Adaptive Video Transmission Schemes Using MPEG-7 Motion Intensity Descriptor

Osama A. Lotfallah, Martin Reisslein, *Senior Member, IEEE*, and Sethuraman Panchanathan, *Fellow, IEEE*

***Abstract*—A variety of error resilience and scalable coding techniques have recently been proposed to facilitate the delivery of video over best-effort networks; a common drawback of these techniques is reduced compression efficiency. Also, MPEG-7 descriptors have recently been developed for the purpose of indexing. In this paper, we propose to employ MPEG-7 descriptors to improve the quality of the video delivered over best-effort networks. In particular, we propose a video transmission system that uses the motion activity descriptors to ensure robust video transmission. A novel motion activity extraction technique is proposed, which relies on a neural network approach. By considering several low-level visual features, our proposed extraction approach achieves high consistency with subjective evaluations of motion activities. In order to demonstrate the benefits of the proposed transmission system, we develop a selective packet dropping scheme that can be applied in case of network congestion. Simulations demonstrate that the reconstruction quality of the proposed congestion scheme can surpass conventional schemes by 1.2 dB. The network performance of the proposed transmission system when video sequences are coded into single layer or scalable layers is presented. We also present a transcoding scheme that achieves the optimal reconstructed quality by exploiting the motion activities of the underlying video sequence.**

***Index Terms*—Congestion control, motion intensity level, MPEG-7 descriptors, transcoding.**

I. Introduction

THE transmission of compressed visual information over unreliable networks that cannot guarantee timely and lossless data delivery, such as the best-effort Internet and wireless networks, can result in poor reconstructed video quality. To improve the video quality, a variety of error resilience techniques has been developed. Data partitioning, resynchronization markers, and intra-refreshment, for instance, represent the common resilient techniques that are available in the MPEG-4/H.264 encoders [1]–[3] to mitigate the error propagation effect that arises when losses affect the intra-coded frames that are used as references for the encoding of the inter-coded frames. Since error resilience cannot completely overcome the error effect and the excessive use of the resilience tools results in poor compression efficiency, more advanced schemes are necessary. Layered coding is a potential solution that splits the video stream into a base layer and a number of enhancement layers. The base layer can be a reduced quality or small frame size video stream and each enhancement layer improves the quality or resolution of the video. In general, unequal error protection schemes are used in conjunction with the layered coding since the base layer is more important than the enhancement layers [4]–[6]. Unfortunately, such layered coding typically results in a significant reduction of the compression efficiency. We also note that a variety of approaches employing per-packet rate-distortion optimization as well as adaptive buffering and scheduling of video packets have recently been proposed to alleviate congestion, reduce packet losses and, thus, improve the reconstructed video quality, see for instance [7]–[10]. These approaches are complementary to our work in that we propose a novel motion activity extraction technique and demonstrate how the extracted motion level can be used for low-complexity mechanisms that enhance the video transmission system, for instance by adaptively dropping packets according to the motion activity level. We note that there have been some efforts on taking video content into consideration in video communication, see for instance [11], [12]. These works employ low-level visual features that are not consistent with the general purpose descriptors recommended by MPEG-7, which we adhere to in this study. Also, while these low-level visual features are useful for specifying the bandwidth needed to transmit the video stream, they do not take the quality of the reconstructed video streams into account, which we strive to maximize with our schemes.

In this paper, we propose novel techniques for improving the video quality, which do not affect the compression efficiency of the underlying encoder. The basic idea of our approach is that during the transmission over an unreliable network a video experiences typically a high variability in the available (channel) bit rate and in the packet loss ratio. At the same time, the motion intensity in the video scenes typically changes from low to moderate or high motion scenes. Importantly, the effect of packet loss is typically different for video scenes of different motion intensity. For low motion intensity scenes, the loss could be successfully hidden by simple error concealment techniques such as copying [13], while error concealment techniques may not be effective for high motion scenes. Our adaptive transmission schemes judiciously consider *not only the channel conditions but also the visual content.*

Video streaming servers commonly employ MPEG-7 descriptors of their multimedia contents so that users can access the media contents with a search engine. In general, some of

Manuscript received February 11, 2003; revised May 10, 2004; accepted March 7, 2006. The work was upported in part by the National Science Foundation through Grant ANI-0136774. This paper was recommended by Associate Editor J.-R. Ohm.

O. A. Lotfallah and S. Panchanathan are with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: Osama.Lotfallah@asu.edu; oslatif@asu.edu; panch@asu.edu).

M. Reisslein is with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: reisslein@asu.edu).

Digital Object Identifier 10.1109/TCSVT.2006.877387

1051-8215/$20.00 © 2006 IEEE

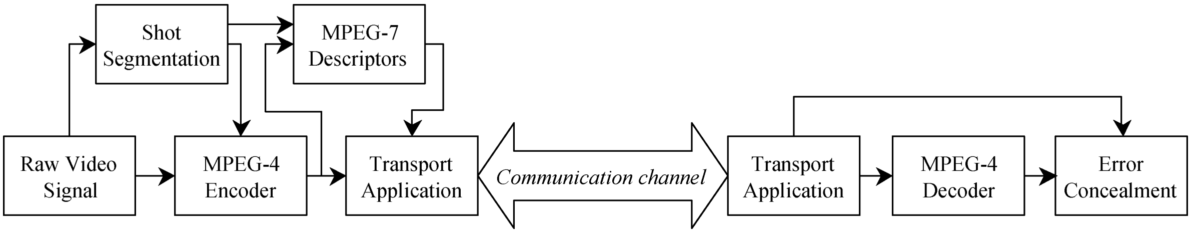


Fig. 1. Framework for proposed video transmission schemes.

MPEG-7 descriptors are highly correlated with the network behavior (e.g., traffic characteristics, loss sensitivity) of the video and media. However, current video transmission schemes do generally not exploit MPEG-7 descriptors. We propose a framework for video transmission that exploits MPEG-7 descriptors for video transmission as shown in Fig. 1. Initially, the video sequence is segmented into a number of shots, which are the smallest logical unit of the video sequence that has a consistent visual content. Subsequently, the original video signal is compressed using MPEG-4 coding schemes that use the shot boundaries to enforce intra-coded frames. Therefore, each video shot is not only independently coded but also randomly accessible. The compressed video streams as well as shot segmentation statistics are used for extracting the MPEG-7 descriptors. The transport application can then adaptively transmit the video according to the current network conditions (channel bandwidth variations) and the visual content of the video. The received video stream is decoded and the impact of any channel losses, which were already kept small by the adaptive transmission, can be mitigated by an appropriate concealment method. The transport application can aid the error concealment by providing information about error location and possibly the underlying MPEG-7 descriptors. One of the benefits of the system architecture depicted in Fig. 1 is it allows for flexible implementation in packet lossy networks, such as the Internet as well as wireless networks.

We consider motion descriptors in this study as motion is a key characterization of the visual content of a shot. Also, shots with different level of motion are more or less amenable to error concealment as noted above, which can be exploited for significantly improving the reconstructed video quality, as demonstrated in this study. The motion descriptors that have currently been selected by MPEG-7 cover an ample range of complexity and functionality and enable MPEG-7 to support a broad range of applications [14]. In this paper, we concentrate on the motion intensity/activity level descriptor. The automatic extraction of the motion descriptors remains a complicated issue because motion is a mixture of object and background (i.e., camera) motions that affect the represented color, shape, and texture features. Therefore, we propose an automatic extraction scheme of the motion intensity/activity level descriptor using an up-to-date artificial neural network. Neural network solutions can achieve high accuracy and also be consistent with human judgments. In order to exploit the extracted MPEG-7 descriptors, a congestion control scheme is proposed that selectively drops packets. We note that the packet prioritization could also be based on nonstandard content descriptors, but we focus on MPEG-7 compliant content descriptors to allow interoperability with other applications, such as video search and retrieval.

This paper is organized as follows. In Section II, we review the existing techniques for extracting the MPEG-7 motion intensity/activity level descriptor of a shot and propose our novel neural network based technique. In Section III, we introduce a packet prioritization scheme that considers the motion activity levels in packet dropping. The proposed congestion control scheme is presented in Section IV and evaluated through simulations. In Section V, we compare the network performance for single-layer and fine granularity scalable (FGS) coding schemes. In Section VI, we examine the effect of motion activity levels in video transcoding applications and introduce a new dimensionality in the transcoding parameters that improve the quality of the received images. Finally, the conclusions are presented in Section VII.

II. Extracting MPEG-7 Motion Intensity Descriptor

Extracting the temporal construction units of a video segment is an essential prerequisite for identifying the features of the underlying visual content. We segment the video into shots, where a shot is defined as a sequence of frames captured by a single camera in a single continuous action in time and space, and for which simple algorithms are available [15], [16]. We employ shot detection algorithms that work in the uncompressed domain, which enables us to code the first frame in every shot as intra-frame. In particular, we combine the color histogram technique [15] and the intensity scaling techniques [17], [18] to detect shots.

In general, a video is composed of a wide range of shots that vary from low to high motion activities [19], [20]. The MPEG-7 standard represents the intensity of the motion activity by an integer lying in the range 1–5 where a high value indicates high activity while a low value indicates low activity [20]. Thus, the five scale levels could be described as: 1) very low intensity; 2) low intensity; 3) medium intensity; 4) high intensity; and 5) very high intensity. In this section, we first briefly review existing schemes for extracting the motion intensity and then introduce our novel extraction approach.

In [21], the macroblock type is used to compute the motion activity levels. More specifically, the ratio of the number of intra-coded macroblocks and macroblock without motion vectors to the total number of macroblocks in a frame is considered. In order to specify the motion activity level between successive frames, this ratio is logarithmically quantized into five levels such that large ratios are assigned to low motion activity levels. Moreover, the distribution of motion intensities inside the

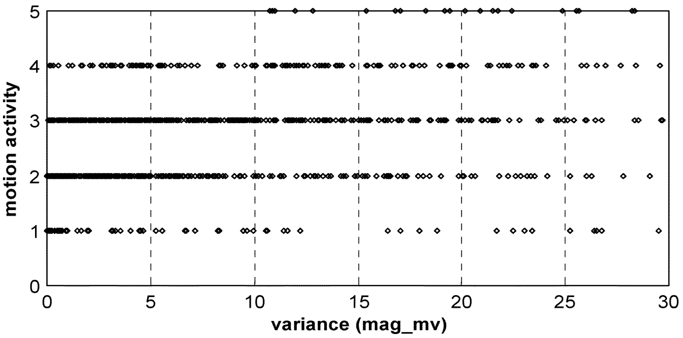


Fig. 2. Distribution of the variance of the motion vector magnitudes of different motion activity level.

shot is expressed as the histogram of the motion activity levels. Therefore, the motion intensity histogram descriptor is represented as $\mathrm{MIH} = [p_1, p_2, p_3, p_4, p_5]$; where p_i is the percentage of P-video object planes (P-VOPs) that correspond to the i^{th} motion activity level [21].

Various statistical analyses of the motion vector magnitude, which is available from the compressed MPEG video, have been proposed for extracting the motion level, see for instance [22], [23]. In these analyses the mean, variance, standard deviation or the maximum of the motion vector magnitudes is quantized into nonuniform ranges corresponding to each motion activity level. Hence, small motion vector magnitudes are matched with low activity level between successive frames. The extension of this procedure to implementation for video shots that contain a number of frames can be achieved by averaging these motion vector magnitudes in the entire shot [23].

Existing procedures for estimating the motion activity level capture the motion activity through analyzing a specific parameter between successive frames. The extracted parameter is a decimal number that maps into a motion activity level. Additionally, the parameter values are divided into five disjoint sets that correspond to the motion activity level. Fig. 2 shows the variance of the motion vector magnitudes for different motion activity levels. The x axis represents the average (over the video shot) of the variance of the motion vector magnitudes (over the video frame). The shot activity levels are based on our ground truth set, which is explained in the following paragraph. We observe from Fig. 2 that splitting the variance of motion vector magnitudes into five disjoint sets results in significant error in specifying the motion activity level. The accuracy of this and other existing techniques is limited because typically only one parameter is used to extract the motion activity level, although motion activities have complex relationships with many visual features. In order to efficiently capture the motion activity level in a shot, *multiple* parameters need to participate in the process. For example, a talk show shot that contains a mixture of background colors can result in high motion vector magnitudes although a human perceives the shot as having low motion activity. Moreover, the extraction method has to guarantee that the motion activity level of a shot is independent of the underlying coding scheme. Therefore, we propose an extraction scheme that avoids these drawbacks and is automated by developing an intelligent neural network that emulates the opinions of human beings.

