

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/3308265>

Object-based transcoding for adaptable video content delivery

ARTICLE in IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY · APRIL 2001

Impact Factor: 2.62 · DOI: 10.1109/76.911163 · Source: IEEE Xplore

CITATIONS

69

READS

32

3 AUTHORS:



[Anthony Vetro](#)

Mitsubishi Electric Research Laboratories

254 PUBLICATIONS **4,686** CITATIONS

[SEE PROFILE](#)



[Huifang Sun](#)

Mitsubishi Electric Research Laboratories

178 PUBLICATIONS **4,577** CITATIONS

[SEE PROFILE](#)



[Yao Wang](#)

Dalian University of Technology

191 PUBLICATIONS **3,568** CITATIONS

[SEE PROFILE](#)

All in-text references [underlined in blue](#) are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Anthony Vetro
Retrieved on: 08 April 2016

Object-Based Transcoding for Adaptable Video Content Delivery

Anthony Vetro, *Member, IEEE*, Huifang Sun, *Fellow IEEE*, and Yao Wang, *Senior Member, IEEE*

Abstract—This paper introduces a new framework for video content delivery that is based on the transcoding of multiple video objects. Generally speaking, transcoding can be defined as the manipulation or conversion of data into another more desirable format. In this paper, we consider manipulations of object-based video content, and more specifically, from one set of bit streams to another. Given the object-based framework, we present a set of new algorithms that are responsible for manipulating the original set of video bit streams. Depending on the particular strategy that is adopted, the transcoder attempts to satisfy network conditions or user requirements in various ways. One of the main contributions of this paper is to discuss the degrees of freedom within an object-based transcoder and demonstrate the flexibility that it has in adapting the content. Two approaches are considered: a dynamic programming approach and an approach that is based on available meta-data. Simulations with these two approaches provide insight regarding the bit allocation among objects and illustrates the tradeoffs that can be made in adapting the content. When certain meta-data about the content is available, we show that bit allocation can be significantly improved, key objects can be identified, and varying the temporal resolution of objects can be considered.

Index Terms—Bit-rate reduction, MPEG-4, MPEG-7, quality of service, transcoding hints, universal access, video objects.

I. INTRODUCTION

DURING the past decade, a number of video-coding standards have been developed for communicating video data. These standards include MPEG-1 for CD-ROM storage, MPEG-2 for DVD and DTV applications and H.261/263 for video conferencing. Newer video coding standards, such as MPEG-4 allow arbitrary-shaped objects to be encoded and decoded as separate video object planes (VOPs) [1]. At the receiver, video objects are composed to form compound objects or scenes.

In the context of video transmission, these compression standards are needed to reduce the amount of bandwidth that is required by the network. Since the delivery system must accommodate various transmission and load constraints, it is sometimes necessary to further convert the already compressed bit stream before transmission [2], [3]. Depending on these constraints, conventional transcoding techniques can be classified into three major categories: bit-rate conversion or scaling,

resolution conversion, and syntax conversion. Bit-rate scaling can accommodate deficiency in available bandwidth. Resolution conversion can also accommodate bandwidth limitations, but is primarily used to account for known limitations in the user device, such as processing power, display constraints, or memory capability. To ensure adaptability across hybrid networks, syntax conversion at the protocol layer is required. Syntax conversions may also be considered at the compression layer to ensure receiver compatibility. A recent paper by Shanableh and Ghanbari [4] examine heterogeneous transcoding (e.g., from MPEG-1 to H.263), which simultaneously considers several of the specific conversions listed above. In this paper, we focus on bit-rate scaling and look to exploit the object-level access of MPEG-4 in new and interesting ways.

Until now, most research has focus on the delivery of frame-based video, which has been encoded using such standards as MPEG-1/2 and H.263. We build on this work by proposing a framework that considers an adaptive means of transcoding each of the objects in the scene based on available bandwidth and complexity of each object. The scheme is adaptive in that various techniques can be employed to reduce the rate depending on the ratio of incoming to outgoing rate, and since the goal is to provide the best overall quality for objects of varying complexity, the degradation of each object need not be the same. In fact, some objects may not even be transmitted.

Within this object-based framework, various transcoding strategies can be employed. For instance, in a conservative approach, all objects would be delivered with varying spatial and temporal quality. However, in an aggressive approach and under severe network conditions, the transcoder may consider dropping less relevant objects from the scene. These distinct approaches are two examples that will be used throughout the paper to characterize the transcoder behavior. Not only do they differ in the actions taken to satisfy constraints on the network or user device, but they also differ in how quality is measured. These points will be elaborated on later.

In this paper, techniques that support both conservative and aggressive transcoding strategies within the object-based framework are proposed. For the conservative strategy, the main objective is to consider the various spatio-temporal tradeoffs among all objects in a scene. To this end, a dynamic programming approach that jointly considers the quantization parameter (QP) selection for each object and frameskip is presented. The bit reallocation among objects is a critical process in the transcoder and is accomplished with the use of modified rate-quantizer models that were originally used for bit allocation in the encoder. The main difference is that the transcoder extracts information directly from the compressed

Manuscript received June 15, 2000; revised December 7, 2000. This paper was recommended by Guest Editor S.-F. Chang.

A. Vetro and H. Sun are with the Mitsubishi Electric Research Laboratories, Murray Hill, NJ 07947 USA.

Y. Wang is with the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY 11201 USA.

Publisher Item Identifier S 1051-8215(01)01857-2.

domain, such as the residual energy of DCT blocks, hence certain parameters must be redefined.

To further enhance the quality that can be achieved with the above method, and also consider a set of tools that would support the aggressive transcoding strategy, we consider a new transcoding method which is based on the emerging MPEG-7 standard, formally called *Multimedia Content Description Interface* [5]. It aims to standardize a set of descriptors and description schemes that can be used to describe multimedia content for a variety of applications, such as content navigation, content management and fast search and retrieval. For video, the descriptors are low-level features of the content, such as color, texture and motion [6]. These low-level features can be organized and referenced by higher-level description schemes. The description schemes may refer to various levels of the program, such as a segment or frame, and may also be used to facilitate video summarization, specify user preferences, and support the annotation of video segments. A more complete specification of the various description schemes that have been adopted by MPEG can be found in [7].

Besides using meta-data for search and browsing applications, it is also of interest to consider meta-data that would be suitable to guide the operation of a transcoder. Such meta-data is referred to as a *transcoding hint* within the MPEG-7 community. In [8], [1], [2], such hints were introduced within the context of a Universal Multimedia Access (UMA) framework. The specifics of this work and extensions of it have been investigated by the MPEG committee, where a complete specification of the transcoding hints that have been adopted by the standard can be found in [7]. The specific extensions that we have proposed for object-based transcoding are discussed in this paper [9].

The use of meta-data for transcoding applications is most applicable in nonreal-time scenarios, i.e., with stored content. For instance, before the content is transmitted or needs to be converted, the meta-data can be efficiently extracted. When a request for the content is presented, any external conditions, e.g., network conditions, can be assessed and the transcoding can be appropriately executed. The advantage that one gains in using meta-data is improved quality and reduced complexity. It should be noted that this meta-data is not limited to descriptions related to the complexity of the scene, e.g., in terms of motion or texture, but may also describe (directly or indirectly) its subjective importance to the user. With this, algorithms have been developed to consider the distribution of available bit-rate to different segments of the video sequence, and within those segments to the different regions or objects in the scene. More generally, we consider algorithms that are capable of adapting the quality of the scene in various ways. Several transcoding hints are proposed in this paper to meet a number of objectives. The first of these objectives is to improve bit allocation among objects, the second is to identify key objects in the scene, and the third is to regulate the variable temporal resolution of objects in the scene.

The rest of this paper is organized as follows. In the next section, the object-based transcoding framework is presented, including a discussion of the two transcoding strategies mentioned above. In Section III, a dynamic programming approach that jointly considers the QP selection and frameskip is presented. In Section IV, a set of transcoding hints are proposed and we

discuss how they can be used by the transcoder. In Section V, we briefly discuss the architecture of an MPEG-4 transcoder. In Section VI, simulation results are presented. Finally, the main contributions of this work are summarized in Section VII.

II. OBJECT-BASED TRANSCODING FRAMEWORK AND STRATEGIES

This section introduces the object-based transcoding framework and discusses two transcoding strategies that will be considered in this paper—a conservative approach and an aggressive approach.

A. Framework

Fig. 1 shows a block diagram of an adaptive transcoding system. The input to the transcoder includes object-based bit streams, where the bit streams have a total rate R_{in} . The outputs of the system are the same bit streams with a reduced total rate $R_{out} < R_{in}$. In a nonreal-time scenario, features are extracted from the bit stream before being processed by the transcoder. The features may include descriptors that can be used by a search engine or other types of meta-data that describe the characteristics of the content. The content and associated meta-data are stored on the server until a request for the content is made. In a real-time scenario, the content must be transmitted right away, therefore any information extracted from the bit stream is of a causal nature. In this way, the feature extraction analysis is limited in that the meta-data produced is instantaneous, i.e., it can only be associated with the current content or content that has already been sent.

The transcoding control unit that we consider in this paper includes three types of analysis units: one responsible for the selection of QPs, another to investigate temporal decisions such as frameskip, and a final one for meta-data analysis. The details of the meta-data analysis will become more clear later in the context of specific transcoding hints, but we should mention that the outcome of its analysis will impact decisions made on the QP and frameskip. Also, the elements in the transcoding control unit are not limited to the ones mentioned here. In fact, as reported in [10], the control unit may include a resize analysis and shape analysis to consider reductions in texture and shape resolution, respectively. In any case, the control information that is passed to each transcoder is relayed from this transcoding control unit so that the bit stream manipulations can be executed. We should note that in practice, there are not separate transcoding units for each object. The figure is only illustrated in this way to convey the general concept. It should be clear though that all of the new degrees of freedom make the object-based transcoding framework very unique and desirable for network applications.

