NETWORK-ADAPTIVE FRAME-EXPANSION-BASED PACKET VIDEO CODING FOR ERASURE CHANNELS

Andrew G. Backhouse, *Irene Y.H. Gu*

Department of Signals and Systems,
Chalmers University of Technology, Gothenburg, 41296, Sweden
{andy, irenegu}@chalmers.se

ABSTRACT

This paper proposes a novel error-resilient packet video coding method which is designed to operate on unreliable IP networks suffering from large bursts of packet losses. The main novelties of the proposed system are: (1) A Bayesian packet loss predictor is employed to dynamically predict the packet loss probabilities; (2) A harmonic frame-expansion is applied to the video stream to add error-resilience. This includes dynamically choosing the size of the expansions. (3) Error-propagation is reduced by filtering out error-prone high-frequency components from reference images. This means that the proposed video codec method does not need to switch between the inter and intra coding modes. (4) Theoretical analysis of the video performance is given. Our preliminary simulations have shown that the frame expansion method can provide excellent resilience to video packet losses. The system has generated high quality videos at low bit-rates (64kbs) on simulated networks with the packet loss probability ranging from 5% to 20%. Visual inspection has shown that under similar PSNRs, the resulting videos have noticeable less visual artifacts as compared with those from the conventional codecs.

Keywords: frame expansions, video communication, error-resilient video coding, joint source-channel coding

1. INTRODUCTION

Efficiently transporting compressed packet video over an error prone IP network remains a challenging issue. Packet losses in IP networks often occur in bursts [7]. Because the transmission of video is usually delay constrained, channel coding can not always absorb these bursts and is therefore not able to guarantee an error-free transmission. This problem is made worse by the reliance on reference frames in the source coding of video. If a single frame is not reconstructed correctly, then several future frames will be corrupted.

A recent study in [1] has provided a method for dynamically predicting the time-dependent packet loss probability distribution. This is done by inferring congestion from the variations in the packet delay times observed at the receiver. From the prediction of network congestion and the packet loss probability, a better network-adaptive packet video compression and transmission strategy may be employed. Based on this method, we present a novel error-resilient video codec using frame expansions, which is designed to utilise knowledge of the full probability distribution of losses.

Recent studies on error-resilience for video include the introduction of multiple description (MD) coding [2]. While MD has many different forms, it is primarily used for encoding data with two data streams. In error-resilient video coding, redundancy has been added in a variety of ways such as correlating transforms, frame expansions, forward error correction (FEC) and MD scalar quantizers. Another approach which does not add redundancy is the domain-based MD coding. Encoding arbitrary sources has been done with harmonic frames [6]. Frame-expansions have been applied to image coding by using filter banks in [3]. In this paper we employ the frame-expansion proposed in [6] for the design of a new end-to-end packet video coder. The frame-expansions operate on sets of DCT-related coefficients. New mathematical expressions are given for the expected errors when the packet loss probability is known. Error-propagation is reduced by filtering out the high-frequency components in reference DCT blocks. The inter/intra coding modes used in conventional video coding are then no longer required. The proposed frame-based video coding system introduces robust packet error correction capability based on the prediction of packet loss rates.

The remaining paper is organized as follows. In Section 2 a basic overview of the video coding system is described. Section 3 describes some main methods that are proposed and employed in the system. These include the prediction of the packet loss probability using a Bayesian framework, DCT domain motion compensation, the harmonic frame expansions. Section 4 is contributed to the theoretical analysis of end-to-end video performance, including the analytic expression of the expected MSE. Further implementation details for network-adaptive video coding are given in Section 5. In Section 6, we describe the simulations of the proposed system, and some results and preliminary performance evaluation are included. Finally discussions are given in Section 7.