We randomly selected 1000 video shots of various durations, extracted from six different video programs, namely three different movies (*Jurassic Park*, *Terminator*, and *Star War*), an animated movie (*Lady Tramp*), and a TV production video with commercials (*Tonight show* and *Football*). Considering such diversity of productions supports the generality of the drawn conclusions and also the proposed schemes. We selectively added 20 more video shots of activity level 5 to increase its representation in the ground truth that is used to train our proposed neural network. We used ten human subjects to evaluate each video shot. A warm up period was conducted before human subjects were shown the video shots. A number of video shots of various activity levels were displayed, accompanied by the anticipated activity levels. We followed the experimental guidelines presented in [23], [24]. Since we used a large ground truth, the assessments with human subjects were conducted in several sessions of about 20 min each over a few days. This avoided human subject fatigue and at the same time allowed for including a large number of video shots in the ground truth.

In general, the human perception of the motion activity in a shot may differ, and not all human subjects may agree on the same judgment. For example, some human subjects may perceive the motion activity in a "car chase" shot as level 4, while others perceive the motion activity as level 5. Table I shows the mean absolute differences for all ten human subjects used to evaluate a given shot, compared to the mean opinion of all human subjects. The average difference for each human subject is denoted "avr_sub." Our subjective experiments produced small differences in human estimates of the motion activity level. This can be attributed to: 1) the large number of video shots used in the ground truth; 2) selectively adding video shots of under represented motion activity levels; 3) using a warm-up period as recommended in [23], [24]; and 4) using five scales to estimate the perceived motion activity level, which is consistent with the MPEG-7 standard. Table II shows the distribution of the human subjective evaluations around the mean opinion point. On average, more than 50% of the human subjects evaluated the motion activity of a video shot to be equivalent to the mean opinion point, see the second column of Table II. In addition, more than 95% of the human subjects evaluated the motion activity of a video shot to be equivalent to ±1 around the mean opinion point, see the third column of Table II.

A. Extracting Visual Features From the Compressed Domain

The number of parameters, from the compressed video domain, that highly correlate with the motion activity level is limited. The amount of motion activity among consecutive frames affects the motion vector values as well as macroblock types. When object movements in a shot are fast and irregular, motion compensation techniques become inefficient, and the number of intra-coded macroblocks increases significantly, i.e., the ratio of intra-coded macroblocks increases as the motion intensity increases. In addition, irregularities in the object movements result in a dispersed motion vector distribution. On the other hand, regular camera/object movement, e.g., a camera pan or panorama

TABLE I
MEAN ABSOLUTE DIFFERENCE OF EACH HUMAN SUBJECT TO THE MEAN VALUE OF THE GROUND TRUTH

movie name	subj1	subj2	subj3	subj4	subj5	subj6	subj7	subj8	subj9	subj10	avr_sub
Jurassic Park	0.3368	0.6737	0.4105	0.4316	0.5684	0.3474	0.2632	0.7263	0.5895	0.6316	**0.4979**
Star War	0.49	0.32	0.645	0.61	0.68	0.355	0.4	0.33	0.375	0.47	**0.4675**
Terminator	0.375	0.335	0.79	0.555	0.43	0.435	0.435	0.45	0.365	0.38	**0.455**
Football	0.425	0.705	0.675	0.53	0.925	0.6	0.41	0.415	0.565	0.35	**0.56**
Tonight	0.3595	0.451	0.4706	0.7843	0.4902	0.3333	0.3529	0.3203	0.2614	0.451	**0.4275**
Lady	0.3632	0.5721	0.607	0.3831	0.6169	0.6169	0.5373	0.4876	0.597	0.4129	**0.5194**

TABLE II
DISTRIBUTION OF HUMAN SUBJECTIVE EVALUATIONS

Movie name	ratio of subjects==mean	ratio of subjects==(mean±1)
Jurassic Park	0.534	0.9628
Star War	0.5655	0.968
Terminator	0.5692	0.9756
Football	0.5109	0.9398
Tonight	0.602	0.9765
Lady	0.5274	0.9537

shot, which is perceived by human subjects as low motion activity results in a dominant motion vector. This dominant motion vector is coded in most macroblocks of the compressed video frames. Therefore, the distribution of the motion vectors around the dominant motion vector is proportional to the human perception of the motion intensity. We capture this effect by extracting the ratio of the macroblocks that have motion vectors equivalent to the dominant motion vector. In order to smooth out small fluctuations around the dominant motion vector that might result from the underlying video coding scheme, a different metric is computed, which accounts for macroblocks that have motion vectors in the range $(\pm 1, \pm 1)$ around the dominant motion vector. This metric not only absorbs the small motion fluctuations but also emphasizes that the proposed extraction scheme can be applied to any coded video stream regardless of the underlying quantization scale or frame rate. The motion related parameters are obtained from P-VOPs, which make reference to previous frames. The following equation expresses the general structure of our proposed metrics:

$$\text{ratio} = \frac{x}{\text{num\_MB}} \tag{1}$$

where num_MB is the total number of macroblocks in the video frame, and x denotes one of the following.

1) The number of intra-coded macroblocks in the current frame.
2) The number of macroblocks with motion vector equal to the dominant motion vector.
3) The number of macroblocks with motion vector in the range of $(\pm 1, \pm 1)$ around the dominant motion vector.

The extracted motion related parameters capture the movement activities between successive frames. The intra-macroblock ratio is directly proportional to the level of the motion activity, while the motion vector ratios are inversely proportional to the motion activity. In order to specify the overall motion activity for a video shot, statistical analysis of these parameters is applied. First order statistics such the mean value, and second-order statistics such as the variance around the mean value are reasonable approximations of the statistical behavior of the motion activity of a shot. Therefore, each video shot is represented by six different parameters [the mean and variance of ratios (1), (2), and (3)] that are used as inputs to the subsequent artificial neural network stage that estimates the shot activity level.

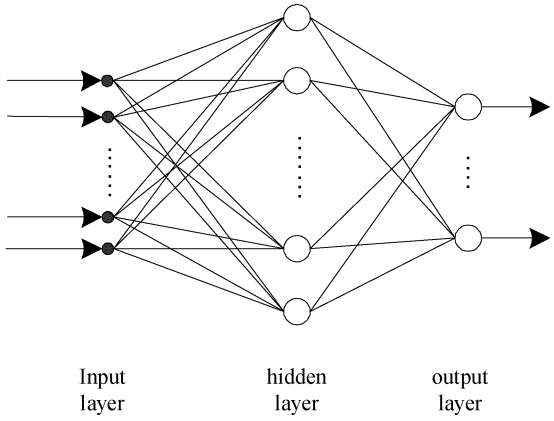


Fig. 3. Architecture of multilayer perceptron network.

TABLE III
MEAN ABSOLUTE DIFFERENCE OF MLP OUTPUTS COMPARED TO THE GROUND TRUTH VALUE $(\text{average MLP} = 0.4781)$

movie name	MLP	movie name	MLP	movie name	MLP
Jurassic Park	0.4211	Terminator	0.465	Tonight	0.3856
Star War	0.475	Football	0.565	Lady	0.5572

B. Artificial Neural Networks (ANN)

Multilayer perceptron (MLP) neural networks are general-purpose, flexible and nonlinear models consisting of a number of simple computational units (neurons) that are organized into multiple layers. In order to design complex models using an MLP, the number of layers and the number of neurons in each layer can be increased. An MLP is capable of accurately predicting the relationship between input vectors and their corre-

TABLE IV
PERFORMANCE OF VARIOUS PARAMETERS USING THE MEAN ABSOLUTE DIFFERENCE TO THE GROUND TRUTH VALUE

Number	Variable Description		Performance
	Per shot	Per frame	
1	Mean	Mean of motion vector magnitudes	1.49622
2	Mean	Variance of motion vector magnitudes	1.82548
3	Mean	Ratio of intra macroblocks	1.05754
4	Variance	Ratio of intra macroblocks	1.23773
5	Mean	Ratio of motion vectors == the dominant motion vector	1.18113
6	Variance	Ratio of motion vectors == the dominant motion vector	0.90565
7	Mean	Ratio of motion vectors (±1, ±1) round the dominant motion vector	1.67359
8	Variance	Ratio of motion vectors (±1, ±1) round the dominant motion vector	0.74339

TABLE V
ERROR RATIO OF VARIOUS SCHEMES USING THE PAIR-WISE COMPARISON

movie name	MLP	Variable Number							
		1	2	3	4	5	6	7	8
Jurassic Park	0.0183	0.0856	0.1131	0.0489	0.0703	0.156	0.1346	0.0948	0.1437
Star War	0.0265	0.0965	0.1377	0.1555	0.1837	0.1004	0.1018	0.1042	0.1126
Terminator	0.0041	0.133	0.1418	0.2763	0.2515	0.2088	0.2046	0.1464	0.2036
Football	0.0799	0.2491	0.1538	0.3337	0.3004	0.3886	0.379	0.2563	0.4029
Tonight	0.0506	0.0537	0.0558	0.2955	0.1374	0.2056	0.1322	0.0537	0.1663
Lady	0.1614	0.0715	0.0469	0.1563	0.2127	0.206	0.158	0.0692	0.1128
Average	**0.0568**	**0.1149**	**0.1082**	**0.211**	**0.1927**	**0.2109**	**0.185**	**0.1208**	**0.1903**

sponding outputs if enough data are available. Therefore, the MLP is known as a universal approximator. Fig. 3 illustrates a typical MLP network, which consists of an input layer, one (or more) hidden layer(s), and an output layer with one neuron per class [25].

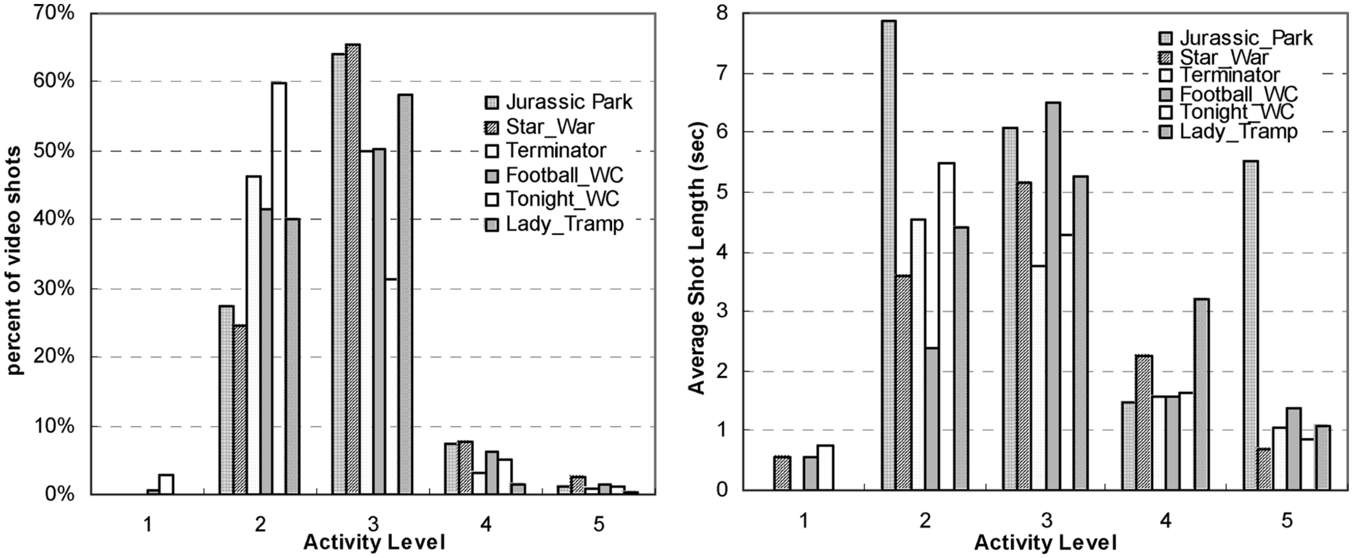
When an input vector is presented at the input layer, the network neurons apply calculations that eventually result in a specific output pattern, which indicates the appropriate class for such input data. Every neuron in a particular layer is connected to every node in the next layer. These connections carry weights, which need to be adjusted during the training phase. In the learning process a set of diverse training examples is input to the MLP network along with their corresponding output values, which are of binary nature in the case of classification problems. Among the algorithms used to learn (or design) the MLP models, the scaled conjugate gradient (SCG) uses second-order information from the neural network. The performance of the SCG has been benchmarked against the performance of the standard backpropagation algorithm [25]. To specify the appropriate number of hidden neurons, it is recommended to have a set of data to train the network and a separate set to test the performance of the network for such particular number of hidden neurons. The optimal number of hidden neurons generates the best performance under the same training set [25]. After designing the MLP network, the network weights are saved and used for the classification of unknown input patterns.