As in the MPEG-2 and H.263 coding standards, MPEG-4 exploits the spatio-temporal redundancy of video using motion compensation and DCT. As a result, the core of the object-based transcoder is quite similar to MPEG-2 transcoders that have been proposed earlier (see, for example, [2], [3]). The major difference is that shape information is now contained within the bit stream, and with regard to texture coding, some new tools have been adopted, such as dc and ac prediction for intra blocks. It

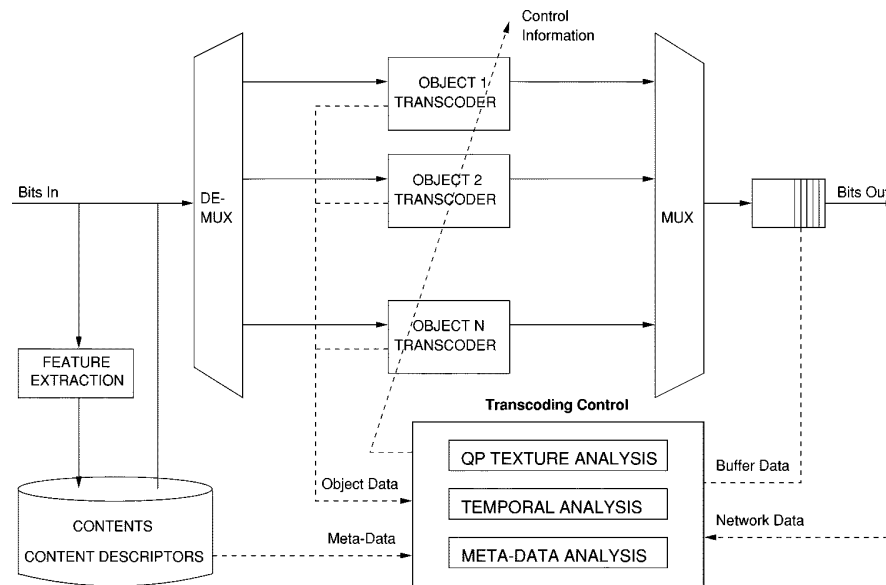


Fig. 1. Adaptive transcoding system using meta-data.

is important to note that the transcoding of texture is indeed dependent on the shape data. In other words, the shape data cannot simply be parsed out and ignored; the syntax of a compliant bit stream depends on the decoded shape data.

In principle though, and at this high-level, the proposed framework is similar to the multiprogram transcoding framework proposed by Sorial *et al.* [11]. The major difference is that the input bit streams do not correspond to frame-based video programs. Also, MPEG-2 does not permit dynamic frame skipping; the GOP structure and reference frames are usually fixed. However, it is possible to skip frames in MPEG-1/2 by inserting dummy frames that do not use any bits for the motion and texture data. In this way, all of the macroblocks in a frame are skipped, where at least 1 bit per macroblock must be spent [12].

The main point worth noting is that the objects themselves need not be transcoded with equal quality. For example, the texture data of one object may be reduced, keeping its temporal rate intact, while the temporal rate of another object be reduced, keeping its texture information intact. Many other combinations can also be considered, including dropping frames or objects. In a news clip, for example, it would be possible to reduce the temporal rate along with the texture bits for the background, while keeping the information associated with a news reader intact.

B. Strategies

Within the object-based transcoding framework, we consider two types of transcoding strategies. One that we refer to as conservative, and another which we refer to as aggressive. These distinct strategies are useful in characterizing the behavior of different transcoders and also in justifying the need for some of the transcoding hints proposed later on.

Both strategies aim to maximize quality, but differ in the actions taken to do so, as well as the measures that indicate quality.

In the conservative approach, quality should be measured objectively, i.e., by traditional PSNR-like measures, whereas in the aggressive approach, overall perceptual quality should be considered. With traditional PSNR-like measures, there is limitation in what can be measured. In other words, given a scene with several objects, PSNR cannot indicate the relevance of each object to a user; it can only measure bit-to-bit differences. To measure the overall perceptual quality is difficult—it requires relevance feedback by the user. The best that a transcoder can do without this feedback is attempt to identify the key objects in a scene through characteristics of the signal itself. Although this is not always possible, since less relevant objects may display relevant object characteristics, it may be possible for certain applications in which the nature of the content is known.

If we define the conservative approach to maximize PSNR quality, this implies that all objects in the scene must be transmitted. Of course, the objects will vary in spatial and temporal quality, but the PSNR between the original and reconstructed scene can be calculated. However, when varying the temporal quality of multiple objects, a major subtlety exists, that is, it is possible that different temporal rates for each object may produce undefined areas in the reconstructed scene. To illustrate, consider two objects, VO1 at a rate of 10 Hz and VO2 at a rate of 30 Hz. If the shape boundary of VO2 changes before the next VO1 is received, such undefined areas will appear in the composite image; we refer to this as a composition problem. To overcome this problem, we propose a measure that indicates the change in the shape boundaries over time, i.e., a shape hint. Such a measure is needed to allow for variable temporal quality of objects, but has some other uses as well, which are discussed later.

If we define the aggressive approach to maximize perceptual quality, this implies that all objects in the scene are not necessarily transmitted, i.e., less relevant objects may be dropped. When objects are dropped, traditional PSNR-like

measures cannot be used, since the contents contained in the original and reconstructed scene have changed. To enable such a transcoding strategy, it is essential that key objects in the scene be identified. For this purpose, we will identify certain transcoding hints that can be used and demonstrate their utility.

III. DYNAMIC PROGRAMMING APPROACH

In the previous section, the object-based transcoding framework has been presented. We now provide the details of a dynamic programming based approach that jointly considers the spatial and temporal quality of objects.

A. Texture Model for Rate Control

We begin our discussion with a brief overview of the texture models used in the encoder for rate control [1], [13], [14]. Let R represent the texture bits spent for a video object (VO), Q denote the QP, (X_1, X_2) the first and second-order model parameters and S the encoding complexity, such as the mean absolute difference. The relation between R and Q is given by

$$R = S \cdot \left(\frac{X_1}{Q} + \frac{X_2}{Q^2} \right). \quad (1)$$

Given the target amount of bits that a VO has been assigned and the current value of S , the value of Q is solved for based on the current value of (X_1, X_2) . After a VO has been encoded, the actual number of bits that have been spent is known, and the model parameters can be updated. This can be done by linear regression using the results of the past n frames.

B. QP Selections in the Transcoder

The transcoding problem is different in that Q , the set of original QPs, and the actual number of bits are already given. Also, rather than computing the encoding complexity S from the spatial domain, we must define a new DCT-based complexity measure $\tilde{S}$. This measure is defined as

$$\tilde{S} = \frac{1}{M_c} \sum_{m \in \mathcal{M}} \sum_{i=1}^{63} \rho(i) \cdot |B_m(i)|^2 \quad (2)$$

where

- $B_m(i)$ ac coefficients of a block;
- m macroblock index in the set $\mathcal{M}$ of coded blocks;
- M_c number of blocks in that set;
- $\rho(i)$ frequency-dependent weighting.

The proposed complexity measure indicates the energy of the ac coefficients, where the contribution of high-frequency components is lessened by the weighting function. This weighting function can be chosen to mimic that of an MPEG quantization matrix.

From the data transmitted in the bit stream, and the data from past VOs, the model parameters can be computed and continually updated. Actually, we can update the model twice for every transcoded VOP; once before transcoding using data in the bit stream, then again after coding the texture with the new set of QPs, Q' . With this increased number of data points, we can expect the model parameters to be more robust and converge faster.

The main objective of our texture analysis is choose Q' which satisfy the rate constraint and minimize distortion. However, it is important to note that optimality is conditioned on Q . Therefore, we must take care in how the distortion is quantified. More specifically, we must consider the *distortion increase* with respect to the operating point defined by the incoming Q .

One way to compute Q' is to utilize the same methodology as used in the rate control problem. In this way, we would first estimate a budget for all VOPs at a particular time instant, adjust the target to account for the current level of the buffer, then distribute this sum of bits to each object. Given these object-based target bit rates, the new set of QPs can be computed with our texture model. The main problem with this approach is that we rely on the distribution of bits to be robust. In general, the distribution is not robust and the ability to control the increase in distortion is lost, since the new QPs have been computed independent of the original ones.

To overcome this problem and attempt to solve for Q' in some way that is dependent on Q , we propose an algorithm based on a dynamic programming. To maintain as close a quality as possible to the original, it can be argued that the QPs of each object should change as little as possible. Given this, we can express the distortion increase as follows:

$$\Delta D(Q'|Q) = \sum_{k \in \mathcal{K}} \alpha_k [D(Q'_k) - D(Q_k)] \quad (3)$$

where k denotes a VOP index in the set of VOPs $\mathcal{K}$, and α_k represents the visual significance or priority of object k . This significance may be a function of the objects relative size and complexity, or may even reflect the semantic importance if it is known. It is important to note that $Q'_k \geq Q_k, \forall k$, therefore the solution space is bounded by the incoming QP for each object Q_k , and the maximum allowable QP $Q_{\max}$, which is usually 31.

