given rate also perform well at other content and bitrates for a particular shot type, we computed the Spearman rank correlation between the SDS (10) and the ranking provided by our method as shown in Table 3, on a new test set. Spearmen rank correlation is a useful metric to measure the performance in rankings [34], and since rankings, instead of absolute values, are important to choose the best operator, we employed Spearmen rank correlation in this comparison. It can be seen that our algorithm finds the best or the second best scaling option from the six scaling options for most cases. Furthermore, the results of the Subjective Test-II confirm that coefficients found for a given shot type in a specific video will work for the same shot type in any other video. We also employed the well-known VQM objective measure, defined in [8, 35], instead of our objective measure (7) in the proposed selection scalability option selection algorithm at several bitrates (see Table 4). Table 4 also illustrates the VQM results for the video with highest visual quality to show the quality range of videos used in the test. Results show that our metric performs better than the VQM, since VQM does not adapt to different contents, and hence *these results show the merit of adapting the coefficients with respect to shot type*.

4.4. Subjective Test-II

In this test, a new test video clip is divided into temporal segments according to the shot types defined above. For each temporal segment, the best scaling option is determined using our proposed method with coefficients determined as described above. The segments extracted with the best scaling option are then cascaded to form test video. It is important to notice that in this subjective test, videos are in cascaded form of different shot types, to show the merit of the proposed system under scaling-type changes from shot to shot, that is, the results of this test also include the end user satisfaction evaluated for the whole video with scaling option jumps. In this test, two comparisons are performed to answer two questions.

Does changing the scalability option with respect to content type really make significant difference in the visual quality of the scaled video when compared to using the same scalability option for the whole sequence? To answer this question, adaptively scaled video is compared to videos decoded at the same rate but obtained with all fixed scaling options, that is, subjects are asked to choose the most pleasing video among seven videos, six obtained from six fixed scaling options and one obtained by adaptively changing scaling type.

Is it useful to consider subject type (i.e., type-A or type-B as defined in Section 4.2) in determining the best scalability option? Changing the scalability option according to subject type requires knowledge of the subject type beforehand which makes the system rather difficult to implement, so learning the extent of the improvement when subject type is used will be beneficial for practical application scenarios. To answer this question, subjects are asked to choose from

Table 3: The normalized coefficients of the cost function for all users, type-A users, and type-B users, respectively.

	Blurriness	Flatness	Blockiness	Jerkiness
Shot-1	0.374 /0.428/0.237	0.2158/0.243/0.191	0/0/0	0.355/0.240/0.627
Shot-2	0.254/0.294/0.209	0.337/0.419/0.221	0/0/0	0.468/0.311/ 0.664
Shot-3	0.498/0.629/0.114	0/0/0	0.096/0.0664/0.191	0.291/0.164/0.837
Shot-4	0.418/0.534/0.250	0.378/0.328/0.407	0/0/0	0.136/0.0216/0.410

Table 4: The performance of our optimal operator selection algorithm: the Spearman rank correlation, the subjective rank of the option that our algorithm finds, and the subjective rank of the option that another objective metric finds (applicable for only all users part), respectively. VQM results show the VQM measure (scale 5) for the video with highest visual quality.

	All users			Type-A users			Type-B users			VQM results		
	100 kbps	200 kbps	300 kbps	100 kbps	200 kbps	300 kbps	100 kbps	200 kbps	300 kbps	100 kbps	200 kbps	300 kbps
Shot-1	0.74/1/1	0.94/1/4	0.77/1/3	0.6/1	0.83/1	0.54/2	0.84/1	0.9/1	1/1	3.62	4.07	4.17
Shot-2	0.31/3/5	0.71/1/1	0.99/1/1	0.17/3	0.37/1	1/1	0.99/1	0.99/1	1/1	2.95	3.60	3.94
Shot-3	0.43/4/3	0.77/1/1	0.49/1/1	0.5/4	0.93/1	0.6/1	0.77/3	0.79/1	0.37/1	3.82	4.47	4.71
Shot-4	0.86/1/4	0.94/1/4	1/1/1	0.93/1	0.84/2	0.69/2	0.81/2	0.9/1	1/1	2.73	3.36	3.86

Table 5: The first row shows percentage of users who prefer the proposed content-aware scaling to all fixed scaling options. The second row shows the percentage of subjects who preferred the adaptive scaling option with respect to subject type rather than constant scaling option with respect to subject type.

