ADAPTIVE MULTIPLE DESCRIPTION CODING FOR INTERNET VIDEO

Osama A. Lotfallah and Sethuraman Panchanathan

Visual Computing and Communication Laboratory
Center for Ubiquitous Computing (CUbiC)
Arizona State University, Tempe, AZ 85287, USA

ABSTRACT

Transmission of compressed video over unreliable networks such as the Internet will inevitably result in serious degradations in video quality. Therefore, multiple description coding (MDC) encodes the video signal into a number of bit streams (descriptions) such that receiving a description improves the video quality. Among the different MDC schemes, spatial/temporal sub-sampling is a low complex encoding /decoding scheme, however these MDC schemes are vulnerable to varying channel conditions and visual content. To guarantee better image quality during transmission, we propose an adaptive scheme that alternates between two simple MDC schemes depending on the channel status and underlying visual content. The proposed scheme achieves substantial robustness over packet lossy networks. Furthermore, a Multi-Layer Perceptron (MLP) network is designed to provide an on-line decision mechanism with 85% accuracy. Experimental results suggest that our approach can achieve a superior video quality for a wide range of packet loss and video content.

1. INTRODUCTION

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 such as http, ftp employ 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 conferencing and video streaming prefer UDP to avoid unacceptable delay introduced by packet retransmission. Transmission of video sequences using UDP is considered a greedy traffic because TCP throttles the transmission rate in the event of network congestion whilst UDP does not have such a control mechanism. In order that both TCP and UDP sessions fairly co-exit in the Internet, “TCP-friendly” rate control is introduced [1]. A TCP-friendly system regulates its data transmission rate according to the network condition, typically expressed in terms of the round–trip-time (RTT) and the packet loss probability, to achieve similar throughput that a TCP connection would require on the same path.

Multiple description coding (MDC) divides the video data into equally importance streams [2]. Each stream is transmitted from source to destination over different communication channels. In its simplest form it is assumed that the source is connected to the destination by two parallel channels. Therefore, the objective of MDC is to encode a source into two bit streams such that a high decoding quality is guaranteed when the two bit streams are received, while a lower, but still acceptable, decoding quality is achievable if only one stream is received. As a result of splitting the sequence into multiple correlated sub-streams, the coding efficiency is decreased. On the other hand, layered (scalable) coding splits the video sequence into a base and number of enhancement layers. Transmission of scalable video is inefficient when channel prioritization is not possible; as in the case of the current Internet transmission. Therefore, MDC is popular in the field of Internet video streaming applications.

There is a great deal of interest in designing practical MDC schemes. MDC using interleaved quantizer is one of these schemes where the coder applies an offset quantization with 2Δ step size for each description [3]. Thus, effectively Δ quantization step is achieved when both descriptions are received. Pair-wise correlating transforms is another MDC scheme such that (A, B) coefficients are unitary transformed to (C, D) [4]. Therefore, when C or D is received, (A, B) coefficients could be estimated using a linear predictor. In addition, multiple description motion compensation splits the sequence into odd and even frames. Allowing the encoder to perform prediction from both past even and odd frames [5].

Unfortunately, most MDC schemes achieve a reasonable balance between redundancy and distortion by modifying several components of the existing single description coding (SDC). Therefore, we are interested in proposing a low complex MDC scheme that could be integrated to existing SDC. Independent descriptions that are coded by the existing SDC represent the simplest MDC scheme. Although spatial/temporal sub-sampling produces independent descriptions, redundancy is an important issue. During the entire period of video transmission over Internet, the packet loss ratio and available bit rate vary widely. Moreover, the video content could change from low to moderate or high motion scenes. Thus, the challenging issue is to design an MDC scheme that intelligently adjusts the coding scheme according to not only the channel conditions but also the visual content. In this paper, we propose such an automatic adaptive scheme that switches between two simple MDCs depending on the channel loss, channel bandwidth and visual activity.

The paper is organized as follows. Section 2 presents a comparison between spatial and temporal sub-sampling MDCs and addresses their transmission issues over an Internet channel. The proposed adaptive transmission scheme is presented and evaluated in Section 3. Finally, the conclusions are presented in Section 4 followed by references.