Given the above quantification for the increase in distortion, we can solve our problem using dynamic programming. Formally, the problem can be stated as

$$\min \Delta D(Q'|Q), \quad \text{s.t.} \begin{cases} R_{\text{Total}} \leq R_{\text{Budget}} \\ Q_k \leq Q'_k \leq Q_{\max} \end{cases} \quad (4)$$

This problem is solved by converting the constrained problem into an unconstrained one, where the rate and distortion are merged through a Lagrangian multiplier, λ . That is, we minimize the cost given by $\Delta D + \lambda R_{\text{Total}}$. Since this cost is additive over all objects, it is possible to solve this problem using dynamic programming. To do this, a 2-D lattice is formed as shown in Fig. 2. The vertical axis represents the candidate QPs, Q'_k , while the horizontal axis denotes the object index k . The incoming set of QPs (Q_k) is highlighted with squares and the valid solution space is the shaded area between these values and $Q_{\max}$. Each node (k, Q'_k) is associated with a cost $\Delta D_{k, Q'_k} + \lambda R_{k, Q'_k}$. It is well known that the distortion with a quantizer step size of Q is proportional to Q^2 , therefore the following approximation is used to quantify the increase in distortion: $\Delta D_{k, Q'_k} = Q'^2_k - Q^2_k$. The rate R_{k, Q'_k} for each node is estimated by (1) using the modified DCT measure of encoding complexity given in (2). This measure of DCT encoding complexity is also used to indicate the visual significance of the VOP, i.e.,

$\alpha_k = \tilde{S}_k$. Given this, the optimal solution can easily be found for any $\lambda \geq 0$. To determine the value of λ that satisfies the constraint on the rate, the well-known bisection algorithm can be used [15]. It is important to emphasize that the search space being considered is much less than MPEG-2 transcoding algorithms that attempt to find the best set of quantizers for every macroblock. Here, we only consider object-based quantizers, hence this approach is quite feasible.

C. Frameskip Analysis

Generally speaking, the purpose of skipping frames is to reduce the buffer occupancy level so that buffer overflow and ultimately the loss of packets is prevented. Another reason to skip frames is to allow a tradeoff in spatial versus temporal quality. In this way, fewer frames are coded, but they are coded with higher quality. Consequently, if the buffer is not in danger of overflowing, the decision to skip a frame should be incorporated into the QP selection process.

Building from the proposed technique for QP selection, which searches a valid solution space for a set of QPs, we can achieve this spatial versus temporal tradeoff by constraining the solution space. In this way, the quantizer constraint in (4) would instead be given by

$$Q_k \leq Q'_k \leq \min\{Q_k + \Delta Q_k, Q_{\max}\} \quad (5)$$

where ΔQ_k indicates the maximum allowable change in the new QP to satisfy this new constraints on the spatial quality. When the constraint cannot be met, the transcoder will skip the current set of VOPs, thereby regulating the overall spatio-temporal quality.

As shown in Fig. 2, a valid path is one in which all elements of Q'_k fall in the constrained area. If one of these elements falls outside the area, the path is invalid in that it is not maintaining some specified level of spatial quality, where the spatial quality is implied by the distortion increase given in (3). It should be noted that different criteria may be used to determine the ΔQ_k for a particular object. For instance, the change can be a function of the objects complexity or simply a percentage of the incoming QP.

In the case of choosing the maximum based on complexity, the transcoder is would essentially limit those objects with higher complexity to smaller QP, since their impact on spatial quality would be most severe. On the other hand, limiting the complexity based on the incoming QP would imply that the transcoder would like to maintain a similar QP distribution as compared to the originally encoded bit stream. Both approaches seem valid, however more research is needed to determine the best way to limit the QP for each object. Such tradeoffs in spatial versus temporal quality have not received significant attention in the literature; some studies have been conducted in [16].

Of course, one of the advantages in dealing with object based data is that the temporal quality of one object can be different from another. In this way, bits can be saved by skipping the background object, for example. However, since objects are often disjoint, reducing the temporal resolution of one object may cause holes in the composed scene. This composition problem

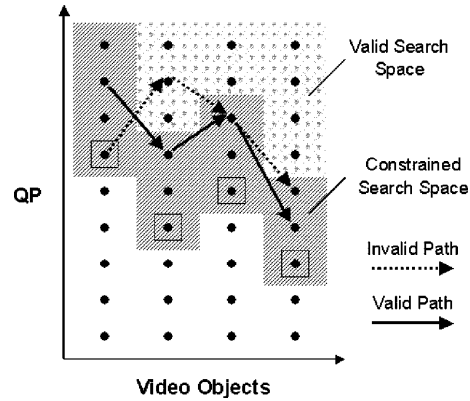


Fig. 2. Illustration of QP search space.

is discussed in more detail in the next section. To solve this problem, a shape hint that indicates the amount of change in the shape boundary over time is proposed. However, the use of such information has not been yet been considered in this framework, but it is surely an interesting topic for further work. Until then, we must impose the constraint that all VOPs have the same temporal resolution.

IV. META-DATA BASED APPROACH

In general, meta-data includes low-level features of the video which describe color, motion, texture and shape, as well as high-level features, which may include storyboard information or the semantics of the various objects in a scene. Of course, not all meta-data is useful for transcoding, but we identify several forms of meta-data that can be used to guide the transcoder. These forms of meta-data are referred to as transcoding hints and are specified in the Committee Draft of the MPEG-7 standard [7]. It is the function of the meta-data analysis to map these hints into the control parameters of the transcoder. The objectives of the transcoding hints that we propose in the following are to: 1) improve the bit allocation among objects; 2) identify key objects in the scene; and 3) regulate the variable temporal resolution of objects in the scene. These tasks will be made possible through a difficulty hint, first proposed by Suzuki and Kuhn in [17], and a shape change and motion intensity hint, first proposed in [9].

The initial motivation of the difficulty hint was to assist the transcoder in performing constant bit-rate (CBR) to variable bit-rate (VBR) conversions. It would do so by providing the transcoder with information related to the complexity of one segment with respect to the complexity of all other segments being considered. In [18], this specific application has been presented in detail; also, several additional motion-based hints have been discussed, including a hint that specifies motion uncompensability and a hint that specifies search range parameters. In this paper, we examine the use of the difficulty hint for improved bit allocation among multiple video objects. Since encoding difficulty is directly related to the distribution of bits among objects, such a hint is expected to yield better results.

Given only the encoding complexity, the task of identifying key objects in the scene would be difficult to do in a reliable

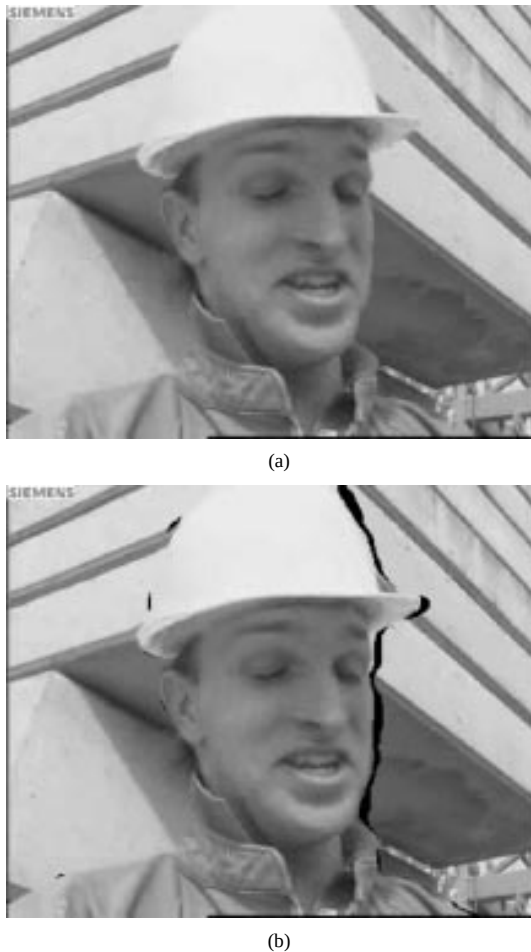


Fig. 3. Illustration of composition problem for encoding multiple objects at different temporal rates. (a) Foreground and background both coded at 30 Hz, no composition problem. (b) Foreground coded at 30 Hz, background coded at 15 Hz, composition problem is shown.

way. The reason is that objects which consume a large number of bits, or even a large number of average bits per block, are not necessarily interesting parts of the scene. To enhance the reliability, we also consider the motion intensity of the objects under consideration.

In the classic transcoding problem, the selection of control parameters implies that we should determine the spatio-temporal tradeoffs that translate into higher PSNR for the scene. However, with object-based transcoding, we allow ourselves to deviate from this traditional notion of quality and begin to consider perceptual quality. In contrast to traditional PSNR measures, this type of measure is more subjective, where the subjective measure is a metric of how well relevant information is being conveyed. Subsequently, it is not only the goal of the meta-data analysis to provide a means of improving the bit allocation, but also to identify key objects in the scene for improved perceptual quality.

A. QP Selection

As discussed in the dynamic programming approach, the process of selecting QPs for each object relies on a model that

defines the rate-quantizer relationship. We again make use of the model defined by $n(1)$ and (2), where the difficulty hint is used to compute the rate budget for each object.

The hint itself is a weight w_i in the range $[0, 1]$, which indicates the relative encoding complexity of each segment. A segment may be defined as a single object or a group of the same object over a specified interval of time. Defining the segment as a group of objects is advantageous to the transcoder in that future encoding complexity can be accounted for. To compute the weights for each object, the actual number of bits used for encoding each object over the specified interval of time are recorded. The weights are then obtained by normalizing the bits used for each object with respect to the total bits used for all objects (over the specified interval of time). Moreover, given that N objects are being considered, the rate for the i th object at time t is simply given by $R_i(t) = w_i \cdot R(t)$, where $R(t)$ is a portion of the total bit rate that has been allocated to the current set of objects at time t . Currently, $R(t)$ is equally distributed across all coding times. However, with the difficulty hint, unequal distribution may provide a means to improve the temporal allocation of bits. However, at this point, we are only interested in the spatial allocation among objects.

It should be noted that the difficulty hint is computed based on the actual number of bits spent during encoding. Although the incoming bits to the transcoder may be sufficient, quantization may hide the true encoding complexity. Therefore, when possible, the difficulty hint should be computed from encoding at fine QPs. The semantics of the MPEG-7 description scheme do not make this mandatory, but it should be noted that results may vary.

B. Key Object Identification

In recent years, key frame extraction has been a fundamental tool for research on video summarization (see, for example, [19]). The primary objective of summarization is to present the user with a consolidated view of the content. The requirement is simple: give the user a general idea of the contents that can be viewed in a short amount of time. In this paper, we would like to extend such concepts to the transcoder. Time constraints for summarization can be likened to bandwidth constraints for the transcoder. Given this, the transcoder should adapt its delivery so that the content is reduced to the most relevant objects.