	100 kbits	200 kbits	300 kbits
Adaptive scaling performance	95%	75%	75%
Bimodal user separation	20%	5%	5%

Table 6: An example of content-adaptive scaling option selection for different subject types.

	Shot-1	Shot-2	Shot-3	Shot-4	Shot-5
Type-A	Option 2	Option 1	Option 1	Option 5	Option 5
Type-B	Option 1	Option 1	Option 1	Option 4	Option 5

videos which are content adaptively scaled with coefficient sets tuned to their specific subject types versus tuned to general type.

The results confirm that content adaptive scaling provides significant improvement over fixed scaling as shown in the first row of Table 5. Majority of the subjects prefer dynamically scaled video to any constant scaling option for all bitrates tested. The performance gain obtained by separating the subjects into two groups, in addition to content adaptivity, is presented in second row of Table 5. The effect of subjective preferences on the scalability operator selection is observed to be somewhat important at low bitrates and not important at higher rates; a result which was observed in the first subjective test also. An example of chosen scaling preferences for different types of subjects is shown in Table 6. Note that in this part, we compare *content* adaptive scaling to *content and subject* adaptive scalings.

This result agrees with the observation that "information assimilation" (i.e, where the lines are, who the players are, which teams are playing) of a video is not affected by the frame rate but "satisfaction" is [36]. At high bitrates, spatial quality is high enough for information assimilation and the best scalability operator is selected mainly from satisfaction point of view which leads to similar choices of scaling option for all users. At low rates, picture quality may not be good enough for information assimilation. Hence, information assimilation plays a key role on optimal operator selection for type-A subjects; where for type-B subjects satisfaction is still more important in determination of optimal scaling choice, resulting in significant clustering among subjects in the subjective evaluation of videos coded at low rates.

5. CONCLUSIONS

In this work we propose a content adaptive scalable video streaming framework, where each temporal segment is coded with the best scaling option. The best scaling option is determined by a cost function which is a linear combination of different distortion measures such as blurriness, blockiness, flatness, and jerkiness. Two subjective tests are performed to find the coefficients of the cost function and to test the performance of the proposed system. Statistical significances of the test variables are analyzed. Results clearly show that best scaling option changes with the content, and content adaptive coding with optimum scaling option results in better visual quality. Although our results and analysis are provided for soccer videos, the proposed method can be applied to other types of video content as well.

ACKNOWLEDGMENTS

A preliminary version of this work has been presented in the Picture Coding Symposium, December 2004 [18]. This work has been done while Emrah Akyol and Reha Civanlar were also with Koc University, Istanbul, Turky. It has been supported by the Eurpean Commission within FP6 under the Network of Excellence Grant 511568 with acronym 3DTV.

REFERENCES

[1] J.-R. Ohm, "Advances in scalable video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42–56, 2005.

[2] J. Reichel, H. Schwarz, and M. Wien, "Scalable video coding - Working Draft 1," *Joint Video Team (JVT)*, Doc. JVTN020, Hong Kong, January 2005.

[3] A. Puri, X. Chen, and A. Luthra, "Video coding using the H.264/MPEG-4 AVC compression standard," *Signal Processing: Image Communication*, vol. 19, no. 9, pp. 793–849, 2004.

[4] R. Kumar Rajendran, M. van der Schaar, and S. F. Chang, "FGS+: optimizing the joint spatio temporal video quality in MPEG-4 fine grained scalable coding," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '02)*, Phoenix, Ariz, USA, May 2002.

[5] C. Kuhmünch, G. Kühne, C. Schremmer, and T. Haenselmann, "Video-scaling algorithm based on human perception for spatio-temporal stimuli," in *Multimedia Computing and Networking (MMCN '01)*, vol. 4312 of *Proceedings of SPIE*, pp. 13–24, SPIE Press, San Jose, Calif, USA, January 2001.