C. Performance of MLP

The training set is composed of input vectors and corresponding output vectors. The dimensionality of the input vector is 6, according to discussion presented in Section II-A. Since our input components are confined in the range from 0 to 1 [see (1)], input preprocessing is not required for our MLP design. The design of the MLP network contains five output neurons, one corresponding to each motion activity level. The target output vectors are expressed in a binary format, whereby only one output neuron is fired (i.e., equal to 1) for each input vector. The rounded mean opinion point of human subject estimates is used to generate such target output vectors.

During the MLP design phase the network performance is evaluated with respect to the target outputs. The outputs of the MLP network after a complete training represent a posterior probability function. For example, an input vector belonging to motion activity i results in a larger value for output neuron number i compared to other output neurons. Correspondingly, if neuron number i corresponds to the maximum output value, a larger number of human subject agrees that the underlying shot has motion activity i compared to the other motion activity classes. We determine therefore the activity level of the applied input vector as the maximum value of the five output neurons. During the model selection of the appropriate number of hidden neurons, the network performance, over the validation set, was calculated as the mean absolute difference to the ground truth value. We used 10-fold validation sets to select the best number of hidden neurons. A regularization parameter of 0.05 was used to avoid over-fitting. The best number of hidden neurons was found to be 8, which resulted in an average of 0.48 differences to the tenfold validation sets. Table III shows the performance of this MLP for each video program.

For sake of comparison, we implemented various single parameter extraction schemes that are presented in [23], in addition to using only the inputs to the MLP network and quantizing the range into disjoint sets. The optimization algorithm presented in [23] was used to obtain the boundary for each activity level. Table IV shows the performance of eight different single parameter extraction schemes. The performance is calculated based on the same tenfold validation sets used for MLP performance evaluation. We should note that the parameters 3–8 are the input vector to the MLP network. We observe that the MLP outperforms all schemes which use a single parameter for



(a)

(b)

Fig. 4. Shot motion activity level histogram of different video sequences.

extracting the motion activity level. This is because the perceived motion activity is a correlation between multiple visual features. The MLP outputs capture the motion activity level by involving six low-level visual features. Another way of evaluating the performance of the different extraction schemes is to use the pair-wise comparison scheme presented in [23]. An ordered list of the motion activity levels of the ground truth shots was produced. The outputs of the motion activity extraction scheme were compared with respect to this ordered list. Table V shows the results of such pair-wise comparison. The error rate of the MLP outputs with respect to the ordered list is about 5.6%, while the best single parameter extraction scheme achieves an error rate of about 10.8%. The MLP outputs are thus found to be superior to existing extraction schemes, using both the mean absolute difference to the ground truth and the pair-wise comparison to the ordered list generated from the ground truth. We have to note that extracting more than one parameter from the compressed video domain linearly increases the computational complexity of the proposed extraction scheme. However, the extraction of these parameters is usually carried out offline and the parameters then used as input to the MLP network. The time delay of MLP to generate the outputs is very small and depends on the number of neurons in the middle layer [25].

The distribution of the motion activity levels for each movie is shown in Fig. 4. The MLP classification outputs for all shots from the six considered videos (including shots that are not represented in the ground truth) are used to produce these distributions. We observe that 75% of movie shots are of activity level 2 or 3, which cover low and moderate activity shots. On the other hand, shots of activity level 5 are not only rare but also have the shortest shot duration.

III. Prioritization of Video Packets

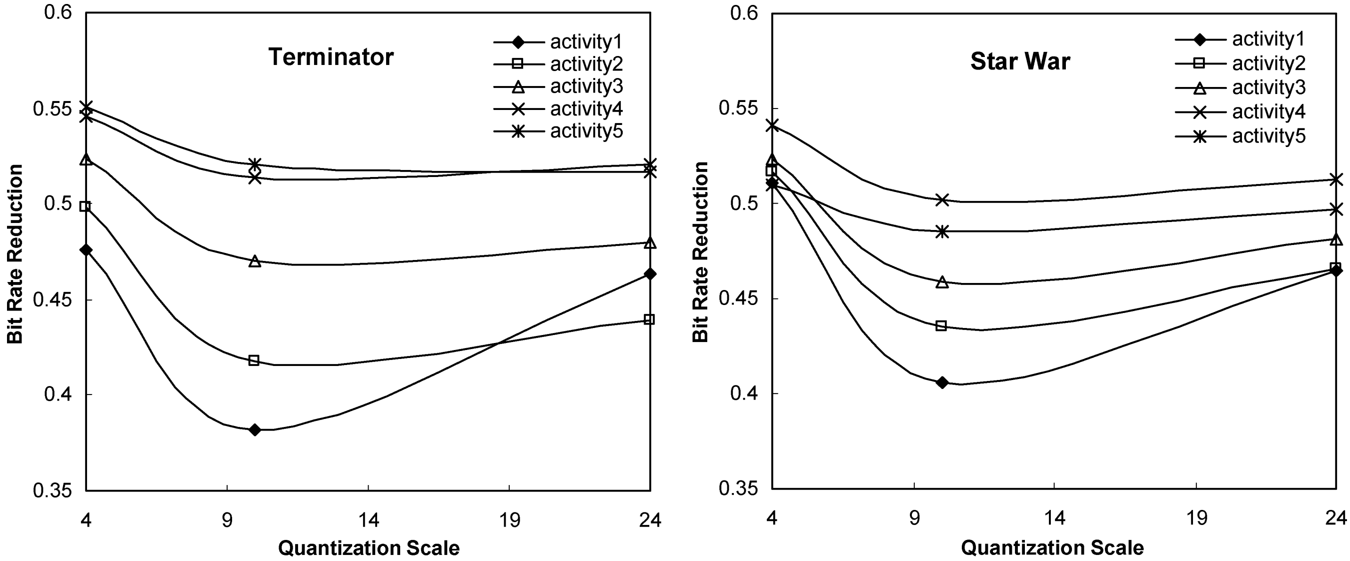
In this section, we present a packet prioritization scheme that exploits the motion activity descriptors extracted with the MLP network of the previous section. There are many options for conveying the MPEG-7 motion activity descriptor to the network nodes. For instance, the motion activity could be carried as differentiated services (DiffServ) code points in the individual packets [26]. Another option could be to convey the motion activity level in real-time control protocol (RTCP) packets [27] to the intermediate nodes. The overhead would be very small since the motion activity information is only required at the shot boundaries (with about 800 shots in a typical 1-hr video sequence). Other approaches could employ one of numerous approaches that are currently being developed for active network paradigm [28], [29].

TABLE VI
Average Bit-Rate Reduction Due to Dropping All B-VOPs

Movie name	Maximum bit rate reduction		
	HQ	MQ	LQ
Terminator	0.515	0.453	0.47
Star War	0.52	0.445	0.474
Lady Tramp	0.573	0.475	0.446
Football_WC	0.557	0.477	0.469

All simulations were performed using the MPEG-4 codec, where 1-hr video sequences were coded at the QCIF resolution with 30 fps. The video streams contained three P-VOPs coded between I-VOPs and two B-VOPs between I-VOP or P-VOPs. Moreover, three different video qualities were generated by adjusting the quantization scale of the encoded streams. The quantization scale parameter was set to 4, 10, and 24 for high-quality (HQ), medium-quality (MQ), and low-quality (LQ) video, respectively. The quality of the video sequence was measured as the peak signal-to-noise ratio (PSNR). The quantization scale settings resulted in average PSNRs of 36, 32, and 28 dB for HQ, MQ, and LQ video, respectively.

For lossy packet networks, congestion control techniques are essential to deliver uninterrupted videos. Among the different techniques that are employed in case of network congestion, frame dropping represents a simple congestion control method



(a) (b)

Fig. 5. Achievable bit-rate reduction due to dropping all B-VOPs per motion activity level.

that can be implemented at the video server, end user or intermediate nodes [30]. Video frames are ordered according to their importance. Packets that carry frames of low importance, such as B-VOPs, are dropped first and if the network congestion is severe some of the P-VOP packets may also subsequently be dropped [31], [32]. However, it is generally recommended to avoid dropping I-VOP or P-VOP packets since severe quality degradation at the decoder would likely occur. In order to avoid dropping I-VOP or P-VOP packets, a sufficient number of B-VOPs needs to be coded in the transmitted video stream. Throughout our experimentations, 2 B-VOPs are coded between I-VOP or P-VOPs, which allows the congestion control scheme to regulate the transmission rate by about 40%–50% of the original rate; see Table VI. Packets belonging to I-VOP and P-VOPs are reliably transmitted while packets belonging to B-VOPs are assigned a priority level according to the motion activity level of the underlying shot.

The B-VOP dropping technique has an upper bound for the video transmission rate reduction, which is achieved when all B-VOPs are dropped. However, the reduction in the original bit rate due to dropping all B-VOPs depends on the motion activity level as well as the quality of the original encoded video stream; see Fig. 5. Shots of higher motion activity level experience higher bit rate reduction. For *Terminator* video sequence and quantization scale 10, the average bit rate reduction for video shots of motion activity level 1, 2, 3, 4, and 5 are 0.37, 0.42, 0.47, 0.52, and 0.53, respectively. This is because shots of high motion activity have a significant difference between consecutive frames, which generates larger bit rates for B-VOPs. On the other hand, the difference between consecutive frames in low motion activity shots is small, which results in relatively lower bit rate reduction.

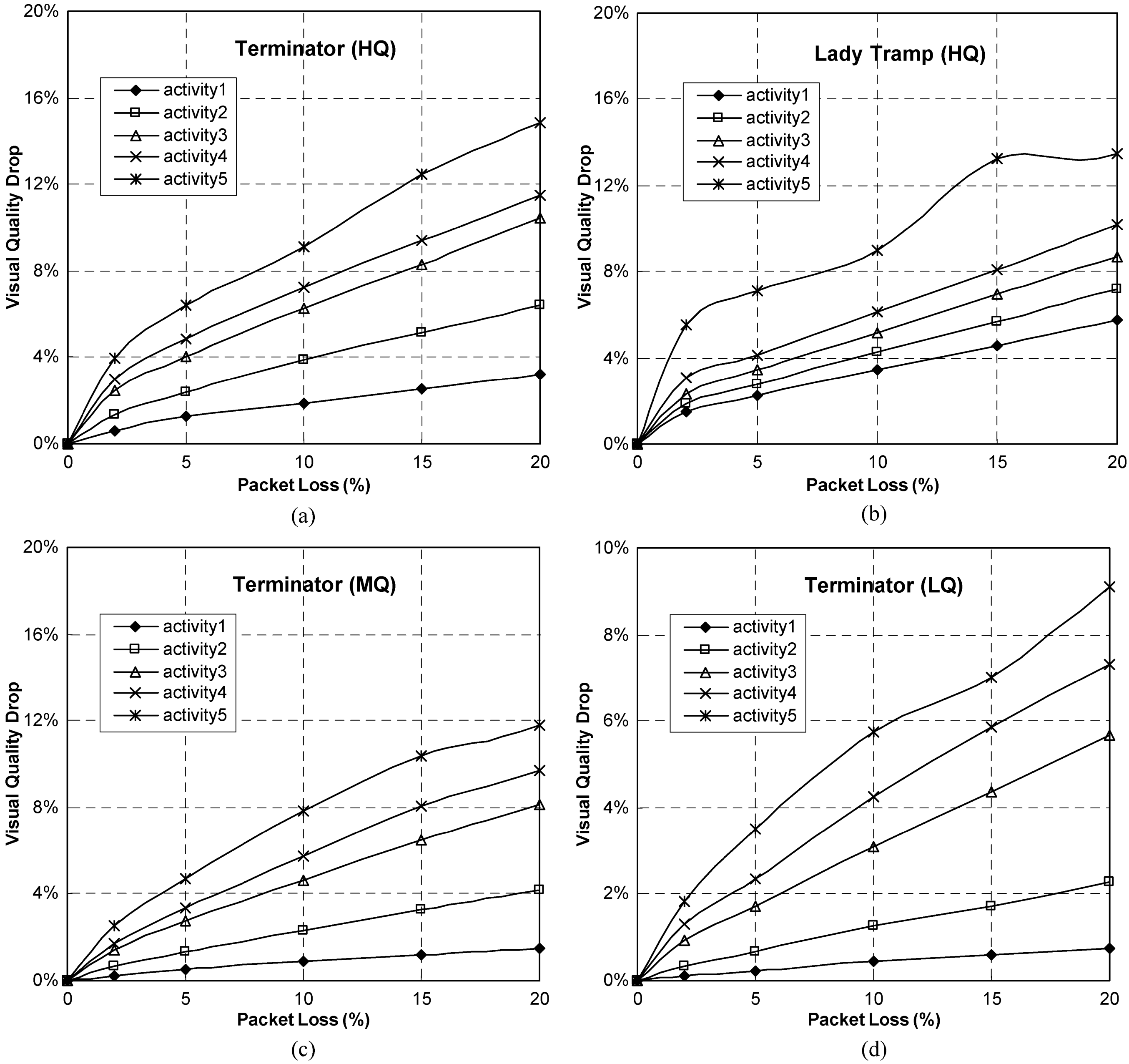
Fig. 5 also shows that the minimal bit rate reduction is achieved when the video shots are encoded at MQ (see quantization scale $=$ 10. For HQ video, B-VOPs are coded with high level of details, which in turn produces a large bit rate for B-VOPs. Therefore, a large bit rate reduction is achieved for HQ video. For MQ video, a lower level of image details is coded, which in turn produces a smaller bit rate for B-VOPs so that the bit rate reduction is reduced. For LQ video, another factor affects the B-VOP bit rate, which is the inefficiency of the motion compensation between reference frames that produces a larger bit rate for B-VOPs compared to MQ video. Therefore, we expect that dropping B-VOPs for videos of medium qualities may not be sufficient, when the congestion is severe. Fortunately, the B-VOP dropping scheme can accommodate up to 40% packet loss during the transmission from the source to destination. This level of packet losses is appropriate for a large number of video transmission scenarios.