In this consolidation process, it is important that the transcoder maintain as much information as possible. It can be argued that poorly coded content conveys less information than high-quality content. Consequently, there is a tradeoff in the quality of the content versus the content that is delivered. This is a new concept that does not exist in summarization since summarization is mainly focused on high-quality frames only.

As one can now imagine, there is scope for a great amount of work in this area centered around the tradeoff in delivered content versus the quality of the content, i.e., transmit more content with lower quality or less content with higher quality. In this tradeoff, the transcoder should consider many variations in the spatial and temporal resolution of each object. Also, some measure would need to be devised to handle the decision-making process. As a first step toward such a measure, a measure of "fidelity" has been introduced by those working on summarization

[20]. It is expected that the transcoding framework would need to extend such ideas to incorporate R-D considerations.

In this paper, we introduce a simplified perspective to justify these broad concepts. Specifically, we consider a binary decision on each object, whether it is sent or not. To determine the importance of each object, a measure based on the intensity of motion activity σ and bit complexity B is proposed

$$\eta = (\sigma + c) \cdot \frac{B}{\gamma} \quad (6)$$

where $c > 0$ is a constant to allow zero motion objects to be decided based on the amount of bits, and γ is a normalizing factor that accounts for object size. Larger values of η indicate objects of greater significance. The intensity of motion activity is defined by MPEG-7 as the standard deviation of motion vector magnitudes [6], [21].

C. Variable Temporal Resolution

In this subsection, we illustrate the composition problem encountered in the encoding and transcoding of multiple video objects with different temporal resolutions. We then touch briefly on measures that indicate the amount of change in the shape boundary. Finally, we discuss the proposed shape hint. The uses of this hint are reserved until the next section when simulation results are discussed.

To motivate our reasons for considering a shape hint, the composition problem is discussed. The foreman sequence is shown in Fig. 3. This sequence has two objects, a foreground and background. The left image shows the decoded and composited sequence for the case when both objects are encoded at 30 Hz. The right image shows the decoded and composited sequence for the case when the foreground is encoded at 30 Hz and the background at 15 Hz.

When these two objects are encoded at the same temporal resolution, there is no problem with object composition during image reconstruction in the receiver, i.e., all pixels in the reconstructed scene are defined. However, a problem occurs when the objects are encoded at different temporal resolutions. When the shape of the object has movement in the scene and these object are being encoded at different rates, it causes undefined pixels (or holes) in the reconstructed frames. These holes are due to the movement of one object, without the updating of adjacent or overlapping objects. The holes are uncovered areas of the scene that cannot be associated with either object and for which no pixels are defined. The holes disappear when the objects are resynchronized, i.e., background and foreground are coded at the same time instant.

We now shift our discussion to measures that defines the shape change or distortion between the shape of an object through time. This measure could serve as an input to control the temporal resolution of objects.

Below, two shape-distortion measures that have been proposed in [22] for key frame extraction are reviewed. The distortion is measured from one frame to the next; however, such shape metrics can also relate to the scene at various other levels, e.g., at the segment level over some defined period of time.

The first difference measure that is considered is the well-known Hamming distance, which measures the number of dif-

ferent pixels between two shapes. As defined in [22], the Hamming distance d is given by

$$d = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \|\alpha_1(m, n) - \alpha_2(m, n)\| \quad (7)$$

where $\alpha_1(m, n)$ and $\alpha_2(m, n)$ are corresponding segmentation planes at different time instances. The second shape difference measure is the Hausdorff distance, which is defined as the maxmin function between two sets of pixel

$$h(A, B) = \max\{\min\{d(a, b)\}\} \quad (8)$$

where a and b are pixels of the sets A and B of two video objects respectively, and $d(a, b)$ is the Euclidean distance between these points. The above metric indicates the maximum distance of the pixels in set A to the nearest pixel in set B . Because this metric is not symmetric, i.e., $h(A, B)$ may not be equal to $h(B, A)$, a more general definition is given by

$$H(A, B) = \max\{h(A, B), h(B, A)\}. \quad (9)$$

It should be noted that the above measures are taken between corresponding shapes at different time instants. For a fixed amount of time, small differences indicate that greater variations in the temporal resolution for each object could be used, whereas larger differences indicate smaller variations. If the duration of time is made larger and the differences remain small, then this indicates a potential for savings through variations in the temporal quality.

To derive the proposed shape hint from the distortion measure described above, these measures must be normalized in the range $[0, 1]$. For the Hamming distance, we may normalize the distortion by the number of pixels in the object. A value of zero would indicate no change, while a value of one would indicate that the object is moving very fast. If the Hausdorff distance is used, we may normalize with respect to the maximum width or height of the object's bounding box or of the frame. Due to the nonsymmetric nature of the Hausdorff distance, it may also be useful in to distinguish differences between a growing shape boundary and a reducing shape boundary, i.e., a shape boundary may grow as the result of a zoom in and reduce as the result of a zoom out. With regard to the composition problem, shape boundaries that grow may be treated differently than those that reduce.

We should note that these difference measures are most accurate when computed in the pixel-domain, however, as in [22], approximated data from the compressed-domain can also be used. The pixel-domain data are readily available in the encoder, but for the transcoder, it may not be computationally feasible (or of interest) to decode the shape data. Rather, the macroblock coding modes (e.g., transparent, opaque, border) could be used. In this case, a macroblock-level shape boundary could be formed and the distortion between shape can be computed.

In summary, there are many possibilities for using the proposed shape hint. Among the uses are:

- 1) the transcoder can determine time instances when it is feasible to vary the temporal resolution among objects

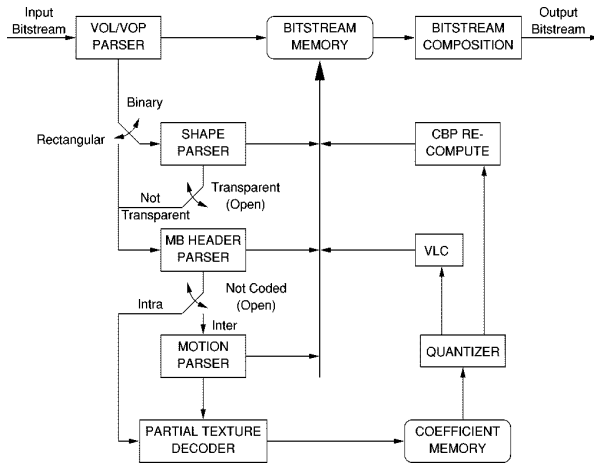


Fig. 4. Architecture to transcode an MPEG-4 object.

to avoid composition problems, i.e., when all shape hints show low movement;

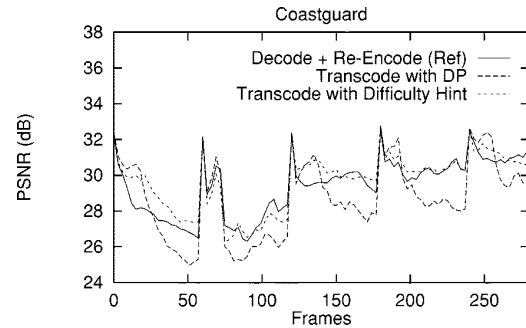
- 2) the shape hint can assist in computing the temporal rate for individual objects, i.e., large difference between a 30- and 10-Hz shape hint could indicate limitations in temporal rate reduction.
- 3) Better bit allocation, since the transcoder is taking full advantage of spatio-temporal redundancy, i.e., coding objects at lower temporal resolution, gives more bits to other objects.

V. TRANSCODER ARCHITECTURE

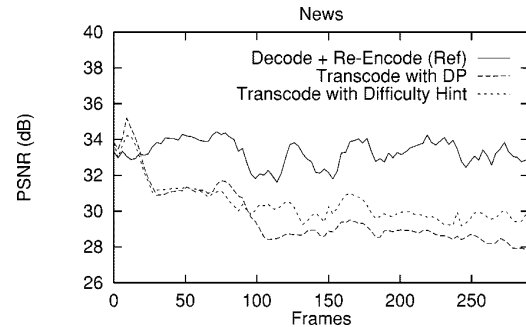
The purpose of this section is to relate the conceptual object-based transcoding system of Fig. 1 with an MPEG-4 transcoding architecture that more closely resembles the flow of data and control information. This architecture is shown in Fig. 4 and resembles an open-loop scheme, which is used in the simulations. It is straightforward to extend this open-loop scheme to a closed one, where the drift in simulations would be significantly decreased.

As with previous transcoding architectures, the syntax of the standard dictates the architecture. One high-level difference is that the bit streams for each object are independent of other bit streams. As a result, each object is associated with a video object layer (VOL) and VOP header. The VOP header contains the quantization parameter (QP) that the object was encoded with. The QP for each object is later used in the modeling and analysis of the texture information. All other bits are stored until the bit stream is recomposed. During the bit stream recombination, information about the temporal rate is received. If a VOP should be skipped or the temporal rate is changed, bits related to the skipped VOP will not be part of the outgoing bit stream. Additionally, fields in the VOL/VOP header that convey information about the timing and referencing may also need to change.

The most significant difference from other standards is that MPEG-4 is capable of coding the shape of an object. From the VOP layer, we find out whether the VOP contains shape information (binary) or not (rectangular). If it is a rectangular VOP,



(a)



(b)

Fig. 5. PSNR comparing reference and transcoded frames using dynamic programming (DP) approach and meta-data based approach with difficulty hints. (a) News, 256 kbits/s $\rightarrow$ 128 kbits/s, 10 fps, $N = 300$, $M = 1$. (b) Coastguard, 512 $\rightarrow$ 384 kbits/s, 10 fps, $N = 20$, $M = 1$.