2. SYSTEM DESCRIPTION

The video codec is roughly based on the H263+ standard and is schematically illustrated in Fig.1. An image sequence is sent by a user across a lossy network to a second user. During the transmission, the second user gathers information based on transmission delays and losses which is transmitted back to the first user. Based on this information, the packet loss probability distribution is derived (Section 3.1). Each image is first split into blocks and a DCT transform is applied to all blocks. Next the DCT coefficients are motion compensated using a simple block-based motion compensation scheme. (Section 3.2). Then, a frame-expansion transformation is applied which maps the low dimensional source codes to

0-7803-9332-5/05/$20.00 ©2005 IEEE

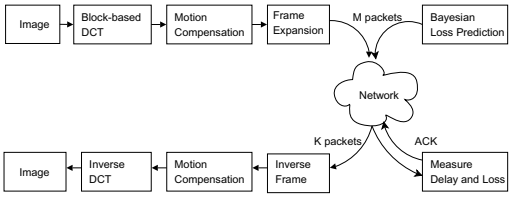

Fig. 1. Block diagram of the proposed video system.

higher dimension ones by introducing redundancy. The choice of transform is based on the time-varying network conditions which are gathered from the feedback process. The data is quantized and encoded and then transmitted to the network. In the receiver side, the inverse of the frame-expansion is applied. The inverse transform is dependent on the number of packets which arrive. The video is then reconstructed with the inverse video decoding methods.

As can be seem from the block diagram of Fig.1, the most essential parts of the proposed system are the Bayesian packet loss prediction, the optimal forward and inverse frame-expansion and the motion compensation.

3. VIDEO CODING

3.1. Bayesian Packet Loss Prediction

In [5], the authors propose an empirically-based method for describing time-varying traffic by measuring variations in end-to-end delay. In [1], this principle is extended to dynamically predict packet losses on the network. One of the essential parts of this work is to make use of this technique to optimise the error-resilience of our video coder. The technique can be briefly summarized as follows. Based on feedback of delays and losses in the network, the packet loss probability in the channel can be dynamically estimated. When packets are passed through a network, they traverse through multiple links and routers. Congestion occurs at routers where data from several links merge. To cope with varying loads, the routers usually maintain buffers. Once a buffer is full, packets will be lost. The probability that the packets are lost can therefore be predicted. Suppose we have transmitted N packets. We denote the set of transmitted packets by $\mathbf{N}$. Let $L(n) = 1$ if the nth transmitted packet is lost. Let $L(\mathbf{N})$ be the corresponding vector. Let $\rho(n)$, denote the ratio of the competing traffic arrival rate to the service rate at the time of the transmission of the nth packet. The expectation of $L(n)$ is then given by

$$E(L(n)) > \max\left(1 - 1/\rho(n), 0\right). \tag{1}$$

If the time interval between the transmission of the $n-1$th and $n$the packet is $\tau(n)$, then the expected variation in delay equals

$$E(d(n)) = (1 - \rho(n))\tau(n). \tag{2}$$

Obtaining a probability distribution on the number of losses during the transmission of the next M packets is obtained by associating (1) and (2). In [1], a Bayesian prior is defined for possible time-varying properties of the channel $\rho(n)$. Here $\rho(n)$ is assumed to be polynomial $\rho(n) = \sum_i a_i n^i$ and an a priori distribution $P(\mathbf{a})$ is defined for $\mathbf{a} = \{a_i\}$. The a posteriori distribution of $\mathbf{a}$ is obtained by solving

$$P(\mathbf{a}|L(\mathbf{N}), d(\mathbf{N})) \propto P(L(\mathbf{N}), d(\mathbf{N})|\mathbf{a})P(\mathbf{a}). \tag{3}$$

Thereby it is possible to obtain the probability that K packets in the next M will be lost

$$P(K|M, L(\mathbf{N}), d(\mathbf{N})) = \int_{\mathbf{a}} P(K|\mathbf{a})P(\mathbf{a}|L(\mathbf{N}), d(\mathbf{N}))d\mathbf{a} \tag{4}$$

3.2. Motion Compensation (MC) in the DCT Domain

In most block-based video encoders, high rates of compression are achieved by MC. For each block, the most similar block in the previous frame is found and only the difference between them is encoded. To prevent mismatch between encoder and decoder, the encoder applies motion compensation against the reconstructed reference frame after quantization. This way, encoder and decoder have identical reference frames. However, when the transmission suffers from data losses, the use of motion compensation will suffer from error-propagation. In block-based coders this effect is reduced by "intra"-coding. However on very poor channels "intra"-coding is required very often. Error-propagation is most significant for high frequency components which cause nasty distortion at the edges of blocks. For this reason, we apply a slight modification to the standard motion compensation scheme.