[6] Y. Wang, M. van der Schaar, S.-F. Chang, and A. C. Loui, "Classification-based multidimensional adaptation prediction for scalable video coding using subjective quality evaluation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1270–1279, 2005.

[7] B.-F. Hung and C.-L. Huang, "Content-based FGS coding mode determination for video streaming over wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1595–1603, 2003.

[8] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in *Proceedings of the Multimedia Systems and Applications II*, vol. 3845 of *Proceedings of SPIE*, pp. 266–277, Boston, Mass, USA, September 1999.

[9] E. C. Reed and J. S. Lim, "Optimal multidimensional bit-rate control for video communication," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 873–885, 2002.

[10] A. Vetro, Y. Wang, and H. Sun, "Rate-distortion optimized video coding considering frameskip," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 3, pp. 534–537, Thessaloniki, Greece, October 2001.

[11] Y. Wang, J.-G. Kim, and S.-F. Chang, "Content-based utility function prediction for real-time MPEG-4 video transcoding," in *Proceedings of IEEE International Conference on Image Processing (ICIP '03)*, vol. 1, pp. 189–192, Barcelona, Spain, September 2003.

[12] P. Yin, A. Vetro, M. Xia, and B. Liu, "Rate-distortion models for video transcoding," in *Image and Video Communications and Processing*, vol. 5022 of *Proceedings of SPIE*, pp. 479–488, Santa Clara, Calif, USA, January 2003.

[13] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 207–220, MIT Press, Cambridge, Mass, USA, 1993.

[14] S. Winkler, C. J. B. Lambrecht, and M. Kunt, "Vision and video: models and applications," in *Vision Models and Applications to Image and Video Processing*, C. J. B. Lambrecht, Ed., chapter 10, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[15] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "Objective video quality assessment system based on human perception," in *Human Vision, Visual Processing, and Digital Display IV*, vol. 1913 of *Proceedings of SPIE*, pp. 15–26, San Jose, Calif, USA, February 1993.

[16] K. T. Tan and M. Ghanbari, "A multi-metric objective picture-quality measurement model for MPEG video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1208–1213, 2000.

[17] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, 2000.

[18] E. Akyol, A. M. Tekalp, and M. R. Civanlar, "Optimum scaling operator selection in scalable video coding," in *Picture Coding Symposium*, pp. 477–482, San Francisco, Calif, USA, December 2004.

[19] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.

[20] A. Kokaram, N. Rea, R. Dahyot, et al., "Browsing sports video: trends in sports-related indexing and retrieval work," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 47–58, 2006.

[21] C. G. M. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[22] S.-F. Chang and P. Bocheck, "Principles and applications of content-aware video communication," in *Proceedings of the IEEE Internaitonal Symposium on Circuits and Systems (ISCAS '00)*, vol. 4, pp. 33–36, Geneva, Switzerland, May 2000.

[23] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247–278, 1998.

[24] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.

[25] L. Shapiro and G. Stockman, *Computer Vision*, Prentice-Hall, Upper Saddle River, NJ, USA, 2000.

[26] F. Pan, X. Lin, S. Rahardja, et al., "A locally adaptive algorithm for measuring blocking artifacts in images and videos," *Signal Processing: Image Communication*, vol. 19, no. 6, pp. 499–506, 2004.

[27] T. Frajka and K. Zeger, "Downsampling dependent upsampling of images," *Signal Processing: Image Communication*, vol. 19, no. 3, pp. 257–265, 2004.

[28] A. M. Tekalp, *Digital Video Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 1995.

[29] A. P. Hekstra, J. G. Beerends, D. Ledermann, et al., "PVQM—a perceptual video quality measure," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 781–798, 2002.

[30] J. Xu, R. Xiong, B. Feng, et al., "3D sub-band video coding using barbell lifting," ISO/IEC JTC/WG11 M10569, S05.