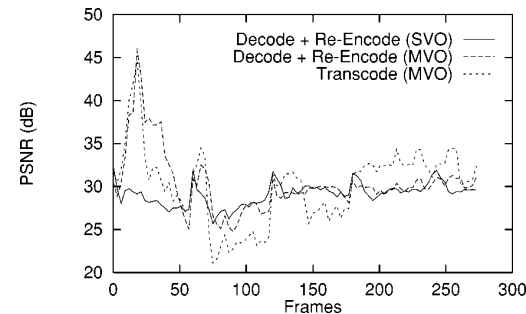


Fig. 6. PSNR comparing SVO and MVO references versus MVO transcoded results for one object in Coastguard. Transcoded results are obtained with the DP approach.

then the object is simply a rectangular frame and there is no need to parse shape bits. In the case of binary shape, we need to determine if the macroblock is transparent or not. Transparent blocks are within the bounding box of the object, but are outside the object boundary, so there is no motion or texture information associated with it.

The remainder of the MPEG-4 syntax is somewhat similar to that of MPEG-2 with a few exceptions. At the MB layer, there exist bits that contain the coded block pattern (CBP). The CBP is used to signal to the decoder which blocks of a macroblock contain at least one ac coefficient. Not only does it affect the structure of the bit stream, but it also has an impact on intra ac/dc prediction. The reason that the transcoder must be concerned

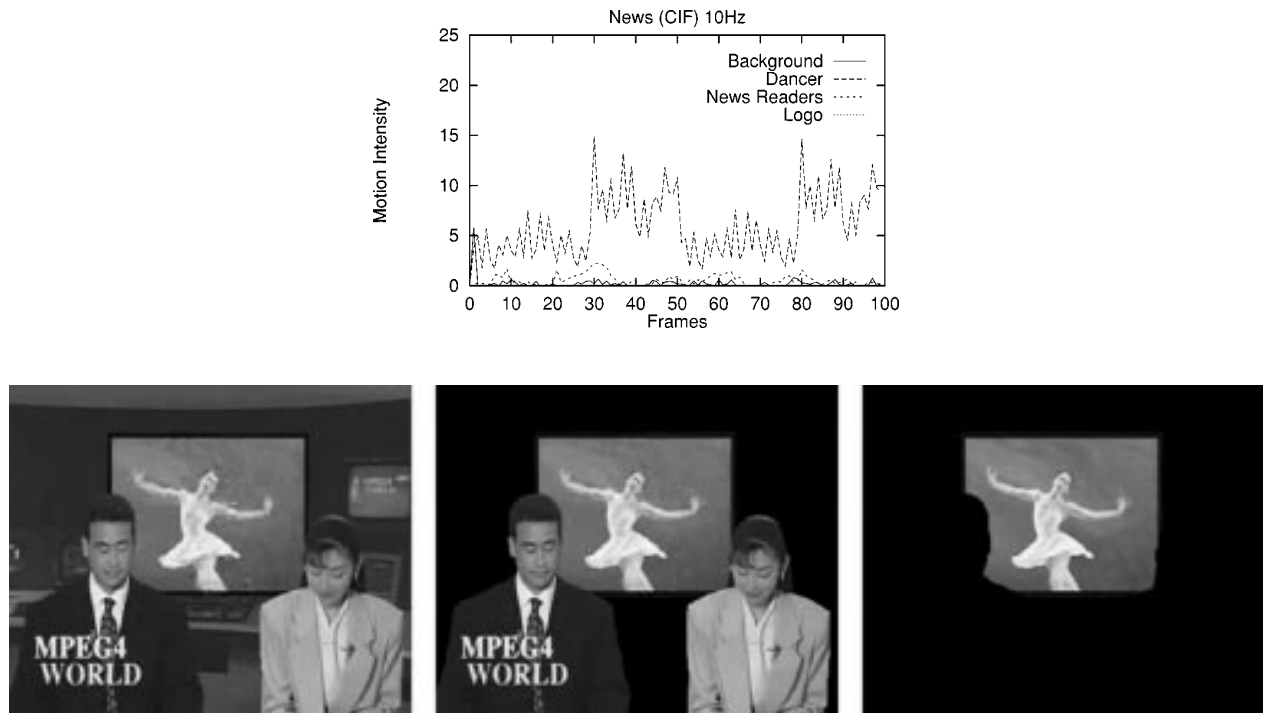


Fig. 7. Illustration of adaptable transcoding of News sequence with key object identification, $384 \rightarrow 128$ kbits/s. The motion intensity and difficulty hints are used to order the importance of each object and the difficulty hint alone is used for bit allocation. Fewer objects can be delivered with noticeable higher quality.

with this parameter is because the CBP will change accordingly with the requantization of DCT blocks. For this reason, we must recompute the CBP.

VI. SIMULATION RESULTS

The results presented in this section demonstrate the various ways that object-based bit streams can be transcoded. First, we provide experimental results for the bit allocation among objects. Next, we show results using key object identification. Then, we discuss the potential use of shape hints for a variety of test sequences. Finally, experimental results that overcome the composition problem and yield gains in coding efficiency are shown.

A. Bit Allocation Among Objects

To study bit allocation in the transcoder, several results are simulated with the Coastguard and News sequences. Both sequences are CIF resolution (352×288), contain four objects each, and are 300 frames long. For all simulations, the original bit stream is encoded at a rate of R_{in} . The rate-control algorithm used for encoding is the one reported in [14] and implemented in [23]. To achieve the reduced rate of R_{out} , three methods are compared. The first method is a brute-force approach that simply decodes the original multiple video object (MVO) bit streams and re-encodes the decoded sequence at the outgoing rate. Recall that this method of transcoding is complex due to the added CPU to perform motion estimation and added frame

memory, but it serves as our reference. The next two methods have similar complexity (much lower than the reference) and make use of the same architecture. Actually, they only differ in the algorithm used to select the QP for each object, so the comparison is very fair. One method simulates the dynamic programming approach presented in Section III-B, while the other method simulates the meta-data based approach that uses difficulty hints, as presented in Section IV-A.

A plot of the PSNR curves comparing the reference method and two transcoding methods is shown in Fig. 5. The Coastguard sequence is simulated with $R_{in} = 512$ kbits/s, $R_{out} = 384$ kbits/s, a frame rate of 10 fps, and GOP parameters $N = 20$ and $M = 1$. Similarly, the News sequences was simulated with $R_{in} = 256$ kbits/s, $R_{out} = 128$ kbits/s, a frame rate of 10 fps, and GOP parameters $N = 300$ and $M = 1$. It is quite evident from the plots that the transcoding method that using difficulty hints is always better than the DP-based approach. This is expected since the difficulty hint can provide information about the scene that is not easy (if not impossible) to obtain. For one, the difficulty hint is computed based on the bit usage (complexity) that is obtained from a fine quantizer, e.g., $Q = 1$. Also, the difficulty hint reflects the bit usage of future frames. In our experiments, the difficulty hint was computed regularly over a period of five frames. We tested the algorithm with different values, e.g., 30 frames or two frames, but the results did not seem to change much. Better results could be expected if the difficulty hint were computed on an irregular basis, where the interval boundaries would be determined in such a way that the difficulty is more consistent over the segment. With more consistency, the hint would not smooth out highs and lows in the bit

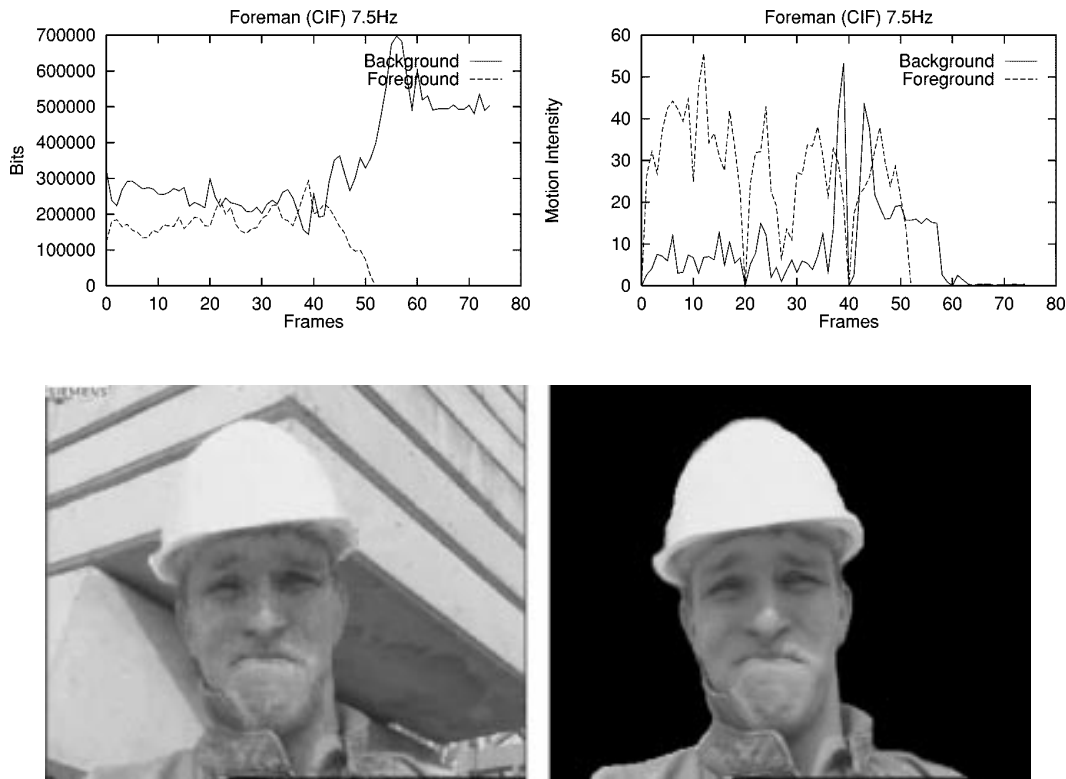


Fig. 8. Illustration of adaptable transcoding of Foreman sequence with key object identification, 256 $\rightarrow$ 192 kbits/s. The motion intensity and difficulty hints are used to order the importance of each object and the difficulty hint alone is used for bit allocation. Fewer objects can be delivered with noticeable higher quality.

usage, resulting in a more accurate description of difficulty for each segment of the program.