After splitting the image-frame into a set of blocks $\mathcal{B}$ sized (8x8), block matching is applied, such that the most similar block in the previous frame with the motion vector (d_x, d_y) is found. Both blocks are DCT transformed. The DCT components of the reference block are then filtered. After they have been filtered, the difference between the new DCT components and the old filtered ones is taken. Mathematically stated, let $s^b_{i,j}(n)$ denote the DCT coefficient corresponding to the ith horizontal and jth vertical frequency of the bth block of image-frame n. Let the corresponding DCT coefficient of the reference block be denoted by $\hat{s}^b_{i,j}(n-1)$. The reference DCT coefficient is then multiplied by a scalar value $\alpha_{i,j}$, $0 < \alpha_{i,j} < 1$, and the difference $x^b_{i,j}(n)$ is encoded,

$$x^b_{i,j}(n) = s^b_{i,j}(n) - \alpha_{i,j}\hat{s}^b_{i,j}(n-1). \tag{5}$$

The scalar value $\alpha_{i,j}$ is chosen differently depending on the frequencies. We shall see later from (12) that because the relative difference between the quantization error and the variance is large for low frequency components, the frame-expansion provides good error-resilience. Additionally, low frequency components are highly correlated in time and therefore there is much to be gained from MC by choosing $\alpha_{i,j}$ close to 1. The opposite is true for the high-frequency components. Reconstructing a particular $s(n)$ is done with the MMSE which can be shown to be,

$$\hat{s}(n) = \gamma\left(\hat{x}(n) + \alpha\hat{s}(n-1)\right), \tag{6}$$

where $\gamma = \frac{\sigma_s^2}{\sigma_s^2 + \epsilon_x + \alpha^2\epsilon_s(n-1)}$ and ϵ_x corresponds to the MSE in the transmission of x, $\epsilon_s(n-1)$ corresponds the the MSE of the reference DCT coefficient and σ_s^2 is the variance of the DCT coefficient. Here the error from motion compensation is absorbed together with the transmission error ϵ_x.

3.3. Network-Adaptive Coding by Using Frame-Expansions

The frame-expansion is then applied to $x^b_{i,j}(n)$ in (5), and can be described as a linear mapping

$$y^b_{i,j}(n) = \mathbf{V}\, x^b_{i,j}(n) \tag{7}$$

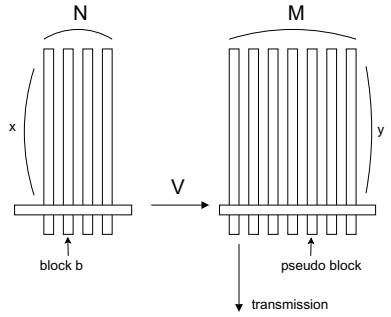

Fig. 2. Applying frame-expansion to the residuals $x_{ij}^b(n)$ in (5).

as illustrated in Fig.2.

The transform is illustrated in Figure 2. Each column on the left side of Figure 2 corresponds to the components $x_{i,j}^b$, $1 \leq i \leq 8, 1 \leq j \leq 8$ belonging to a single block. It is assumed that the different DCT coefficients within a block have unequal variances. However, it is assumed that the DCT coefficients corresponding to identical frequencies in different blocks have equal variances. Therefore the frame-expansion is independently applied to groups of coefficients corresponding to identical frequencies. This is illustrated in Figure 2 by the transform applied to the horizontal slice. The effect of the frame-expansion is to generate a new set of texture blocks $y_{i,j}^b(n)$. These are illustrated by the rightmost columns. These are then individually run length encoded and transmitted in separate packets. For a QCIF (144×176) image block, if a single image-frame is split into blocks of 8×8, then there will be 396 such blocks, which will in general be greater than the number of packets. The transform is therefore applied to sets of blocks. Every N blocks are mapped into M blocks. i.e. Let the image-frame be split into a set of blocks $\mathcal{B}$. In order to apply the frame expansion, these blocks of $x_{i,j}^b(n)$ are grouped into non-intersecting subsets $B_q \subset \mathcal{B}$, each having at most N elements. Let