IV. Proposed Congestion Control Scheme

In this section, we first conduct a number of experiments to evaluate the impact of packet losses for each packet priority level presented in the previous section. Subsequently, a congestion control scheme is presented as an optimization problem that maximizes the average reconstruction qualities. Finally, the proposed optimal congestion control is simulated over various packet loss scenarios.

A. *Degradation of Visual Quality Due to Packet Loss*

In our simulation experiments, the Gilbert model is used to emulate the Internet packet loss behavior through defining two states namely loss and no-loss state with probabilities $1-p$ and $1-q$ to transit between states [33]. As a result of video transmission, some video packets are lost; the impact of these losses is typically alleviated with error concealment at the decoder [13]. Since packet losses affect only B-VOPs in our evaluation, we adopt concealment by copying from the closest correctly received I-VOP or P-VOP. The human visual system is highly sensitive to contrast in the image quality, which occurs when the correctly received portion of B-VOPs is displayed along with the concealed (copied) portions. We avoid this by copying



(a) (b) (c) (d)

Fig. 6. Reduction of visual quality $Q_\alpha(L)$ as a function of the packet loss ratio L for the different activity levels α.

the entire B-VOP in the case of a partial frame loss. The effect of channel losses is thus perceived as a frame rate reduction, which is better than displaying a diverse image quality at the decoder. Moreover, transmitting packets of small and fixed length tends to minimize not only the delay but also the delay variations (jitter) at the decoder. The MPEG-4 encoder's capability of inserting resynchronization markers separated by an almost fixed number of bits is exploited. Resynchronization markers are fixed length sequences of a specified bit pattern, which are mainly used to restore the normal decoding operation after an error. In our simulations, each video packet carries 512 data bytes, which is specified to the MPEG-4 encoder as the distance between the resynchronization markers. In order to minimize the overhead of the transport protocol headers, the resynchronization markers might be removed since video packets always begin with these markers.

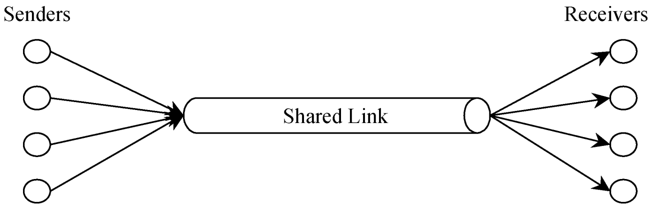
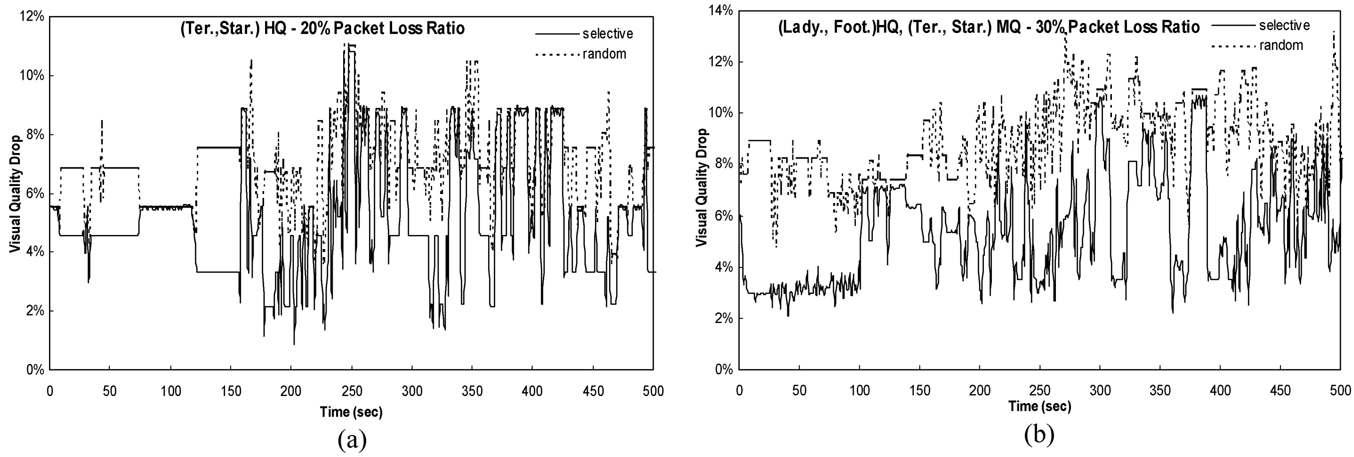


Fig. 7. Multiple senders/ receivers sharing the same link.



(a)

(b)

Fig. 8. Comparison between proposed selective (adaptive) packet dropping and conventional random packet dropping.

The average degradation (drop) in visual quality due to packet loss ratio L for motion activity level α is expressed as

$$Q_\alpha(L) = \frac{1}{N_\alpha} \sum_{n=1}^{N_\alpha} \frac{\mathrm{PSNR}(n,\alpha) - \overline{\mathrm{PSNR}}(n,\alpha,L)}{\mathrm{PSNR}(n,\alpha)} \tag{2}$$

where $\mathrm{PSNR}(n,\alpha)$ represents the average image qualities (measured as PSNR) for shot number n of motion activity level α, while $\overline{\mathrm{PSNR}}(n,\alpha,L)$ represents the average image qualities for shot number n of motion activity level α after concealing a channel loss caused by packet loss ratio L. Moreover, we denote the number of shots of motion activity level α in the video sequence as N_α.

In Fig. 6, $Q_\alpha(L)$ is shown for two video sequences where five different curves are depicted corresponding to the average quality drop for each motion activity level. For *Lady Tramp* (HQ) video sequence at 10% loss of transmitted packets, the average visual quality drops for shots of motion activity level 1, 2, 3, 4, and 5 are 3.4%, 4.2%, 5.2%, 6.2%, and 9%, respectively. We observe that the visual quality drop for high motion activity shots is high since error concealment techniques are inefficient in overcoming such packet losses. Low activity shots experience little movement between successive frames, so that concealing the loss in these shots by copying is acceptable. Due to the logarithmic function that is used to compute the PSNR, the visual quality drop decreases as the quality of the transmitted video decreases. Thus, we notice that video shots of motion activity level 3 of *Terminator* video sequence are degraded by 8.3%, 6.5%, and 4.4% for HQ, MQ, and LQ, respectively, when the video transmission experience 15% packet loss. Fig. 6 also shows that substantial quality drop is experienced for small packet losses; see packet loss ratio in the range 0%–2%. Small packet losses result in some B-VOPs being undecodable, which has a significant impact on the reconstructed quality compared to loss free transmission. In the case of high packet losses, a large number of B-VOPs is lost, and any increase in the packet losses monotonically increases the degradation in visual quality.

B. Link Optimization Problem

In this and the subsequent subsection, we investigate a method to maximize the average reconstructed qualities when several video streams share a common communication resource, e.g., a shared link at the streaming server or an intermediate network node (see Fig. 7). As the video bit rates fluctuate, the shared communication link may experience congestion and some video packets need to be dropped. It is likely that the link buffer contains packets of various video sequences. Additionally, the packets of currently transmitting shots have typically different motion activity levels. According to Fig. 6, the reconstructed quality degradation depends on the motion activity level. In the case of congestion or rate regulation, the number of dropped packets from each video sequence can be specified such that the overall visual quality degradation is minimized. Unless there is severe network congestion, I-VOP and P-VOP packets are reliably transmitted. Hence, such a link allocation process becomes a constrained optimization problem of the weighted sum of the overall visual quality degradation of the J ongoing video streams. The visual quality drop for activity level α and video stream j is represented as $Q_{j,\alpha}(L)$ which is function of the underlying packet loss ratio L. We denote the packet loss ratio for activity level α and video stream j as $L_{j,\alpha}$ and the number of packets in the link buffer of activity level α and video stream j as $B_{j,\alpha}$. We further denote the total link loss ratio as L_T, the total link buffer size as B_T, and the maximum rate reduction that packets of activity level α and video stream j can afford as $\overline{L_{j,\alpha}}$. The optimization problem is formulated as

$$\text{Minimize } Q = \frac{1}{J} \sum_{j=1}^{J} \sum_{\alpha=1}^{5} \frac{B_{j,\alpha}}{B_T} Q_{j,\alpha}(L_{j,\alpha}) \tag{3}$$

Subject to :

$$L_T = \sum_{j=1}^{J} \sum_{\alpha=1}^{5} \frac{B_{j,\alpha}}{B_T} L_{j,\alpha} \tag{4}$$

$$B_T = \sum_{j=1}^{J} \sum_{\alpha=0}^{5} B_{j,\alpha} \tag{5}$$

$$0 \le L_{j,\alpha} \le \overline{L_{j,\alpha}}. \tag{6}$$

In order to minimize (3) under the constraints of (4)–(6), packet prioritization as explained in Section III is employed. In addition, sample points of the relationship $Q_{j,\alpha}(L)$ between the