From the plots in Fig. 5, a number of other things are worth discussing. First, for the Coastguard sequence, the transcoded results using the difficulty hint provide a very close match to the reference method. In fact, for many frames it is even slightly better. This indicates that with the use of noncausal information, it is possible to overcome some of the major drawbacks that conventional low-complexity open-loop transcoders suffer from: mainly drift. Speaking of drift, it is clear from both sequences that it is significantly reduced in comparison to the DP-based approach. The News simulation shows an extreme coding situation in which only 1 I-VOP is used at the beginning of each coded sequence. In this case, the drift can clearly be seen. As reported in earlier works [2], [3], using a closed-loop transcoding architecture could also minimize the effects of drift. In [24], an additional study regarding the cause of drift is presented. Any of these architectures and techniques could be applied here to achieve further gains. We present our results on the simple open-loop system so that gains are easily seen and insight regarding the bit allocation among objects is not clouded by other factors. Overall, the DP-based approach, does demonstrate comparable results to the reference method with a significant reduction in complexity. Furthermore, when the difficulty hint is available, further gains over the DP-based approach can be achieved.

As an additional test, we would like to compare object-based transcoding to frame-based transcoding. To do so, we consider

a single video object (SVO) simulation of the Coastguard sequence. The SVO reference, i.e., decoded and re-encoded, is compared to the MVO simulations described above, with all simulation parameters equal. However, rather than plotting the PSNR over the entire frame, we only consider the PSNR for the primary object in the scene—the large boat. Both the alpha plane for this large boat and the PSNR plots are shown in Fig. 6. It can be seen from these plots that the PSNR of the object as it first enters the scene is dramatically increased with the MVO simulations, and on average better than the SVO result. The MVO transcoded result that is plotted in this figure obtained with the DP-based approach. The main points that these results support are: 1) that it is possible (even with the DP-based approach) to improve the quality in more relevant parts of the scene and 2) the quality of MVO simulations can outperform SVO simulations. Of course, only one case is illustrated here, but these results encourage further study of this topic.

B. Results with Key Object Identification

To demonstrate how key object identification can be used to adapt content delivery and balance quality, the News sequence is considered at CIF resolution, 10 fps. The News sequence contains four objects of varying complexity and of distinctly varying interest to the user. The sequence was originally encoded at 384 kbits/s, then transcoded down to 128 kbits/s.

Fig. 7 illustrates several transcoded versions of frame 46 from the News sequence. The versions differ in the number of objects that are actually transmitted. In the left-most version, all

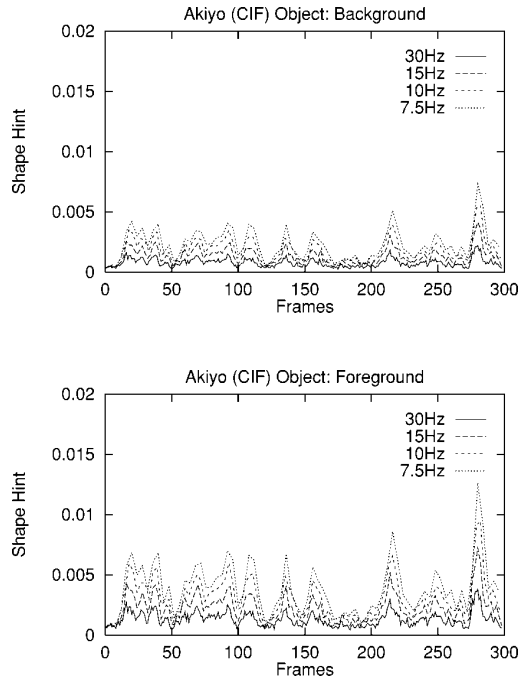


Fig. 9. Shape hints for Akiyo.

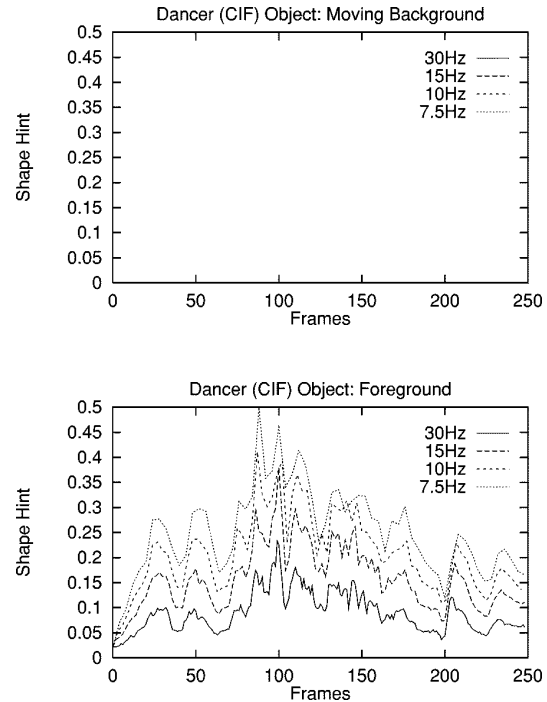


Fig. 10. Shape hints for Dancer.

objects are transmitted with relatively low quality, whereas in the right-most version, only one object is transmitted with relatively high quality. The center image lies somewhere between the other two, in that only the background is not sent. It can be considered a balance between the two in terms of quality and content. For all versions, the objects are identified using the computed measure of importance based on motion and difficulty hints, while the bit allocation for objects under consideration were done using the difficulty hint only.

Similar results are also obtained for the Foreman sequence, as shown in Fig. 8. The sequence was originally coded at a rate of 256 kbits/s and transcoded down to 192 kbits/s. Here, since there are only two objects, the choice is much simpler. Again, based on the motion intensity and difficulty hints, the relevant object is identified. It should be observed from the figure that the increase in quality is quite clear, as all the additional bits are used for the foreground.

Of course, this method of identifying key objects will not be suitable for all scenes, especially ones in which the nature of the content is not known and meaningful segmentation is difficult to achieve. However, for certain scenes that can be segmented with some confidence, possibly using some a priori information about the contents of the scene, this added flexibility of dropping objects from the scene can be very useful to improve the quality of delivered content. A new optimization problem can also be considered. This optimization can be viewed as an extension of the bit allocation problem over N objects, where the N objects must be selected from a total of $M \geq N$ objects. Choosing the best set of N objects is the added dimension. However, this could be difficult to formulate since there is a subjective element to choosing/excluding objects. Also, bit-to-bit differences with the original scene are meaningless in the absence of one or more objects.

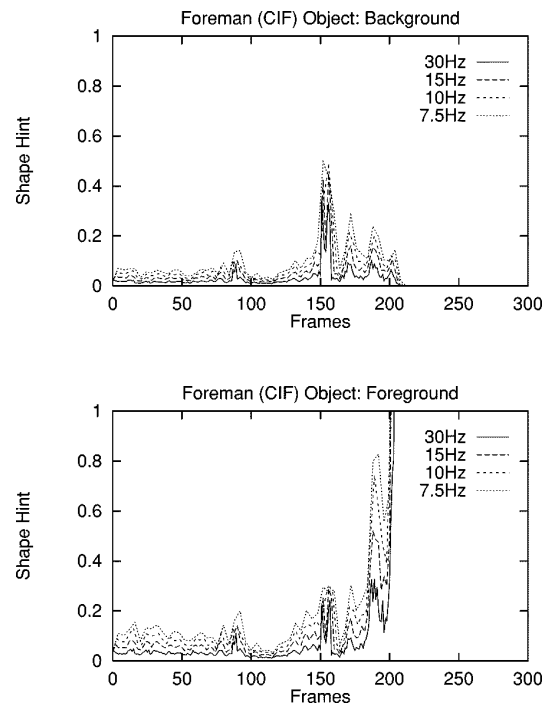


Fig. 11. Shape hints for Foreman.

C. Discussion of Shape Hints

The shape hints are extracted for each object in a number of sequences. For each of the sequences, we interpret how the shape hint could be used by a transcoder. In many cases, the final

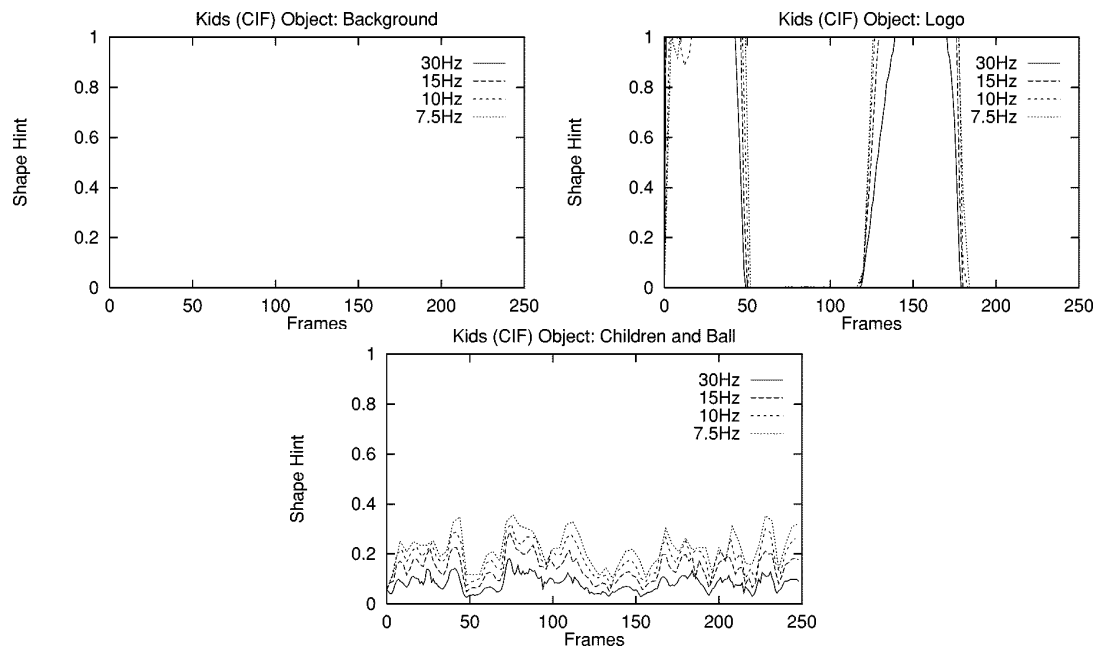


Fig. 12. Shape hints for Children.

outcome and transcoder decision is based on the motion or difficulty of the objects as well. The plots of the shape hint are given for every frame and across a number of frame-rates. However, in practice and according to the current draft of the MPEG-7 standard, the shape hint will only be specified for one frame-rate, which is equal to the frame-rate of the content to be encoded or transcoded. Also, the shape hint can be specified over a period of time, where the average of the shape hints at each time may be used. Finally, it should be noted that the differences in the shape hints between different frame rates are for illustrative purposes only to demonstrate that the shape hints yield fairly consistent results independent of the input frame rate. However, further examination may show that the differences in shape hints across frame rates could be a useful metric to decide on the reduced temporal rates for each object.