$$x_{i,j}^{B_q}(n) = \left[x_{i,j}^{b_{q_1}}(n) \ \cdots \ x_{i,j}^{b_{q_n}}(n) \right]^T \tag{8}$$

where $B_q \subset \mathcal{B}$, $B_q = \{b_{q_1} \ \cdots \ b_{q_n}\}$. Let $\mathbf{V}$ be the operator of the frame expansion that maps an N-dimensional vector $x(n)$ onto an M-dimensional vector $y(n)$, the mapping in (7) can be described by,

$$y_{i,j}^{B_q}(n) = \mathbf{V}\, x_{i,j}^{B_q}(n) \tag{9}$$

A real harmonic tight frame that maps an N-dimensional space onto an M-dimensional space is defined by the following vectors [6]. Let f_{k+1}, $k = 0, 1, \ldots, M-1$ denote the row vector of $\mathbf{V}$ and be given as follows,

$$f_{k+1} = \begin{cases} \sqrt{\frac{2}{N}} \Big[\cos\frac{k\pi}{M}, \cos\frac{3k\pi}{M}, \ldots, \cos\frac{(N-1)k\pi}{M}, \\ \sin\frac{k\pi}{M}, \sin\frac{3k\pi}{M}, \ldots, \sin\frac{(N-1)k\pi}{M} \Big]^T, \text{ for even N} \\ \\ \sqrt{\frac{2}{N}} \Big[\frac{1}{\sqrt{2}} \cos\frac{2k\pi}{M}, \cos\frac{4k\pi}{M}, \ldots, \cos\frac{(N-1)k\pi}{M}, \\ \sin\frac{2k\pi}{M}, \sin\frac{4k\pi}{M}, \ldots, \sin\frac{(N-1)k\pi}{M} \Big]^T, \text{ for odd N} \end{cases} \tag{10}$$

3.4. Network Losses and Optimal Inverse Frame Expansions

Suppose M packets are transmitted. There are 2^M combinations of possible packet arrivals from a binary channel. Let Z denote the set of all possible packet arrivals, $z \subset Z$ denote a particular subset of arrived packets, and k_z denote the cardinality of z. We assume that the probability that k_z packets are lost, $P(k_z)$, is known in advance, or predicted using the method described in Section 3.1. The scalar quantization of each vector y which is transmitted yields $\hat{y}$. Suppose only the set of packets z arrive, then we denote the received vector by $\hat{y}_z$ which is a sub-vector of $\hat{y}$ and can be viewed as being generated by an operator $\mathbf{V}_z$ which is a sub-matrix of $\mathbf{V}$. Given the arrival of $\hat{y}_z$, the inverse frame operator which is optimal under the MMSE criterion can be derived,

$$\mathbf{V}_z^- = \mathbf{C}_x \mathbf{V}_z^T \left(\mathbf{V}_z \mathbf{C}_x \mathbf{V}_z^T + \mathbf{C}_{w_z} \right)^{-1}. \tag{11}$$

The residual vectors $x(n)$ can then be reconstructed by $\hat{x}(n) = \mathbf{V}_z^- \hat{y}_z(n)$ and subsequently one can reconstruct the DCT coefficients $\hat{s}_{i,j}(n)$ from (6).

4. THEORETICAL ANALYSIS OF THE EXPECTED MSE

4.1. End-to-End Expected MSE of $x(n)$

Let $y(n)$ be the frame-expansion of the residual vector $x(n)$ from (7) which are i.i.d. with zero-mean. Let $\hat{y}_z(n)$ be the received $y(n)$, $\hat{x}_z(n)$ be the reconstructed residual vector after applying the inverse frame expansion in (11). Then the end-to-end expected MSE between $\hat{x}(n) = \mathbf{V}_z^- \hat{y}_z(n)$ and $x(n)$ can be shown to be,