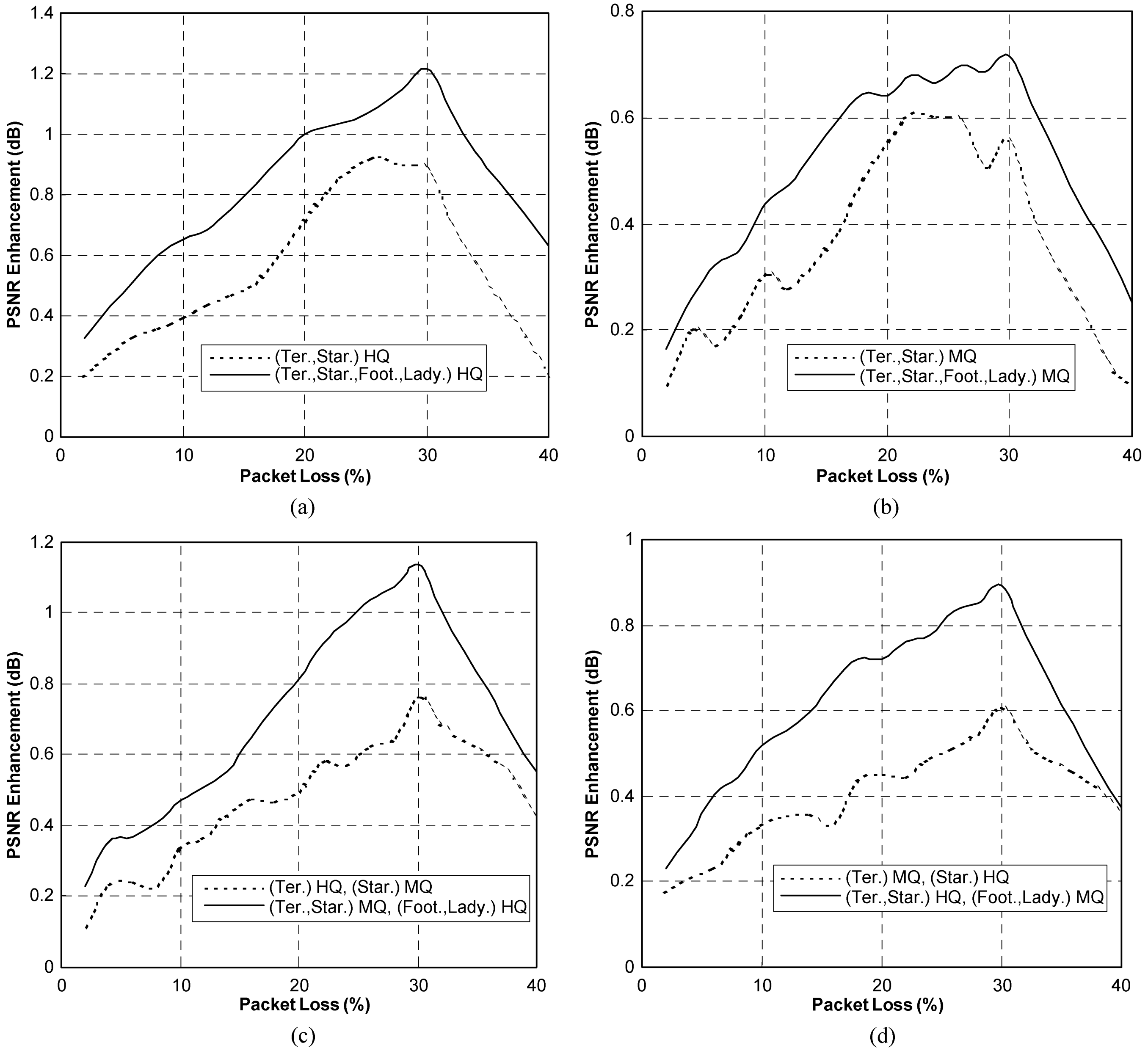
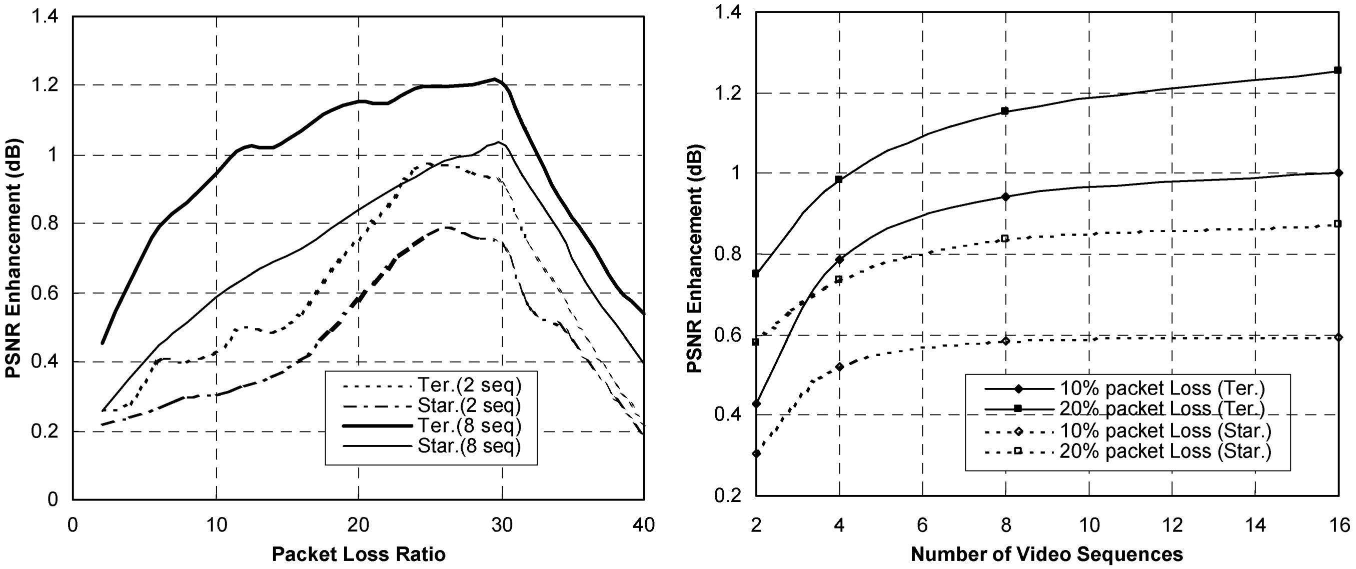


Fig. 9. Performance enhancement of the selective dropping scheme over random B-frame dropping for video streams of various qualities.

visual quality drop and packet loss ratio need to be available at the video streaming server and intermediate nodes that perform link control mechanisms. These sample points can be conveyed during the video connection setup, using RTCP control packets, and the bandwidth overhead can be reduced by approximating the relationship $Q_{j,\alpha}(L)$ using a linear function. Fig. 6 shows that the relationship $Q_{j,\alpha}(L)$ can reasonably be approximated by the slope of a linear function running through the origin of the $Q_{j,\alpha}(L)$ coordinate system. The linear approximation of $Q_{j,\alpha}(L)$ reduces not only the bandwidth overhead, but also reduces the computational complexity of determining the optimal packet loss ratios $L^*_{j,\alpha}$, which minimize (3). Hence, the link optimization problem (3)–(6) can be solved by employing a simple linear search algorithm using for instance dynamic programming [34]. The proposed congestion control algorithm simply drops packets of motion activity α_1 until the quality reduction $Q_{j_1,\alpha_1}(L_{j_1,\alpha_1})$ exceeds the quality reduction $Q_{j_2,\alpha_2}(L_{j_2,\alpha_2})$ of dropping packet of motion activity α_2, where $\alpha_2 > \alpha_1$ and $j_1, j_2 = 1, 2, \ldots, J$. It is found that the optimal packet loss ratios $L^*_{j,\alpha}$ guarantee that the J video sequences experience the same overall visual quality drop during the entire video transmission session, which emphasizes the fairness of the proposed congestion control scheme.

C. Simulation Results

To validate the effectiveness of the proposed congestion control scheme, four different simulation experiments, each using the approximation of the $Q_{j,\alpha}(L)$ relationship by a linear function, have been performed. The first experiment compares the proposed congestion control scheme, to conventional schemes that drop the video packets carrying B-VOPs without any discrimination between these packets (also referred to as random packet dropping). The second and third experiment target scenarios with high user demand for a particular video



(a) (b)

Fig. 10. Performance enhancement of the selective dropping scheme for variable number of video streams.

sequence. The impact of such increasing demands is evaluated in the second experiment, while the impact of the transmission start time is estimated in the third experiment. Finally, the fourth experiment addresses the decision time granularity that is used by intermediate nodes to update the congestion status.

1) Experiment 1: In order to investigate the influence of the packet dropping scheme, a number of video sequences coded at different qualities are transmitted over a shared communication resource that suffers various degrees of contention. In general, the communication resource updates its congestion status every round-trip time (RTT). It is assumed that the communication buffer (B_T) is large enough to allocate the packets of different video streams that are received during every RTT. In this experiment, we select the RTT to be 1 s. Moreover, video transmissions are assumed to start at the same time. These assumptions are relaxed in the subsequent experiments. Fig. 8 shows the average visual quality drop for each video sequence, when selective and random packet dropping are used to resolve the network congestion. It is immediately clear from these results that selective packet dropping achieves smaller quality degradation for each video sequence. When shots of the same motion activity level share the network resource, there is a possibility that both the selective and random dropping schemes result in the same quality degradation. Fortunately, it is rare that two shots of two different video sequences have the same motion activity level and this possibility decreases as the number of video sequences sharing the network resource increases. Fig. 8 also shows that congestion control by selective dropping accomplishes significant improvements, even if the underlying video sequences have different qualities.

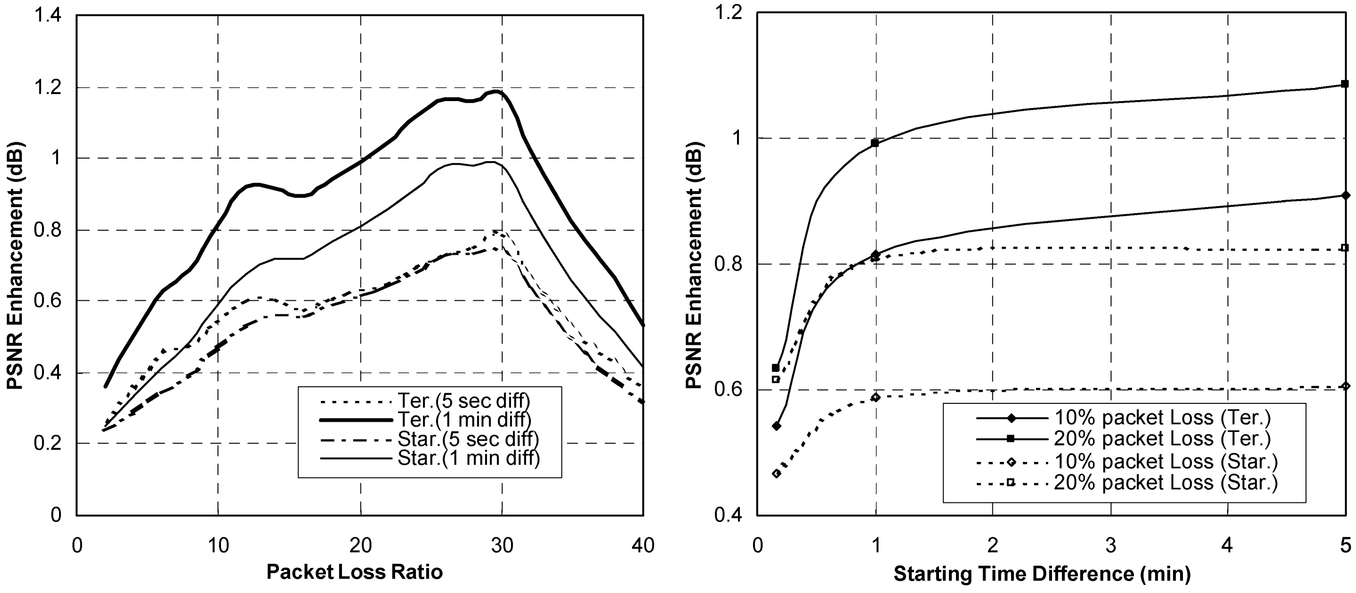
In order to examine the improvements of the proposed selective packet dropping scheme, we compute the average visual quality enhancements compared to random dropping schemes for a packet loss ratio L as

$$\Delta(L) = \frac{1}{NJ}\sum_{j=1}^{J}\sum_{n=1}^{N}\overline{\mathrm{PSNR1}}_j(n,L) - \overline{\mathrm{PSNR2}}_j(n,L) \quad (7)$$

where the visual quality (measured in PSNR) of frame number n and video sequence number j after concealing L packet loss ratio is denoted as $\overline{\mathrm{PSNR1}}_j(n,L)$ if the communication channel applies the proposed selective dropping scheme in the event of congestion, or as $\overline{\mathrm{PSNR2}}_j(n,L)$ if the communication channel applies a random packet dropping scheme in the event of congestion. We further denote the number of video frames as N and the number of video sequences as J. In the following simulations, the visual enhancement $\Delta(L)$ is evaluated for 1-hr movies where N equals 30 (f/s) ×60 (s/min)×60 (min/hr). In addition, the packet loss ratio varies from 2% to 40%.

Fig. 9 shows the visual enhancement $\Delta(L)$ for different sets of video sequences (2 and 4) of various visual qualities that contend for a shared communication resource. For low packet loss ratios, the visual quality drop in all motion activities is small and, hence, selective packet dropping achieves a limited enhancement. However, when the packet loss ratio increases, the enhancement due to selective dropping increases until it reaches a maximum value at around 30% packet loss ratio. The enhancement of the selective dropping decreases for higher packet loss ratios, and the enhancement becomes zero when all B-VOPs are dropped. Both selective and random packet dropping schemes suffer significant packet loss in most motion activity levels when the channel loss is high, so that the enhancement of selective dropping is small at such high losses. However, increasing the number of video sequences results in more packet diversity at the shared resource, which provides the congestion control scheme with a vast amount of dropping scenarios. Therefore, it is observed in Fig. 9 that the enhancement increases when the number of video sequences involved in the congestion situation increases. For example, when two HQ videos share a communication resource that drops 10% of existing packets, our proposed congestion control scheme improves the visual quality by 0.4 dB. On the other hand, when four HQ videos share such a communication resource, enhancement of 0.65 dB is achieved for each video sequence.

2) Experiment 2: To gain insight into situations where the user demands for a particular video dominate the network



(a)

(b)

Fig. 11. Performance of the selective dropping scheme for four video sequences starting transmission at different times.

traffic, a number of simulations have been conducted. Such situations can occur during the weekends for popular movies. In order to emulate the randomness of user demands, we start the video transmission from the streaming server at instances that are uniformly distributed. To assure the statistical confidence of the achieved results, the experiments were implemented 500 times, and the results represent the average of these runs. Similar to pervious experiments, we select the RTT to be 1 second.

