STREAMING VIDEO WITH OPTIMIZED RECONSTRUCTION-BASED DCT

Xiao Su and Benjamin W. Wah

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: {xiao-su, wah}@manip.crhc.uiuc.edu
URL: http://www.manip.crhc.uiuc.edu

ABSTRACT

One fundamental problem with streaming video data over unreliable IP networks is that packets may be dropped or arrive too late for real-time playback. Traditional error-control schemes are not attractive because they either add redundant information that may worsen network traffic, or rely solely on decoders with inadequate error concealment. This paper presents a joint sender-receiver approach in designing transforms for multiple-description coding in order to conceal network losses in streaming real-time video over the Internet. In the receiver side, we adopt a simple interpolation-based reconstruction, as sophisticated concealment techniques cannot be employed in software-based real-time playback. In the sender side, we design an *optimized reconstruction-based discrete cosine transform* (ORB-DCT) with the objective of minimizing the mean squared error, assuming that some of the descriptions are lost and that the missing information is reconstructed by simple averaging at the destination. Experimental results show that our proposed ORB-DCT performs better than the original DCT in real Internet tests. Future research includes finding perceptual-based quantization matrix based on extended basis images derived for reconstruction, and incorporating the effects of quantization and inverse quantization in the design.

1. INTRODUCTION

Increases in bandwidth and computational speed lead to growing interests in real-time video transmissions over the Internet. Unlike circuit-switched telephone networks, the Internet is a packet-switched, best-effort delivery service, with no guarantee on the quality of service. As a result, packets carrying video frames may be dropped or arrive too late to be useful for real-time playback.

Traditional coding algorithms for video compression are not robust to transmission errors. The sole objective of coding is to maximize coding gain, assuming error-free channels. Most video coding schemes rely on temporal-difference coding to achieve coding efficiency, thereby introducing a pervasive dependency structure into the bit stream. Losses due to dropped packets or late arrivals result in not only the loss of a bit stream itself but also the loss of subsequent dependent frames. Therefore, visual artifacts resulted from losses can be long lasting and annoying.

PROC. IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO., JULY 31-AUG. 2, 2000

Research supported by a grant from Conexant Systems, Inc.

To deliver video over the Internet in real time with high quality, an area of active research is to develop simple, effective and robust coding and error-concealment strategies. Most schemes found in the literature can be roughly grouped into two classes: layered coding and multiple-description coding.

In networks that provide transport prioritization, *layered coding* is effective for concealing network losses. In layered coding, video data is partitioned into a base layer and a few enhancement layers. The base layer contains visually important video data that can be used to produce video output of acceptable quality, whereas the enhancement layers contain complementary information that allows higher-quality video to be generated. In networks with priority support, the base layer is normally assigned a higher priority so that it has a larger chance to be delivered error free when network conditions worsen. Layered coding has been popular with ATM networks but may not be suitable for Internet transmissions for two reasons. First, the Internet does not provide priority delivery service for different layers. Second, when the packet-loss rate is high and part of the base layer is lost, it is hard to reconstruct the lost bit stream since no redundancy is present.

Unlike layered coding, *multiple-description coding* (MDC) [2, 3, 6] divides video data into equally important streams such that the decoding quality with any subset is acceptable, and that better quality is obtained by more descriptions. It is assumed that losses happen to different descriptions are uncorrelated, and that the probability of losing all the descriptions is small.

MDC has been implemented in several ways. In a scalar-quantizer design [3], two side-scalar quantizers are applied to produce two descriptions. A proper subset of index pairs formed from side quantizers are mapped to central-quantizer intervals in such a way that if both descriptions were received, the reconstruction error is minimized. The difficulties with this approach are that optimal index assignments are hard to achieve in real time, and that suboptimal approaches, such as A2 index assignment [3], introduce a large overhead in bit rate [7]. Instead of putting each pixel in every description, a *pair-wise correlating-transform* (PCT) [6] approach has been proposed to introduce correlations in each pair of transformed coefficients. The two coefficients resulted from PCT are put into two descriptions. This approach has high coding efficiency when both descriptions are available but has mediocre reconstruction quality with one description. However, in an error-prone environment like the Internet, the ultimate perceived quality is dominated by the reconstructed quality from one description.