- 1) *Shape Hints for Akiyo (Fig. 9):* The shape hints for this sequence indicate very low movement of objects. This would be a very good candidate for variable temporal reduction. Depending on the motion and/or difficulty, a transcoder may reduce the resolution of either object.
- 2) *Shape Hints for Dancer (Fig. 10):* The shape hint for the background indicates no change (it is a full rectangular frame), while the foreground object is experiencing large movement in shape. Is it possible to have variable resolutions with this scene, since the full background is coded and it will never cause holes, but it is probably not a good idea since the shape movement of the foreground indicates very large movement.
- 3) *Shape Hint for Foreman (Fig. 11):* The movement of the background and foreground agree with each other and show large movement for the first 200 frames. After that, the shape hint indicates that the foreground no longer exists and the background eventually fills the entire frame.

- 4) *Shape Hint for Children (Fig. 12):* As with the Dancer sequence, the background is taking up the entire rectangular frame and is unchanging. From the shape hint of the logo, we see that it has extremely fast movement (or is not in the scene), then sits still for a while, then moves very quickly again and sits still until the end. This is a very good example where the shape hint can be used to identify key frames to code or identify localized segments where the temporal resolution can be dramatically reduced. The final object, the children playing with the ball, has very large movements throughout, as indicated by its shape hints.
- 5) *Shape Hints for News (Fig. 13):* Overall, the shape hints for this sequence indicate very low movement. There is one place where one of the newsreaders nods her head; this is picked up between frames 75 and 100. But, for the most part, variable temporal resolution can be tolerated for this sequence. In fact, the logo object does not move at all and can be coded once at the very start, without interfering with other objects.

D. Results with Varying Temporal Resolution

Given that the shape hint will identify scenes (or segments) that can take advantage of variable temporal resolution, we provide results to demonstrate the gain. In the previous subsection, the Akiyo sequence has been identified as a scene that can be encoded or transcoded with variable temporal resolution, without creating disturbing artifacts due to composition. To examine this further, we generate the R-D curves for two cases. The first case is our reference and simply encodes both foreground and background objects at 30 Hz with a constant QP. This is referred to as the fixed case. As a comparison, we maintain the temporal resolution of the foreground at 30 Hz, but reduce the temporal

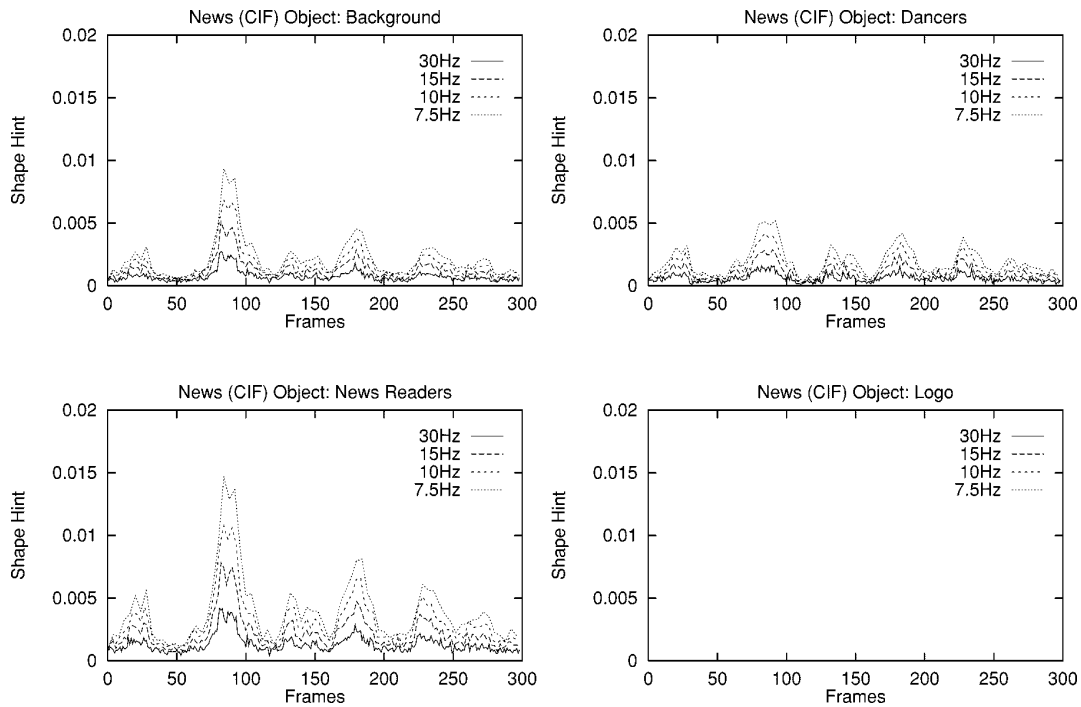


Fig. 13. Shape hints for News.

resolution of the background to 7.5 Hz. This case is referred to as the variable case.

A comparison of the fixed and variable R-D curves is shown in Fig. 14. The set of QPs that were used to generate the curves were $Q = 9, 15, 21$, and 27 . The finest QP was 9 since reducing the temporal resolution for high bandwidth simulations is not of much interest. It should be emphasized that the PSNR for all simulations are computed with respect to the original 30-Hz sequence. The curves in this figure show that a 25% reduction in bit rate can be achieved by reducing the resolution of the background only, with no loss of quality. As indicated by the shape hints for this sequence, the movement among shape boundaries is very small. Consequently, any holes in the reconstructed image are negligible and do not impact the PSNR values.

Due to limitations in space, it is not feasible to test the various ways that the content can be encoded and transcoded with variable temporal resolution. However, we should mention that the indication of the shape hint to allow variable temporal resolution without holes may not always provide a benefit. For example, the background of the Children sequence does not contain any motion and is a full rectangular frame, so the border of other objects does not affect it. In this case, you can send this background once in the beginning; however, the encoder would not spend many bits on the updates anyway, i.e., it would only use 1-bit per macroblock to signal that a block is skipped. As a general rule, gains will be possible if there is sufficient information that is being coded, but does not need to be. We do however expect gains in other sequences besides Akiyo, such as News and others like it.

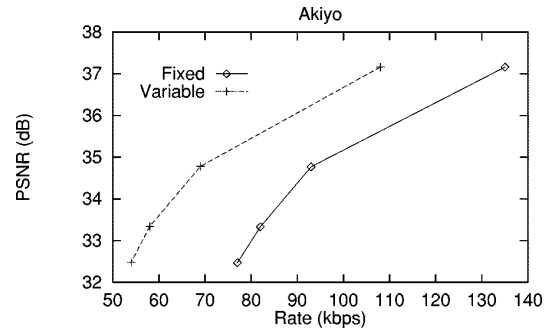


Fig. 14. R-D curves comparing Akiyo sequence coded for fixed and variable temporal resolution. With fixed temporal resolution, both foreground and background are coded at 30 Hz. With variable temporal resolution the foreground is still coded at 30 Hz, but the background is coded at 7.5 Hz.

The gains discussed above use traditional PSNR measures as a major part of the evaluation. Such measures are suitable to test the gains in variable temporal resolution when the movement of the object is low. The reason is that comparisons are made with respect to the original sequence, and so losses between a 30-Hz reference and a 7.5-Hz object rendered at 30 Hz remain small. However, for faster moving objects, PSNR is not a suitable measure to quantify the differences between the same object coded at different temporal resolutions. The bit-to-bit differences will seem large, while the perceptual loss may only be noticeable to a very trained eye. For this reason, measures to evaluate the quality of objects at different temporal resolutions are needed. With such measures, the gains in varying the temporal resolution of objects will become more evident.

VII. CONCLUSIONS AND FUTURE WORK

This paper introduced a new object-based transcoding framework. Within this framework, two distinct transcoding strategies were introduced: a conservative and an aggressive strategy. Transcoding techniques that support these two strategies were proposed and the flexibility of object-based transcoding for adaptable delivery of content has been shown.

The first approach presented was a dynamic programming based approach that selected the QP for each object based on an estimate of the distortion increase and rate budget. In the context of this approach, the similarities and differences between object-based encoding and transcoding were discussed. We focused specifically on the texture model used and different information that is available to each. Since the transcoder only has access to partially decoded information, a new DCT-based complexity measure was defined to model the texture. To consider spatial versus temporal tradeoffs, we discussed how the search space could be constrained.