2. COMPARISON BETWEEN TEMPORAL/SPATIAL SUB-SAMPLING MDC

The framework of sub-sampling MDC schemes is illustrated in Fig1. The original video sequence is coded into a number of independently decodable streams, each with its own prediction process. Accordingly, the interleaving component could split the original video sequence into even and odd frames (temporal sub-sampling) or split the original video image into even and odd lines (spatial sub-sampling) [6]. Subsequently, each stream could be coded independently using two codecs or serially using a single codec that preserves the last two reference frames. As a result of transmission over packet lossy networks, a description or parts of it might be lost. Therefore, appropriate error concealment is applied which has access to both descriptions. The availability of all descriptions to the concealment procedure guarantees the best reconstruction quality [7].

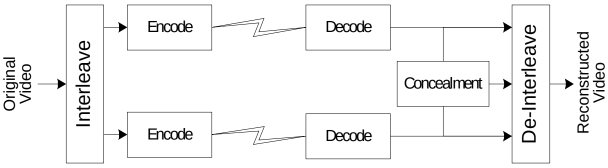

Fig. 1 General interleaving description video communication system

Intra-coded frames are placed at regular time instances so that error propagation due to inter-frame compensation is minimized. For spatial sub-sampling descriptions, I-frames are coded at the same instance as the SDC. While for temporal sub-sampling descriptions, I-frames are coded every even SDC I-frame. To allow intermediate nodes further regulating the rate in the event of congestion, the structure of the Group Of Picture (GOP) includes sufficient number of B-frames between anchor frames (I- & P-frame). In our simulation experiments, 2 B-frames have been placed between anchor frames for SDC scheme and spatial sub-sampling descriptions. To guarantee similar temporal dependency to SDC scheme, one B-frame has been placed between anchor frames for temporal sub-sampling descriptions. On the other hand, the transmission bit rate is split evenly between the two descriptions.

The performance of spatial/temporal sub-sampling MDC schemes using redundancy-rate-distortion (RRD) curves is shown in Fig 2. It is obvious that temporal sub-sampling MDC scheme is better than spatial sub-sampling by 1.5 dB for *Foreman* sequence and 0.5 dB for *Coastguard* sequence. Therefore, we conclude that temporal sub-sampling is an appropriate MDC scheme for loss-free transmission.

2.1. Packet lossy networks results

Gilbert model is widely used to emulate the Internet packet loss behavior [8]. For our simulation experiments, we vary the loss ratio from 0 to 0.15 and transmit both descriptions over similar channel conditions. To efficiently exploit the MDC scheme, an appropriate loss concealment algorithm is used. In the case of spatial sub-sampling MDC, a description contains either the even or odd image lines. Therefore, the average value of two consecutive lines of a correctly received description is used to predict the missing lines of the other description. Copying from the correctly received description is a simple concealment to missing blocks when temporal sub-sampling MDC is adopted. On the other hand, if both descriptions are lost, each description will be concealed independently. In this event, copying from the closest anchor frame conceals the missing blocks. The number of intra-macroblocks in a frame could detect a scene change. In the event of scene change, concealing missing macroblocks by copying from the closest anchor frame is inefficient. Thus, these missing macroblocks could be predicted from surrounding correctly received pixels such as interpolative concealment [7].

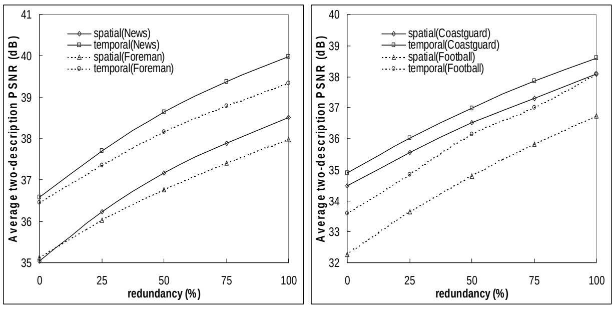

Fig. 2 RRD curves of spatial/temporal sub-sampling MDC for different video sequences

Our video sequences were MPEG-4 coded at CIF-resolution in progressive 4:2:0 YUV format at 25 frames/s. To ensure similar packet length, the resynchronization markers are approximately repeated every 4098 bits. Moreover, a TM5 rate control algorithm is used to adjust the bit rate. Since the visual content of a scene might contain natural or synthetic objects where their positions change gradually or rapidly, our test video set includes videos with different genres and content such as: head and shoulder, camera pan, zoom in, zoom out, waterfall, sport sequences, text scrolling, etc.