$$\epsilon_x = \frac{1}{N} \sum_{i=1}^{N} \frac{\sigma_x^2}{1 + \frac{\sigma_x^2}{\sigma_w^2} \lambda_{z,i}}. \tag{12}$$

where σ_x^2 is the variance of x, σ_w^2 is the variance of the quantization noise, and $\lambda_{z,i}$ are the eigenvalues of $\mathbf{V}_z^T \mathbf{V}_z$. As one can see, obtaining an expression of the error requires finding the eigenvalues. Although in principle these can be solved and stored, the number of matrices $\mathbf{V}_z$ corresponding to possible combinations of arrived packets makes it impractical. Therefore, we approximate the sum in (12) by an integral and the eigenvalues by a piecewise linear approximation $\hat{\lambda}_z(u)$. When the packet loss probability is given, the full MSE ϵ_x can be solved integrating the approximation of (12) with respect to the packet loss probability, i.e.,

$$\epsilon_x = \frac{1}{N} \int_{z \in Z} \int_0^N \frac{\sigma_x^2}{1 + \frac{\sigma_x^2}{\sigma_w^2} \hat{\lambda}_z(u)} du \, P(z) dz, \tag{13}$$

which can be solved analytically.

4.2. End-to-End MSE of $s(n)$ and PSNR

From the analytical approximation of the end-to-end MSEs of residual vector $x(n)$ in Section 4.1, we can obtain the end-to-end expected MSEs between the original DCT coefficients $s(n)$ and the reconstructed ones $\hat{s}(n)$ by using (6). The MSE of an individual DCT coefficient in this case can be shown to be

$$\epsilon_{s_{ij}}(n) = \frac{\left(\epsilon_{x_{ij}} + \alpha^2 \epsilon_{s_{ij}}(n-1) \right) \sigma_{s_{ij}}^2}{\epsilon_{x_{ij}} + \alpha^2 \epsilon_{s_{ij}}(n-1) + \sigma_{s_{ij}}^2}, \tag{14}$$

where $s_{ij}(n)$ denotes the DCT coefficient in the ith row and jth column of the nth frame. Subsequently, we can obtain the average expected MSE of a (8×8) block as, $MSE(n) = \frac{1}{64} \sum_{i,j=1}^{8} \epsilon_{s_{ij}}(n)$.

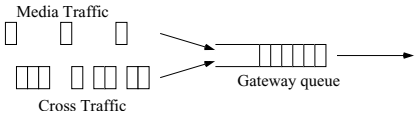

Fig. 3. The simple network model used in the simulations

5. NETWORK ADAPTIVE VIDEO CODING

For each received transmitted packet, the receiver is sends an acknowledgement (ACK). The round trip time (RTT), defined as the time duration between the transmission and the ACK of a packet, can vary significantly. When there is significant congestion RTT can be as large as 300 ms. Therefore, it is often the case that no ACKs have been received for several transmitted video-frames. Let r be the latest packet for which ACKs have been received, and n be the current frame. To obtain $\epsilon_{s_{i,j}}(n-1)$, the set of errors $\epsilon_{s_{i,j}}(r)$ are first updated using (12) and (14), $\epsilon_{s_{i,j}}(n-1)$ are then obtained using (13) and (14). To minimize the MSE of $\epsilon_{s_{i,j}}(n)$, the quantization error and scaling parameter α are jointly optimized. For a given expected number of packet losses, the frame-expansion ratio M/N is chosen such that $(M-N)/M$ is set as approximately the middle value within the range of $[p, 2p]$, which was chosen empirically.