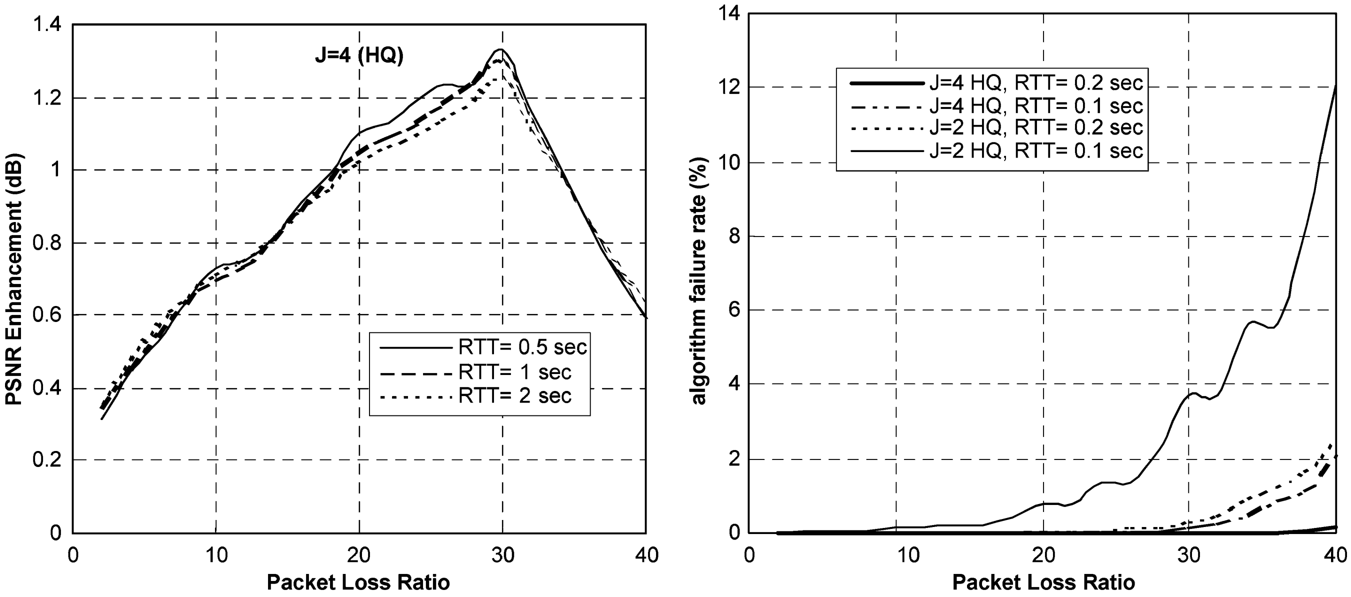
In Fig. 10(a), the visual enhancements $\Delta(L)$ for the *Terminator* and *Star War* video sequences are shown. The depicted curves are the cases of 2 and 8 video sequences sharing the communication resource, where the transmission starts at a random time. In addition, Fig. 10(b) shows the visual enhancement as a function of the number of video streams [J in (7)]. The results demonstrate that the proposed selective dropping scheme achieves additional improvement, when the number of video streams J increases. As a result of increasing the number of video streams, the congestion control scheme has access to video packets that carry a variety of motion activities between overlapped video shots. Such packet variety provides the proposed congestion scheme with many scenarios that can optimize the visual qualities.

3) Experiment 3: In this experiment, the impact of the transmission start time of a video sequence on the performance of the proposed congestion scheme is evaluated. As in the second experiment, we assume that only one video sequence experiences contention at the communication resource. On the other hand, the transmission of a video sequence starts at a specified time interval. Similar to pervious experiments, we select the RTT to be 1 s.

Fig. 11(a) shows the visual enhancements when four video sequences start the transmission 5 s and 1 min apart. In addition, Fig. 11(b) shows the visual enhancement as a function of the transmission starting difference. It is observed that larger visual improvements are achieved, if the transmission start times differ by a longer period of time. With larger starting time difference the overlapped shots are approximately independent and provide the congestion schemes with multiple dropping options.

4) Experiment 4: Typically congestion control schemes receive feedback information about the status of the communication link every RTT and we conduct therefore experiments to estimate the impact of the RTT on our proposed selective packet dropping as well as its impact on any B-VOP dropping scheme. In these experiments, the video transmissions are assumed to start at the same time for all the video sequences. The transmission involves different video sequences coded at HQ.

Fig. 12(a) shows the visual enhancement due to using three different RTT for four video sequences sharing the communication resource. The results reveal that the RTT has little influence on the outcomes of the proposed selective dropping scheme. However, as presented in Section III, any B-VOP dropping scheme may provide the congestion control scheme with insufficient number of packets especially when the congestion is severe. In Fig. 12(b), the failure rates of B-VOP dropping schemes are shown for multiple RTTs and various numbers of video streams. We define a failure to occur when the required bit rate reduction exceeds the maximum bit rate reduction achievable with B-VOP dropping. The results confirm that for large packet loss ratios, such as 40%, the B-VOP dropping schemes suffer larger failure rates compared to smaller channel losses. Such algorithm failure for high channel losses is alleviated for large RTTs. For small RTT, small segments of the video streams overlap at the shared communication resource. Such small video segments as well as the same group of picture that is employed for all coded videos imply that packets carrying B-VOPs might be scarce at the communication resource. In addition, Fig. 12(b) demonstrates that increasing the number of video sequences J reduces the failure rate of B-VOP dropping schemes. More video sequences provide the congestion scheme with sufficient number of packets carrying B-VOPs.



(a) (b)

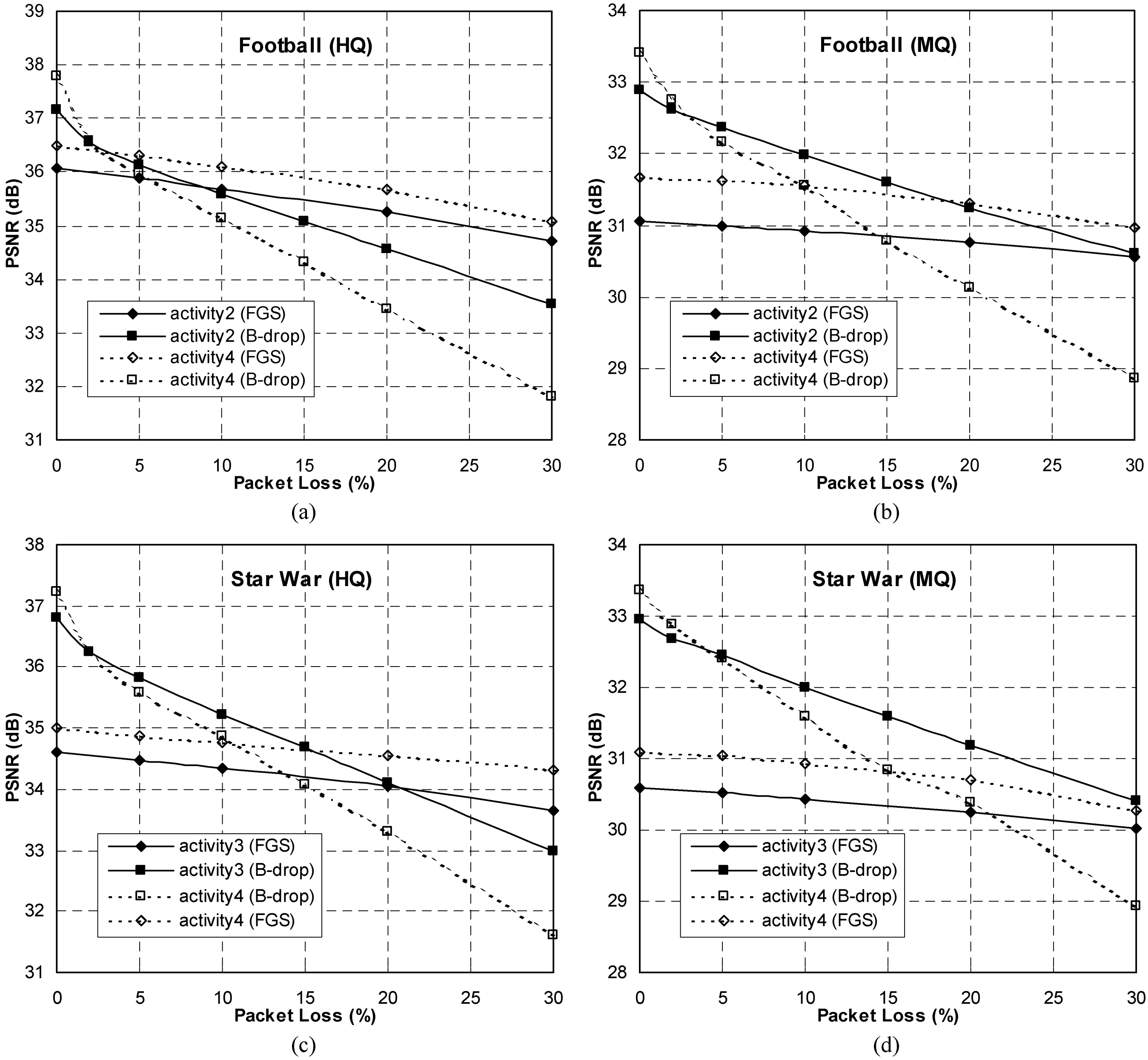
Fig. 12. Performance enhancement of the selective dropping scheme for communication links of variable RTT.

V. Network Performance Comparison Between Single Layer and FGS Coding

In this section, we examine the proposed transmission scheme, shown in Fig. 1, in the context of applications that code the video sequence using more than one coding scheme. The main idea is to select the coding scheme according to the channel conditions as well as the visual content so as to maximize the reconstructed video quality. We consider an application that codes the video either into a single layer and employs selective B-frame dropping (as explained in Section IV) or codes the video using the FGS codec. The FGS coding scheme has recently been included in the MPEG-4 standard [3]. With FGS coding the base layer is coded at the lowest acceptable video quality and requires reliable delivery. The enhancement layer bitstream contains the bit plane coded residual coefficients of the underlying base layer coefficients. Due to the bit plane coding, FGS can support the finest data cutoff in the enhancement layer. Assigning the enhancement layer packets the same priority level (which is lower than for the base layer), results essentially in random packet dropping in the case of network congestion. Random packet dropping may result in packets carrying more significant bit planes being dropped before packets carrying lower significant bit planes. This in turn makes the corresponding packets of lower significant bit planes undecodable. Instead of assigning a single priority level to the enhancement layer packets, we can distinguish between these packets according to their visual effect. The more significant bit planes are more important than the less significant ones, so that we assign packets a priority level according to the underlying bit plane significance. In the event of congestion, packets that carry the least significant bit planes are dropped first.

We proceed to examine the role of the MPEG-7 motion activity descriptor in selecting the coding scheme (single-layer with selective B-frame dropping or FGS) that results in the better reconstructed quality. FGS coding is generally very robust to packet losses (since an enhancement layer frame is coded with reference to the corresponding base layer frame, keeping packet losses localized) but suffers from low coding efficiency (splitting the video sequence into a base and enhancement layer and furthermore splitting the enhancement layer into a number of bit planes). For loss free transmission, FGS coding produces lower reconstructed quality, compared to single layer coding scheme. In the case of network congestion, the quality degradations of single layer coding are produced due to B-VOP packet dropping. FGS does not require an error concealment technique, while B-VOP losses of single layer coding are concealed by copying from the closest frame. According to previous simulation results, the lowest video quality is achieved with a quantization scale of 24, so that we use this quantization scale to code the base layer. For the sake of comparison, the sender truncates the bit rate of the enhancement layer such that the overall bit rate is equivalent to the transmitted bit rate of the single layer. Fig. 13 shows the reconstructed qualities due to various packet losses of video sequences coded using FGS and single-layer schemes. The visual qualities are measured using PSNR, and the average value for each motion activity is used to draw the curves. Fig. 13 also shows the reconstructed qualities for two different video qualities of the same video sequence.