The average perceived visual quality (measured by PSNR) is illustrated in Fig 3. Based on these characteristic curves, we conclude the following points:

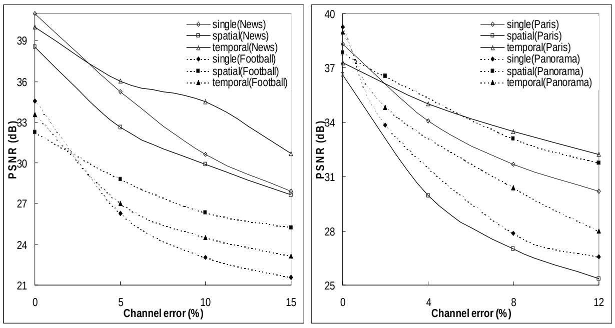

Fig.3 Characteristic curves of video sequences encoded at 1 Mbps (*news*), 3.5 Mbps (*football*), 1.5 Mbps (*paris*) and 2.5 Mbps (*panorama*)

1. The visual quality of SDC coding drops significantly with any packet loss ratio.

2. When the packet loss ratio increases, spatial sub-sampling MDC achieves better reconstruction quality in the case of moderate to high activity video sequences.
3. For specific video sequences, temporal sub-sampling MDC is better than spatial sub-sampling MDC until a specific channel loss ratio, which refers to as the critical point.
4. The value of the critical point changes for different bit rates and video contents.

Since TCP/UDP connections receive an acknowledgement packet from the destination every RTT, this information is used to predict the loss ratio and the transmission rate. In our simulation results, 50 frames are transmitted during the RTT (where RTT=2 seconds for 25 frames/s). Fig. 4 illustrates a detailed chart of the visual quality of the Foreman test sequence in each RTT. It is observed that adapting the coding scheme every RTT will enhance the video quality even if the channel status remains the same.

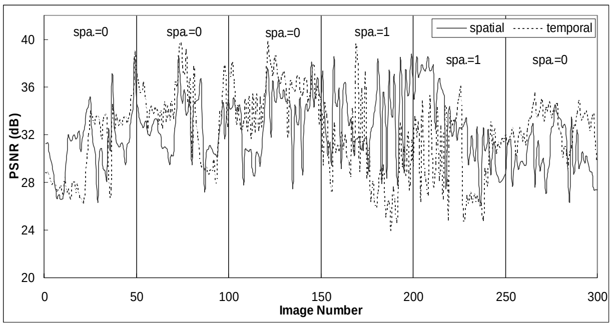

Fig. 4 Foreman sequence encoded at 1.25 Mbps and affected by 5% packet loss. Moreover, spa.=1 indicates that spatial MDC is better than temporal MDC scheme during RTT of 50 images.

These experiments reveal that significant robustness is guaranteed when MDC scheme is updated. Contrary to RRD results, the performance of temporal sub-sampling MDC no longer outperforms spatial sub-sampling MDC for every channel and visual conditions. Furthermore, packet loss ratio, transmission rate and visual content are basic parameters to determine the appropriate MDC scheme. Unfortunately, these parameters are not linearly related and hence require a sophisticated decision mechanism.

3. NEURAL NETWORK AS A DECISION SYSTEM

Artificial neural network techniques have been applied to solve complex problems in the field of image processing and image compression. Among the numerous neural networks, the multiplayer perceptron (MLP) network is particularly an efficient model for classification and prediction problems [9]. An MLP model, see Fig. 5, contains several hidden layers and the function of the hidden layer neurons is to arbitrate between the input and output of the neural network. The input vector is first fed into the source nodes in the input layer of the neural network. The neurons of the input layer constitute the input signals applied to the neurons of the hidden layer. The output signals of the hidden layer are used as inputs to the next hidden layer. Finally, the output layer produces the output results and then the neuron computing process is terminated. 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 SCG is benchmarked against the performance of the standard backpropagation algorithm [9]. On the other hand, it is a challenging issue to specify the optimal number of inputs that could produce the target outputs.