[31] L. Luo, F. Wu, S. Li, Z. Xiong, and Z. Zhuang, "Advanced motion threading for 3D wavelet video coding," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 601–616, 2004, *special issue on Subband/Wavelet Interframe Video Coding*.

[32] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "Three-dimensional embedded subband coding with optimized truncation (3-D ESCOT)," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 290–315, 2001.

[33] "Methodology for the subjective assessment of the quality of television pictures," Recommendation ITU-R BT.500-10, ITU Telecommunication Standardization Sector, Geneva, Switzerland, August 2000.

[34] J. Devore, *Probability and Statistics for Engineering and the Sciences*, Duxbury Press, Pacific Grove, Calif, USA, 1999.

[35] VQM software, http://www.its.bldrdoc.gov/n3/video/vqm-software.htm.
[36] S. R. Gulliver and G. Ghinea, "Changing frame rate, changing satisfaction? [Multimedia quality of perception]," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 1, pp. 177–180, Taipei, Taiwan, June 2004.

Emrah Akyol received the B.S. degree in electrical and electronics engineering in 2003 from Bilkent University, Ankara, Turkey, and the M.S. degree in electrical and computer engineering in 2005 from Koç University, Istanbul, Turkey. He is currently a Ph.D. student at University of California, Los Angeles (UCLA) and Intern at NTT DoCoMo Communication Laboratories, Palo Alto, Calif. His research interests include video compression and streaming and the signal processing aspects of dynamic voltage scaling. Between June 2006 and November 2006, he was a Research Intern at HP Labs, Palo Alto, where he worked on complexity-constrained media compression. He has coauthored several international publications and one pending patent application.

A. Murat Tekalp received the Ph.D. degree in electrical, computer, and systems engineering from Rensselaer, Troy, New York, in 1984. He has been with Eastman Kodak Company, Rochester, New York, from 1984 to 1987, and with the University of Rochester, Rochester, New York, from July 1987 to June 2005, where he was promoted to Distinguished University Professor. Since June 2001, he is a Professor at Koç University, Istanbul, Turkey. His research interests include video compression and streaming, motion-compensated video filtering for high-resolution, content-based video analysis and summarization, multicamera video processing, and protection of digital content. He was named as Distinguished Lecturer by IEEE Signal Processing Society in 1998. He has served as an Associate Editor for the IEEE Transactions on Signal Processing (1990–1992), IEEE Transactions on Image Processing (1994–1996). He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (Jan. 1996–Dec. 1997). He was appointed as the Technical Program Co-Chair for IEEE ICASSP 2000, and the General Chair of IEEE International Conference on Image Processing (ICIP) 2002. He is the Editor-in-Chief of the EURASIP Journal Signal Processing: Image Communication published by Elsevier since 1999. He authored the Prentice Hall textbook Digital Video Processing (1995). He holds seven US patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards. He is an IEEE Fellow

M. Reha Civanlar is VP and Media Lab Director in DoCoMo USA Labs. He was a Visiting Professor, Computer Engineering, Koc University, Istanbul, for four years starting from 2002. He also led a multinational research project on 3DTV transport. He is on the advisory boards of Argela Technologies and Layered Media Inc. Before Koç, he was heading Visual Communications Research Department at AT&T Research, where he worked since 1991. Prior to that, he was at Pixel Machines in Bell Laboratories. His career started as a Researcher in CCSP upon receiving his ECE Ph.D. degree in 1984 from NCSU. He received his B.S. and M.S. degrees in EE from METU, Turkey. He has numerous publications, contributions to international standards, and over forty patents. He is an IEEE Fellow and is a recipient of 1985 Senior Award of IEEE, ASSP. Dr. Civanlar is a Fulbright Scholar and a Member of Sigma Xi. He served as an Editor for IEEE Transactions on Communications, Transactions on Multimedia, and JASP. He is currently an Editor for EURASIP Image Communications. He served on MMSP and MDSP technical committees of the IEEE SP Society. His research interests include networked video emphasizing the Internet and wireless networks and video coding.