First, we observe that using the motion activity descriptors with FGS encoded video would have a rather insignificant effect on the reconstructed quality. To see this, note that in Fig. 13, the FGS encoded video experiences rather limited quality degradations due to packet losses. The visual quality of the loss free transmission (0% packet loss) depends on the color complexity of the underlying video sequence. Also, note that the curves of FGS video streams (with ◊ and ♦ markers in Fig. 13) are almost horizontal and never intersect. On the other hand, the comparison between FGS and B-VOP dropping of single layer coding shows a significant relationship with the motion activity of the



(a)

(b)

(c)

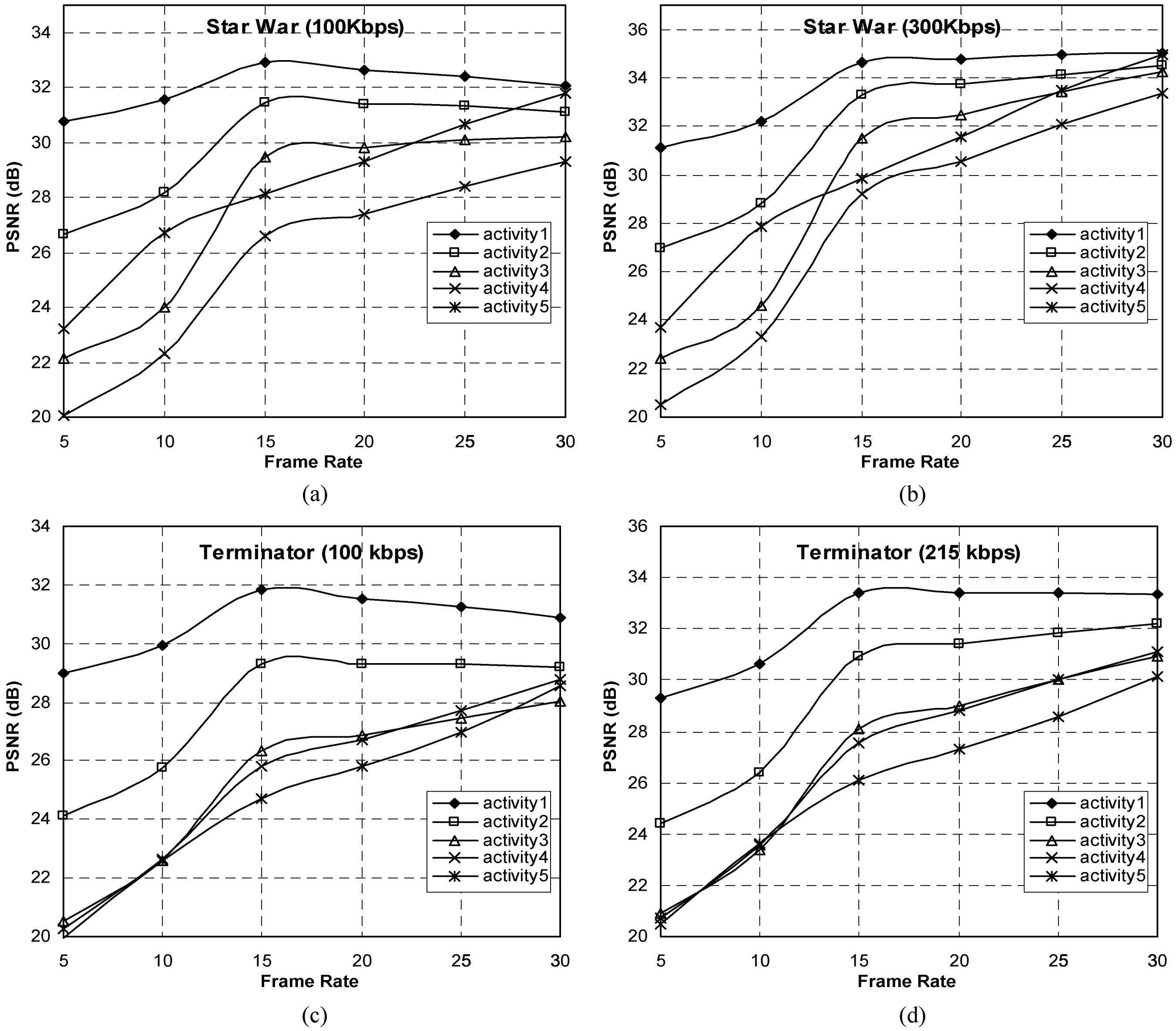
(d)

Fig. 13. Comparison between B-VOP and FGS enhancement layer dropping.

underlying video shots. We observe that the coding efficiency of FGS is smaller than the single layer scheme; compare 0% packet loss in Fig. 13. However, the FGS coding scheme is very robust to packet losses, where the visual quality monotonically decreases by smaller values compared to the B-VOP dropping of single layer videos. At a certain packet loss ratio that is denoted as the *critical point*, the FGS coding scheme performs better than single layer coding. The motion activity level plays an important role in specifying the value of the critical points. For low motion activity shots, video coded using single layer coding can be efficiently concealed by copying, which results in a higher value of the critical point. Thus, the value of the critical point is inversely proportional to the motion activity of the underlying shot. In the case of the *Star War* video sequence, shots of motion activity level 3 and 4 have critical points at 20% and 11% packet loss, when the video is coded at HQ. FGS coding suffers a large loss in the coding efficiency in the case of low visual quality. Consequently, the value of the critical point increases as the quality of the transmitted video decreases. In Fig. 13, for the *Football* video sequence, the critical points are 9% and 30% for high and medium quality video, when the video shots are of low motion activity level $(= 2)$. We note that ANNs can be designed to efficiently detect the critical points, shown in Fig. 13 and, hence, the coding scheme can be alternated between single layer coding and FGS coding schemes. Additional details of designing this neural network can be found in [35].

VI. Transcoding the Video Streams

This section is dedicated to evaluating the performance of rate shaping algorithms that can be applied to the proposed transmission scheme, which employs the motion activity descriptors



(a)

(b)

(c)

(d)

Fig. 14. Impact of frame rate on the perceived visual quality.

of MPEG-7. Since current Internet transmission does not guarantee quality of service (QoS), each application chooses the preferred transport protocol to achieve the required performance. For example, traditional data applications employ the transmission control protocol (TCP) that accomplishes loss-free data transfer by means of window-based rate control and retransmissions. On the other hand, loss-tolerant applications such as video streaming prefer the user datagram protocol (UDP) which has no rate control mechanism and, thus, avoids the delays introduced by packet retransmission. In order that both TCP and UDP sessions fairly coexist in the Internet, “TCP-friendly” rate control has been introduced [36]. A TCP-friendly system regulates its data transmission rate according to the network condition, typically expressed in terms of RTT and the packet loss probability, to achieve similar throughput as a TCP connection would on the same path.

Existing rate-control schemes can be classified into three categories: source-based, receiver-based, and hybrid rate control [30], depending on whether the source, the receiver, or both adjust the transmission rate. Rate shaping techniques are required to match the transmission rate of a preencoded video stream to the target rate constraint. Transcoding is a rate shaping technique, where the original video stream is decoded and then re-encoded to the target transmission rate [38], [39]. Depending on the video coding scheme, transcoding could be simplified without full decoding and re-encoding, which enables the online implementation. Conventional transcoding techniques as well as rate controls at the source node adjust the quantization scale of the encoded video stream to achieve the target transmission rate. In [40], a bit-allocation scheme was proposed that distributes the available bit budget among various video objects based on the complexity and motion intensity information. However, this scheme is only beneficial for scalable object coding schemes. In addition, this scheme does not specify the method of extracting the content descriptors or even specifying the role of these content descriptors. In our work, we propose a transcoding scheme using the motion activity descriptors of MPEG-7 standard to regulate the frame rate of a given bit budget in order to improve the reconstructed visual qualities.

In our approach, a number of video streams are encoded at various bit rates, using for instance the TM5 rate control scheme [41]. Moreover, the frame rate of the encoded video stream is considered as a variable parameter to the rate control technique. Although the video transmission rate is controlled using the TM5 scheme, the observations are consistent with any transcoding scheme. Fig. 14 shows the average reconstructed visual quality for each motion activity level, encoded at different frame rates. Each video sequence is coded into two different bit rates, corresponding to low and medium video quality. For frame rates less than 30 fps, the decoder up-sampled the video stream into 30 fps by repeating the closest video frames. Thus, the reconstructed video quality is calculated as the difference between the up-sampled video stream and the original video stream (i.e., without frame repeating).

We observe that shots of low motion activity can be coded into higher visual quality, if the frame rate is decreased. Decreasing the frame rate allows the encoder to allocate a larger bit budget for each encoded frame. In the case of low motion activity shots, the un-encoded frames contain little motion information that can be compensated by repeating the closest frame. Therefore, the optimal frame rate for encoding shots of motion activity 1 or 2 at 100 kbps is around 15 fps (see Fig. 14). However, increasing the transmission rate shifts the optimal frame rate to a higher value. When the available bit rate is sufficiently large, video shots of various motion activity levels can be coded at the maximum frame rate (i.e., 30 fps).

VII. Conclusions and Future Work

In this paper, we have proposed methods for employing the MPEG-7 descriptors in current video transmission schemes. The proposed methods can be easily implemented in conjunction with existing MPEG-4 codecs, with negligible increase in complexity. Motion activity descriptors show a strong correlation with the reconstructed video quality after losses during network transport. Therefore, we have designed up to date neural network technologies to automatically extract the motion intensity level for a video shot. Our extraction scheme achieves high consistency with human judgments. Additionally, the extracted descriptors are exploited by developing a selective packet dropping scheme that is used in the case of network congestion. Visual improvements due to the proposed congestion control scheme are substantial. Enhancements of 1.2 dB per video sequence are achieved for some packet loss ratios. A comparison between single layer and FGS coded video streams is also presented. Our simulation results demonstrate that the FGS schemes are robust to high packet loss ratios, especially when the visual content belongs to moderate or high motion video sequences. Furthermore, considering the motion activity levels of the video sequences during the transcoding application can result in visual improvements. Simulation results suggest that minimizing the frame rate increases the reconstructed visual quality, in the case of low motion activity sequences and low transmission rates. It is possible to extend the proposed schemes to be implemented in conjunction with H.264 codecs, which achieve higher compression ratios than MPEG-4 codecs.

Acknowledgment

The authors would like to thank the members of the Cubic laboratory and the members of the video trace project of Arizona State University, Tempe, for their help in the subjective evaluations of the motion activity levels.

References

[1] A. Vetro, J. Xin, and H. Sun, "Error resilience video transcoding for wireless communications," *IEEE Wireless Commun.*, vol. 12, no. 4, pp. 14–21, Aug. 2005.

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] *Information Technology- Generic Coding of Audio-Visual Objects—Part 2: Visual*, ISO/IEC 14 496-2, 2001.

[4] Y.-C. Lee, J. Kim, Y. Altunbasak, and R. M. Mersereau, "Layered coded versus multiple description coded video over error-prone networks," *Signal Process.: Image Commun.*, vol. 18, no. 5, pp. 337–356, May 2003.

[5] S. Kang, H. Y. Youn, Y. Lee, D. Lee, and M. Kim, "The active traffic control mechanism for layered multimedia multicast in active network," in *Proc. 8th Int. Symp. Modeling, Anal. Simul. Comput. Telecommun. Syst.*, 2000, pp. 325–332.

[6] Q. Zhang, G. Wang, W. Zhu, and Y.-Q. Zhang, "Robust scalable video streaming over internet with network-adaptive congestion control and unequal loss protection," presented at the Packet Video Workshop, Kyongju, Korea, Apr. 2001.

[7] P. Chou and Z. Miao, Rate-Distortion Optimized Streaming of Packetized Media Microsoft Res. Tech. Rep. MSR-TR-2001-35, Feb. 2001.

[8] J. Chakareski and P. Chou, Application Layer Error Correction Coding for Rate-Distortion Optimized Streaming to Wireless Clients Microsoft Rese. Tech. Rep. MSR-TR-2002-81, Aug. 2002.

[9] I. Bajic, O. Tickoo, A. Balan, S. Kalyanaraman, and J. Woods, "Integrated end-to-end buffer management and congestion control for scalable video communications," in *Proc. IEEE Int. Conf. Image Process.*, 2003, vol. 3, pp. III-257–III-260.

[10] S. Kang, Y. Zhang, M. Dai, and D. Loguinov, "Multi-layer active queue management and congestion control for scalable video streaming," in *Proc. 24th Int. Conf. Distribut. Comput. Syst.*, 2004, pp. 768–777.

[11] A. Dawood and M. Ghanbari, "Content-based MPEG video traffic modeling," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 77–87, Mar. 1999.

[12] S.-F. Chang and P. Bocheck, "Principles and applications of content-aware video communication," in *Proc. IEEE Int. Symp. Circuits Syste.*, 2000, vol. 4, pp. 33–36.

[13] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: a review," *Proc. IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[14] *MPEG-7 Visual Part of the XM 4.0*, ISO/IEC MPEG99/W3068, Dec. 1999.