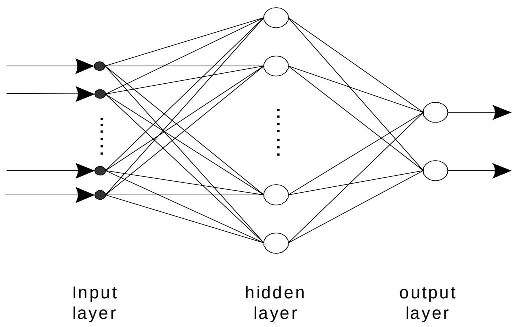

Fig. 5 MLP network composed of: 8-variable input vector, arbitrary number of hidden neurons, and two output neurons.

3.1. Video feature extraction

The activity between successive scenes could be captured through the motion vectors and the number of intra macro-blocks. The number of intra macro-blocks is an indicator of new objects or scene changes that occur, while the motion vectors express the type and direction of the movement. Moreover, these parameters are already generated by the MPEG-4 encoder and could be calculated off-line. For each image, the motion vector magnitude and ratio are calculated as follows.

$$mv = (x\_disp, y\_disp)$$

$$mv\_mag = \sqrt{x\_disp^2 + y\_disp^2}$$

$$mv\_ratio = \frac{mv\_max\_cnt}{num\_mb}$$

where *mv_max_cnt* is the number of *mvs* that are equal to the dominant *mv* and *num_mb* is the number of macro-blocks in the image.

The mean and variance of the number of intra macro-blocks, *mv_mag* and *mv_ratio* over 50 frames (i.e. RTT) is calculated. Furthermore, instead of expressing the bit rate in terms of the absolute value, it is recommended to express the bit rate as a ratio with respect to the maximum rate (i.e. when the quantization scale = 1).

$$rate\_ratio = \frac{current\_rate}{max\_rate}$$

Therefore, the MLP input vector is composed of 8 parameters: *rate_ratio*, channel loss ratio, the mean and variance of the number of intra macro-blocks, the mean and variance of the dominant motion vector, and the mean and variance of the ratio of the dominant motion vector. On the other hand, the output vector is composed of two binary values that represent the decision on the better coding scheme. Thus when the average perceived quality (over RTT) of spatial sub-sampling MDC is greater than temporal sub-sampling MDC, the output vector will be (1,0).

3.2. Simulation Results

Our video set is composed of 12 different CIF video sequences. The MLP network is trained through a set of examples where each example is composed of an input vector and the corresponding output results. Using the simulation results of Section 2, we could specify the output results for each input vector. In addition, the input vectors cover diverse range of values for any video sequence. Subsequently, the training examples of 11 video sequences are selected to train the MLP network off-line. When the network is fully trained, the input vectors of the test (not included in the train set) video sequence are applied to the MLP network. Hence, the network outputs could be compared against the correct decisions obtained through simulation in Section 2. To measure the performance of the MLP network, the percent of correct decisions over different rates and loss ratios are calculated. Table 1 summarizes the accuracy of the network outputs over different test sequences. The network decisions constitute the adaptive MDC scheme. Fig 6 illustrates the characteristic curves including the results of the adaptive scheme. It is obvious that the MLP network is capable of detecting the critical point and switches from temporal to spatial sub-sampling MDC, see *foreman* and *football* sequence results on Fig 6. Moreover, the adaptive scheme considers the underlying visual activity in determining the appropriate MDC scheme. The proposed adaptive scheme outperforms the SDC by about 2.2 dB at 2% packet loss of the *panorama* sequence and this difference becomes greater as the packet loss increases. Although, the average accuracy of correct classification is only 85%, the characteristic curves demonstrate the superiority of the proposed scheme. For example, the network performance of the *foreman* sequence is 67%, while the adaptive scheme gains about 0.5 dB more than temporal or spatial sub-sampling in the case of 5% channel loss, see Fig 6.