6. SIMULATION RESULTS

In order to simulate the proposed video coding system, a simple network model is employed, as shown in Fig.3. To model the competing traffic in a real-network, cross-traffic is fed into the network that is generated by a time varying Poisson process. The traffic was manufactured so that the packet loss probabilities would range between $[0\%, 20\%]$ with a mean-value of 10%. ACKs were transmitted back to the receiver through a separate queue which was set to suffer minor losses. The RTT depended on the two queue size but varied between 10 and 210 ms. Tests have been performed on the 'foreman' image sequence in QCIF format (image size 176×144) and the sequence is encoded at a frame rate 15 fps. The average packet sizes was set to be 200 bytes which corresponds to 12 packets per frame. The factors $\alpha_{i,j}$ and N were chosen as described in Section 5. Fig.4 (a) shows the PSNR per frame, for the image sequence 'foreman' resulting from the simulation. As comparison, we also include the results from the MDMC coding scheme (See Fig. 10b in [4]) in Fig. 4(b). Comparing the results from the 2 methods, it is observed that there are much smoother fluctuations in the PSNR for the proposed method. Further, visual inspection was performed for assessing the reconstructed video quality. In Fig. 5, a reconstructed frame of the 'Foreman' sequence and the difference between the original is shown. We have observed that rather good video quality has been obtained. More performance evaluations will be conducted.

7. CONCLUSION

A novel network-adaptive packet video coding method is proposed which is designed to operate on unreliable erasure channels such as IP networks potentially suffering from large bursts of packet losses. The use of harmonic frame expansions has been introduced for enhancing the resilience of video streams against packet losses in erasure channels. Dynamically predicted packet loss probabilities are used to determine the best size for the frame expansion. Error-propagation is reduced and "intra" coding is avoided by filtering out the high frequency components from the reference frames. The proposed video codec is shown to be an efficient way of encoding packet video over erasure channels at low bit-rates. The derived analytic expressions of expected MSE allow the codec to maintain an estimate of the reconstructed image quality. These estimates can be used to dynamically adapt the coding parameters. The resulting PSNRs from simulation are shown to be rather smooth when video was transmitted over IP channels with a range of packet loss probabilities.

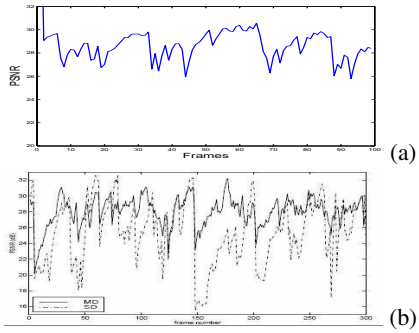

Fig. 4. Comparison for 'Foreman' , 10% packet loss. (a) Y frames encoded at 48kb/s using proposed method; (b) Entire video encoded at 64kb/s using MDMC, MD: Multiple Description, SD: Single Description.

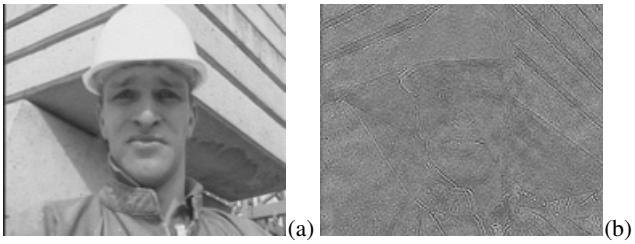

Fig. 5. (a) The reconstructed 60th image frame from the sequence 'foreman', the resulting PSNR was 30dB; (b) The residual image of (a) as compared with the original one.

8. REFERENCES

[1] Andrew Backhouse and Irene Y.H. Gu. A bayesian framework-based end-to-end packet loss prediction in ip networks. *IEEE Sixth International Symposium on Multimedia Software Engineering*, 2004.

[2] Vivek K. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74–93, 2001.

[3] Ravi Motwani. Tree-structured oversampled filterbanks as joint source-channel codes: Application to image transmission over erasure channels. *IEEE transactions on signal processing*, september 2004.

[4] A. Reibman, H. Jafarkhani, Y. Wang, M. Orchard, and R. Puri. Multiple description video coding using motion-compensated temporal prediction, 2002.

[5] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk. Multifractal cross-traffic estimation, 2000.

[6] J. Kelner V. Goyal, J. Kovacevi'c. Quantized frame expansions with erasures. *Journal of Appl. and Comput. Harmonic Analysis*, 2001.

[7] Maya Yajnik, Sue B. Moon, James F. Kurose, and Donald F. Towsley. Measurement and modeling of the temporal dependence in packet loss. In *INFOCOM (1)*, pages 345–352, 1999.