The second approach that we discussed is an approach that is based on the availability of meta-data. We identified several transcoding hints that could be used for different purposes. The difficulty hint was shown to be useful for the allocation of bits among the multiple objects in the scene. Then, using difficulty and motion, it was shown that key objects could be identified and coding gains for the most relevant objects can be achieved. Finally, with the shape hint, we discussed several ways to use the shape hint for adaptive content delivery, then gave an example of the gains that can be achieved.

As far as future work goes, we feel that there are still many open issues within the object-based framework. Certainly, we have not fully exploited all of the degrees of freedom that are possible, but we hope to at least raise some of the issues here and encourage future work in this area.

One area for great potential is to incorporate the ability to resize objects. For mobile appliances with limited display size, this will surely be a necessary technique and combining this with some of the approaches presented in this paper would also be of great interest. However, even for appliances with sufficient display size, it is still of interest to resize objects before transmission, then reconstruct the original resolution on the receiver side. In the past, this has been difficult to do, since the transcoder should conform to a standard coding scheme. By reducing the resolution, e.g., from CIF to QCIF, one would experience problems with the resampling of motion vectors and regaining MB structure, among other things. However, with the adoption of the Dynamic Resolution Conversion (DRC) tool into Version 2 of the MPEG-4 standard [25], it becomes possible to further exploit gains that would result in down-sizing a background object, for example. The tradeoffs between QP selection, temporal rate, and object resizing would require further investigation. Most likely, a multiresolution texture model would be needed to estimate impacts on the rate and distortion.

Another avenue of future work is to identify other transcoding hints that improve the bit allocation or provide a new set of transcoding functionality. However, even with the existing hints, there is still scope to optimize the use of them. For example, it would be interesting to examine if irregular partitions, or in-

tervals of time, for computing and applying the difficulty hint would lead to improved results. Also, we feel that the potential to encode and transcode with variable temporal resolution has not yet been fully exploited.

Finally, a very challenging problem is to extend the measure of fidelity that has been defined by those working on summarization to include R-D considerations. This is most applicable to transcoders that operate with an aggressive transcoding strategy. Given such a measure, a robust control mechanism can be designed to account for the wide range of transcoding options. Such a measure is consistent with the content-aware framework proposed by Chang and Boeckel [26]. In this framework, features of the objects are used to generate utility functions and classify objects into utility classes. Given these functions and classes, various scaling operations can be applied to the content according to current network behavior. Being able to make such predictions allow us to eventually improve upon the problem of dynamic resource allocation, which has been studied in [27] within the context of an end-to-end solution.

REFERENCES

- [1] *Information Technology—Coding of Audio/Visual Objects*, ISO/IEC 14496-2:1999.
- [2] H. Sun, W. Kwok, and J. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 191–199, Apr. 1996.
- [3] P. Assuncao and M. Ghanbari, "A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bitstreams," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 953–967, Dec. 1998.
- [4] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats," *IEEE Trans. Multimedia*, vol. 2, pp. 101–110, June 2000.
- [5] MPEG Requirements Group, Seoul, Korea, "Overview of the MPEG-7 Standard," ISO/IEC N2729, La Baule, France, Oct. 2000.
- [6] *Information Technology—Multimedia Content Description Interface*, ISO/IEC 15938-3:CD, Oct. 2000.
- [7] *Multimedia Description Schemes, Information Technology—Multimedia Content Description Interface*, ISO/IEC 15938-5:CD, Oct. 2000.
- [8] R. Mohan, J. R. Smith, and C.-S. Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, pp. 104–114, Mar. 1999.
- [9] A. Vetro, H. Sun, and A. Divakaran, "Adaptive object-based transcoding using shape and motion-based hints," Geneva, Switzerland, ISO/IEC M6088, May 2000.
- [10] A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for scalable quality of service," in *Proc. IEEE Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000.
- [11] H. Sorial, W. E. Lynch, and A. Vincent, "Joint transcoding of multiple MPEG video bitstreams," in *Proc. Int. Symp. Circuits and Systems*, Orlando, FL, May 1999.
- [12] S. Gopalakrishnan, D. Reininger, and M. Ott, "Realtime MPEG system stream transcoder for heterogeneous networks," in *Proc. Packet Video Workshop*, New York, Apr. 1999.
- [13] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 246–250, Feb. 1997.
- [14] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 186–199, Feb. 1999.
- [15] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Image Processing*, vol. 2, pp. 160–175, Apr. 1993.
- [16] F. C. Martins, W. Ding, and E. Feig, "Joint control of spatial quantization and temporal sampling for very low bit rate video," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, May 1996, pp. 2072–2075.

- [17] T. Suzuki and P. Kuhn, "A Proposal for Segment-Based Transcoding Hints," Noordwijkerhout, Netherlands, ISO/IEC M5847, Mar. 2000.
- [18] P. Kuhn and T. Suzuki, "MPEG-7 metadata for video transcoding: Motion and difficulty hint," in *Proc. SPIE Conf. Storage and Retrieval for Multimedia Databases*, San Jose, CA, Jan. 2001.
- [19] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1280–1289, Dec. 1999.
- [20] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1269–1279, Dec. 1999.
- [21] A. Divakaran and H. Sun, "A descriptor for spatial distribution of motion activity," in *Proc. SPIE Conf. Storage and Retrieval from Image and Video Databases*, San Jose, CA, Jan. 2000.
- [22] B. Erol and F. Kossentini, "Automatic key video object plane selection using shape information in the MPEG-4 compressed-domain," *IEEE Trans. Multimedia*, vol. 2, pp. 129–138, June 2000.
- [23] *Information Technology—Coding of Audio/Visual Objects*, ISO/IEC 14496-5:2000, 2000.
- [24] J. Youn, J. Xin, and M.-T. Sun, "Fast video transcoding architectures for networked multimedia applications," in *Proc. IEEE Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000.
- [25] MPEG Video Group, "MPEG-4 Video Verification Model (v14.0)," ISO/IEC N2932, Melbourne, Australia, Oct. 1999.
- [26] S. F. Chang and P. Boeck, "Principles and applications of content-aware video communications," in *Proc. IEEE Int. Symp. Circuits and Syst.*, Geneva, Switzerland, May 2000.
- [27] Q. Zhang, Y. Q. Zhang, and W. Zhu, "Resource allocation for audio and video streaming over the Internet," in *Proc. IEEE Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000.
- [28] T. Ebrahimi and C. Christopoulos, "Can MPEG-7 be used beyond database applications?," ISO/IEC M3861, Atlantic City, NJ, Oct. 1998.
- [29] C. Christopoulos, T. Ebrahimi, V. V. Vinod, J. R. Smith, R. Mohan, and C. S. Li, "MPEG-7 application: Universal access through content repurposing and media conversion," ISO/IEC M4433, Seoul, Korea, Mar. 1999.



Anthony Vetro (S'92–M'96) received the B.S. and M.S. degrees in electrical engineering from Polytechnic University, Brooklyn, NY, in 1996, where he is working toward the Ph.D. degree in electrical engineering.

In 1996, he joined Mitsubishi Electric Research Laboratories, Murray Hill, NJ, and is currently a Principal Member of the Technical Staff. Upon joining Mitsubishi, he worked on algorithms for down-conversion decoding and was later involved in the development of a single-chip HDTV receiver that employed these algorithms. More recently, his work has focused on the encoding and transport of multimedia content and content description, with emphasis on video content scaling, rate-distortion modeling and optimal bit allocation. Since 1997, he has been an active participant in ISO-MPEG, contributing to the development of the MPEG-4 and MPEG-7 standards. He holds three U.S. patents, with several more pending, and has published a number of papers in these areas.



Huifang Sun (S'83–M'85–SM'93–F'01) received the B.S. degree in electrical engineering from Harbin Engineering Institute, Harbin, China, in 1967, and the Ph.D. degree in electrical engineering from the University of Ottawa, Ottawa, Canada, in 1986.

From 1982 to 1986, he was with the Electrical Engineering Department, University of Ottawa. In 1986, he joined Fairleigh Dickinson University, Teaneck, NJ, as an Assistant Professor, and was subsequently promoted to Associate Professor in Electrical Engineering. From 1990 to 1995, he

was with the Sarnoff Corporation (formerly David Sarnoff Research Center), Princeton, NJ, as a Member of the Technical Staff, and was later promoted to Technology Leader of Digital Video Technology, where his activities were MPEG video coding and Grand Alliance HDTV development. He joined Mitsubishi Electric Research Laboratories, Murray Hill, NJ, in 1995, as a Senior Principal Technical Staff, where his work focused on advanced television development and digital-video related projects. In 1997, he was promoted to Deputy Director. He holds 11 U.S. patents and has several pending, and has authored or co-authored more than 100 journal and conference papers.

Dr. Sun currently serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is also a member of the Technical Committee on Visual Signal Processing of the IEEE Circuits and Systems Society.



Yao Wang (S'86–M'90–SM'98) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1983 and 1985, respectively, and the Ph.D. degree in electrical engineering from University of California at Santa Barbara in 1990.

Since 1990, she has been on the faculty of Polytechnic University, Brooklyn, NY, and is presently a Professor of Electrical Engineering. From 1992 to 1996, she was a part-time consultant with AT&T Bell Laboratories, Holmdel, NJ, and has been with AT&T

Labs - Research, Red Bank, NJ, since 1997. She was on sabbatical leave at Princeton University, Princeton, NJ, in 1998. Her current research interests include image and video compression for unreliable networks, motion estimation, object-oriented video coding, signal processing using multimodal information and image reconstruction problems in medical imaging.

Dr. Wang is presently an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA and for the *Journal of Visual Communications and Image Representation*. She has previously served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. She is a member of the Technical Committee on Multimedia Signal Processing of the IEEE Signal Processing Society and the Technical Committee on Visual Signal Processing and Communications of the IEEE Circuits and Systems Society. She has served on the organizing/technical committees of several international conferences and workshops and as Guest Editor for several special issues related to image and video coding. She won the New York City Mayor's Award for Excellence in Science and Technology for the Year 2000 in the young investigator category.