Table 1. Performance of the MLP network

Test Sequence	% correct decisions	Test Sequence	% correct decisions
News	93	Panorama	73
Foreman	67	Paris	99
Coastguard	83	Susie	87
Football	92	Fast food	70
Zoom in	88	Water fall	92
Scroll text	83	Water sport	90

4. CONCLUSION

In this paper, we propose an approach to dynamically switch the coding scheme based on the underlying visual content as well as channel conditions. The proposed scheme is resilient to packet-losses, since multiple description coding can be easily employed to provide correlated or independent layers that are transmitted over distinct channels. In our approach, we favor simple MDC schemes such as temporal/spatial sub-sampling MDC. Therefore, the proposed method can be easily implemented in conjunction with existing MPEG-4 codecs with negligible increase in complexity. Our simulation results demonstrate that under varying channel conditions and visual contents, adapting the MDC scheme provides a substantial improvement in performance over existing static MDC schemes. In addition, a neural network is designed to aid in the process of determining the appropriate MDC scheme. Through a video set which contains multiple objects and movements, we can achieve 85% accuracy of network decisions. The proposed scheme suggests not only updating the transmission rate every RTT but also modifying the coding scheme. In the future, we plan to conduct further studies to enhance the accuracy of the proposed system and to involve more than two MDC schemes in the process of decision-making.

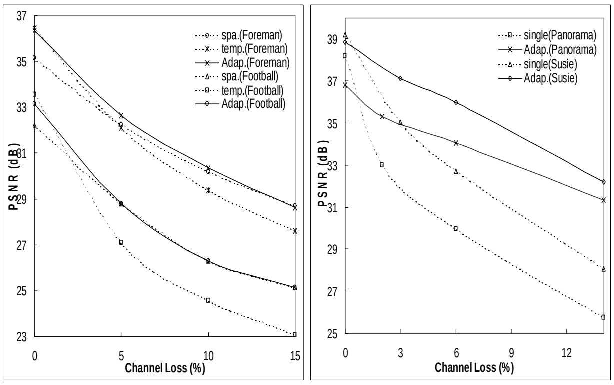

Fig.6 Characteristic curves of video sequences encoded at 1.25 Mbps (*foreman*), 3.5 Mbps (*football*), 1.75 Mbps (*panorama*) and 1.5 Mbps (*susie*)

5. REFERENCES

[1] Naoki Wakamiya, Masaki Miyabayashi, Masayuki Murata and Masayuki Murata, “MPEG-4 Video Transfer with TCP-Friendly Rate Control,” Springer-Verlang Berlin 2001 Volume 2216, pp 29-42

[2] V.K. Goyal, “Multiple description coding: compression meets the network,” IEEE Signal Processing Magazine, Volume: 18 Issue: 5, Sept. 2001 Page(s): 74 –93

[3] V. Vaishampayan and S. John, “Balanced interframe multiple description video compression,” in Proc. IEEE Int. Conf. Image Processing (ICIP99), Kobe, Japan, Oct. 1999.

[4] Yao Wang, M.T. Orchard and A.R. Reibman, “Multiple description image coding for noisy channels by pairing transform coefficients,” IEEE First Workshop on Multimedia Signal Processing 1997, Page(s): 419 –424

[5] Yao Wang and Shunan Lin, “Error-resilient video coding using multiple description motion compensation,” IEEE Transactions on Circuits and Systems for Video Technology, Volume: 12 Issue: 6, 2002 Page(s): 438 –452

[6] J.G. Apostolopoulos, “Error-resilient video compression through the use of multiple states,” Proc. of International Conference on Image 2000, Page(s): 352 -355 vol.3

[7] Y. Wang and Q.-F. Zhu, “Error control and concealment for video communication: A review,” Proc. IEEE, vol. 86, pp. 974–997, May 1998.

[8] S. Wenger, “Error patterns for internet experiments,” in ITU Telecommunications Standardization Sector, Oct. 1999, Document Q15-I-16r1.

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