[15] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *SPIE Storage and Retrieval for Still Images and Video Databases IV*, vol. 2664, pp. 170–179, 1996.

[16] X. U. Cabedo and S. K. Bhattacharjee, "Shot detection tools in digital video," in *Proc. Non-linear Model Based Image Analysis*, 1998, pp. 121–126.

[17] A. Hampapur, R. C. Jain, and T. Weymouth, "Production model based digital video segmentation," *Multimedia Tools Applicat.*, vol. 1, no. 1, pp. 9–46, Mar. 1995.

[18] N. Vasconcelos and A. Lippman, "A Bayesian video modeling framework for shot segmentation and content characterization," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Jun. 1997, pp. 59–66.

[19] A. Divakaran, "An overview of MPEG-7 motion descriptors and their applications," in *Proc. 9th Int. Conf. Comput. Anal. Images Patterns*, 2001, vol. 2124, pp. 29–40.

[20] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 720–724, Jun. 2001.

[21] X. Sun, B. S. Manjunath, and A. Divakaran, "Representation of motion activity in hierarchical levels for video indexing and filtering," in *Proc. ICIP*, Sep. 2002, vol. 1, pp. I-149–I-152.

[22] A. Divakaran, A. Vetro, K. Asai, and H. Nishikawa, "Video browsing system based on compressed domain feature extraction," *IEEE Trans. Consum. Electron.*, vol. 46, no. 3, pp. 637–644, Aug. 2000.

[23] K. A. Peker and A. Divakaran, "Framework for measurement of the intensity of motion activity of video segments," *J. Vis. Commun. Image Represent.*, vol. 15, no. 3, pp. 265–284, Sep. 2004.

[24] VQEG (Video Quality Expert Group), RRNR-TV Group: Test Plan ver. Draft Version 1.7, ITU-T/COM-T/COM12/C [Online]. Available: http://www.its.bldrdoc.gov/vqeg/, May 2004

[25] S. Haykin*, Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[26] K. Chan, R. Sahita, S. Hahn, and K. McCloghrie, Differentiated services quality of service policy information base, request for comments (RFC3317) Mar. 2003.

[27] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, RTP: A transport protocol for real-time applications request for comments: RFC3550, Jul. 2003.

[28] A. Patel, "Active network technology," *IEEE Potentials*, vol. 20, no. 1, pp. 5–10, Feb.–Mar. 2001.

[29] R. Balakrishnan and K. Ramakrishnan, "Active router approach for selective packet discard of streamed MPEG video under low bandwidth conditions," in *Proc. IEEE Int. Conf. Multimedia Expo,*, 2000, vol. 2, pp. 739–742.

[30] W. Deng, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha, "Streaming video over the internet: approaches and directions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 282–300, Mar. 2001.

[31] M. R. Ito and V. Bai, "A packet discard scheme for loss control in IP networks with mpeg video traffic," in *Proc. 8th Int. Conf. Commun. Syst.*, Nov. 2002, vol. 1, pp. 497–503.

[32] Z.-L. Zhang, S. Nelakuditi, R. Aggarwa, and R. P. Tsang, "Efficient server selective frame discard algorithms for stored video delivery over resource constrained networks," in *Proc. IEEE INFOCOM'99*, pp. 472–479.

[33] S. Wenger, Error patterns for internet experiments ITU Telecommunications Standardization Sector, Oct. 1999, Doc. Q15-I-16r1.

[34] O. Lotfallah, "Content-aware video transmission systems," Ph.D. dissertation, Arizona State Univ., Tempe, 2004.

[35] O. Lotfallah and S. Panchanathan, "Adaptive scheme for internet video transmission," in *Proc. ISCAS 2003*, vol. 2, pp. 872–875.

[36] N. Wakamiya, M. Miyabayashi, and M. Murata*, MPEG-4 Video Transfer With TCP-Friendly Rate Control*. Berlin, Germany: Springer-Verlag, 2001, vol. 2216, pp. 29–42.

[37] N. Yeadon, F. Garcia, D. Hutchison, and D. Shepherd, "Filters: QoS support mechanisms for multipeer communications," *IEEE J. Sel. Areas Commun.*, vol. 14, pp. 1245–1262, Sep. 1996.

[38] E. Jammeh, M. Fleury, and M. Ghanbari, "Smoothing transcoded MPEG-1 video streams for internet transmission," *Proc. IEE. Vis., Image Signal Process.*, vol. 151, no. 4, pp. 298–305, Aug. 2004.

[39] R. Puri, K.-W. Lee, K. Ramchandran, and V. Bharghavan, "An integrated source transcoding and congestion control paradigm for video streaming in the internet," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 18–32, Mar. 2001.

[40] A. Vetro, A. Divakaran, H. Sun, and T. Poon, "Transcoding system based on MPEG-7 metadata," in *Proc. IEEE Pacific-Rim Conf. Multimedia*, Sydney, Australia, Dec. 2000, pp. 420–423.

[41] *MPEG Video Test Model 5*, ISO/IEC JTC1/SC29/WG11, MPEG93/457, Apr. 1993, Draft.

Osama A. Lotfallah received the B.S. and M.S. degrees in computer engineering from Cairo University, Egypt, in July 1997 and July 2001, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, Tempe, in December 2004.

During his Masters study, he worked as Teacher Assistant in the Computer Science Department, Cairo University, Cairo, Egypt. Currently, he is a Postdoctoral Research Associate, Department of Computer Science and Engineering, Arizona State University, Tempe, since January 2005. He was actively involved in the teaching and research activities in the field of digital signal processing. He was also an active member of the video traces research group of Arizona State University. His research interests are in the fields of advanced video coding, digital video processing, visual content extraction and video streaming, with a focus on adaptive video transmission schemes. He has two provisional U.S. patents in the field of content-aware video streaming. He is a regular reviewer of many international conferences in the field of visual communication, as well as periodical journal and magazines in the field of multimedia and signal processing.

Martin Reisslein (A'96–S'97–M'98–SM'03) received the Dipl.-Ing. (FH) degree from the Fachhochschule Dieburg, Dieburg, Germany, in 1994 and the M.S.E. degree from the University of Pennsylvania, Philadelphia, in 1996, both in electrical engineering, and the Ph.D. degree in systems engineering from the University of Pennsylvania in 1998.

He is currently an Associate Professor in the Department of Electrical Engineering at Arizona State University (ASU), Tempe. During the academic year 1994–1995, he was a Visitor at the University of Pennsylvania as a Fulbright scholar. From July 1998 to October 2000, he was a scientist with the German National Research Center for Information Technology (GMD FOKUS), Berlin, Germany, and a Lecturer at the Technical University Berlin. From October 2000 to August 2005, he was an Assistant Professor at ASU. His research interests are in the areas of Internet quality of service, video traffic characterization, wireless networking, optical networking, and engineering education.

Since January 2003, Dr. Reisslein has served as Editor-in-Chief of the *IEEE Communications Surveys and Tutorials*. He has served on the Technical Program Committees of IEEE Infocom, IEEE Globecom, and the IEEE International Symposium on Computer and Communications. He has organized sessions at the IEEE Computer Communications Workshop (CCW). He has been co-recipient of Best Paper Awards at the SPIE Photonics East 2000—Terabit Optical Networking Conference and the 2006 IEEE Consumer Communications and Networking Conference (CCNC).

Sethuraman Panchanathan (S'87–M'89–SM'96–F'01) received the B.Sc. degree in physics from the University of Madras, India, in 1981, the B.E. degree in electronics and communication engineering from the Indian Institute of Science, Bangalore, India, in 1984, the M. Tech degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 1986, and the Ph.D. degree in electrical engineering from the University of Ottawa, Ottawa, ON, Canada, in 1989.

He is currently a Professor and Director of the School of Computing and Informatics, Chair of the Computer Science and Engineering Department, Director of the Institute for Computing & Information Sciences & Engineering, and Director of the Research Center on Ubiquitous Computing (CUbiC) at Arizona State University, Tempe. He is also the Interim Director of the new Biomedical Informatics department and an Affiliate faculty in the University of Arizona College of Medicine, Phoenix program. He is also an Affiliate Professor in the Department of Electrical Engineering at ASU. He is co-founder and President of a start-up company MotionEase Inc. which is focused on developing video based motion capture solutions for rehabilitative applications. He leads a team of Researchers and Graduate students working in the areas of Ubiquitous Multimedia Computing, Visual Computing and Communications; Media Processor Designs; Content-based and Compressed Domain Indexing and Retrieval of Images and Video. Multimedia Communication, Face/Gait Analysis and Recognition, Genomic Signal Processing; Ubiquitous Computing Environments for Blind Persons. CUbiC's flagship project iCARE designs assistive technologies and accessible environments for individuals who are blind. The iCARE project won the Governor's Innovator of the Year-Academia Award in November 2004. He was an Adjunct Professor in the School of Information Technology and Engineering at the University of Ottawa (1997–2004). He was an Honorary Visiting Professor at the University of New South Wales in Sydney, Australia, from 1997 to 1999. He was the Chief Scientific Researcher of Obvious Technology (1995–2004) and was a Scientific Advisor for Luxxon Corporation, San Jose, CA (2000–2003). He is a member of the electronic health steering committee appointed by the Governor of Arizona. He was a member of the design team for the Phoenix campus of the University of Arizona medical school and is also a member of the Academic Working group of the medical school. He is currently a PI/Co-PI of projects funded by the National Science Foundation (NSF), National Institute of Health (NIH), Department of Economic Security (Rehabilitation Services Administration), Banner Health and industry. He was a Principal Investigator/ Co-Investigator of projects funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), Federal Centers of Excellence on Telecommunications Research (CITR) and Microelectronics Network (MICRONET), the Provincial Center of Excellence on Telecommunications Research (TRIO), Corel Corporation, and Canadian Marconi, Telesat Canada, Callisto Corp. He was also a PI of projects funded by Intel, Motorola, ARM, and SUN. He. He has published over 280 papers in refereed journals and conferences. He was the Co-Chair of the IS&T/SPIE Digital Video Compression-Algorithms and Technologies'96 and Multimedia Hardware Architectures'97 Conferences. He was also the Tutorials Chair of the IEEE International Conference on Multimedia Systems '97. He was the Symposium Chair of Electronic Imaging '98 Symposium and the Chair of the Multimedia Hardware Architectures'98 Conference. He was the Co-Chair of the Multimedia Storage and Archiving Systems III and IV Conferences in Photonics East. He was also the Co-Chair of the Media Processors 1999, 2000, 2001, and 2002 conferences. He was also a Co-Chair of the Workshop on Parallel and Distributed Computing in Image Processing, Video Processing, and Multimedia in 2000, 2001, 2002, and 2003. He was the Co-Chair of the Internet Multimedia Management Systems 2000, 2001, 2002, and 2003 Conferences. He was the Co-General Chair of the IEEE International Symposium on Circuits and Systems (ISCAS2002). He is currently the Co-Chair of the Visual Communications and Image Processing (VCIP) 2003 Conference. In addition, he is a program committee member of numerous conferences, organizer of special sessions in several conferences, and an invited panel member of special sessions. He has a chapter on "Compressed/Progressive Search" in the book on *Image Databases, Search and Retrieval of Digital Imagery* (Wiley, 2001).

Dr. Panchanathan is the Editor-in-Chief of the *IEEE Multimedia Magazine.* He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, Area Editor of the *Journal of Visual Communications and Image Representation*, and an Associate Editor of the *Journal of Electronic Imaging, International Journal on Artificial Intelligence Tools Architectures, Languages, Algorithms, International Journal on Systemics, Cybernetics and Informatics*. He was also an Associate editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT) and the *IEEE Multimedia Magazine*. He guest edited a two-part special issue (December 1996 and June 1997) on "Indexing, Storage, Browsing and Retrieval of Images and Video" for the *Journal of Visual Communication and Image Representation*. He was a Guest Editor of the special issue on "Visual Computing and Communications" for the *Canadian Journal of Electrical and Computer Engineering*. He was a Guest Editor of a three-part special issue (September 1998, November 1998, and February 1999) on "Image and Video Processing for Emerging Interactive Multimedia" in the IEEE TCSVT. He was also a Guest Editor of a special issue on "Conceptual and Dynamical Aspects of Multimedia Content Description" in the IEEE TCSVT which appeared in September 2002. He was also a Guest Editor of a special issue on "Emerging H.264/AVC Video Coding Standard" in the *Journal of Visual Communication and Image Representation* He is a Fellow of the International Society for Optical Engineers (SPIE), and a member of the European Association for Signal Processing (EURASIP), the Association of Computing Machinery (ACM), and ASEE.