Our proposed approach in this paper is MDC-based, with the goal of providing robust transmissions over the Internet without

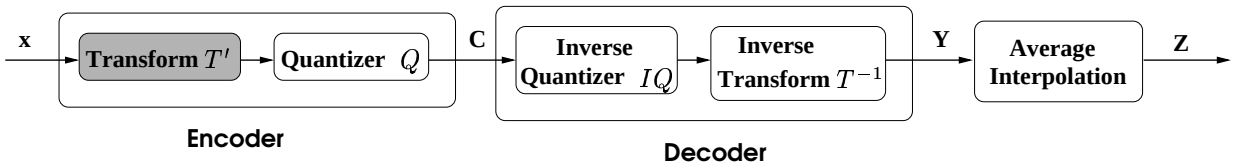


Figure 1: Basic building blocks of a modified codec. (The shaded block is our proposed ORB-DCT.)

supports for priority transmissions. We split adjacent pixels into multiple streams and code each separately. In contrast to previous approaches that design coding schemes at the sender, independent of reconstruction methods at the receiver, we design them in a joint fashion. At the receiver, we adopt a simple reconstruction algorithm based on average interpolation to facilitate real-time playback. At the sender, we design an *optimized reconstruction-based discrete cosine transform* (ORB-DCT) to minimize reconstruction error when some of the streams are lost and reconstructed using average interpolation from other streams at the receiver. The design of the codec is, therefore, tightly coupled to the reconstruction scheme to maximize the reconstruction performance. This approach leads to high reconstruction quality than that using one description, with only moderate increase in bit rate (about 20% to 30%) as compared to single-description coding.

This paper is organized as follows. Section 2 describes in details the design of the optimized reconstruction-based DCT. Section 3 presents the evaluation of the proposed ORB-DCT in both synthetic scenarios and realistic Internet transmission tests.

2. OPTIMIZED RECONSTRUCTION-BASED DCT

Assume that video data is partitioned into two descriptions, and that one of them is lost during transmission. In such a scenario, the original DCT and quantizer designs are not necessarily the best for reconstructing the lost data.

Figure 1 shows a simplified diagram of the basic building blocks in most state-of-the-art codecs. Our goal is to find a new transform T' in order to minimize the reconstruction error after interpolation, based on fixed quantization Q, inverse quantization IQ, and inverse DCT T^{-1}. That is,

$$\mathcal{E}_r = \| \underbrace{Interpolate(T^{-1}(IQ(\mathbf{c})))}_{decompression + reconstruction.} - \mathbf{x} \|^2 . \quad (1)$$

We assume inverse quantization IQ and inverse DCT T^{-1} to be fixed in order to keep our decoders standard-compliant. Consequently, our proposed transform coder can be used in real-time video-on-demand applications with standard-compatible decoders.

With quantization in place, the minimization of $\mathcal{E}_r$ becomes an integer optimization problem, where $\mathbf{c}$ in (1) takes integer values. Such optimizations are computationally prohibitive in real time. In the following, we derive an approximate solution that does not take into account quantization effects. Specifically, the objective to be optimized in the following approximation is:

$$\mathcal{E}_r = \| Interpolate(T^{-1}(\mathbf{c})) - \mathbf{x} \|^2 . \quad (2)$$

2.1. ORB-DCT for Intra-Coded Blocks

Assume that the original frame is divided into blocks of size 8×16 pixels. After ORB-DCT, block $\mathbf{X}$ is transformed into two blocks $\mathbf{C}_1$ and $\mathbf{C}_2$, each of size 8×8, corresponding to blocks of odd-numbered and even-numbered pixels, respectively. Since the derivations are similar, we only show the derivations for $\mathbf{C}_1$.

Our objective is to find $\mathbf{C}_1$ to minimize $\mathcal{E}_r$. After inverse DCT, output $\mathbf{Y}_1$ can be calculated as follows:

$$\mathbf{Y}_1 = \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i\mathbf{b}_j^T , \quad (3)$$

$$\text{where} \quad \mathbf{b}_i = \left\{ \frac{1}{2}\alpha_i \cos\frac{(2k-1)(i-1)\pi}{16} \right\}_{k=1,2,\ldots,8},$$

C_{ij} is the $(i,j)^{th}$ element in $\mathbf{C}_1$, $\mathbf{b}_i$ is the i^{th} basis vector of DCT, $\alpha_1 = \frac{1}{\sqrt{2}}$, and $\alpha_i = 1$ for $i = 2, 3, \ldots, 8$.

Putting (3) in matrix form gives:

$$\mathbf{Y}_1 = (\mathbf{p}_1 \quad \mathbf{p}_2 \quad \ldots \quad \mathbf{p}_8)_{8\times 8} , \quad (4)$$

$$\text{where} \quad \mathbf{p}_k = \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i b_{jk} \qquad k = 1, 2, \ldots, 8, \quad (5)$$

b_{jk} is the k^{th} component of basis vector $\mathbf{b}_j$. The interpolated pixels $\mathbf{Z}$ is then obtained by inserting even-numbered columns as the average of columns from $\mathbf{Y}_1$, with the boundary column duplicated:

$$\mathbf{Z} = \left(\mathbf{p}_1 \quad \frac{\mathbf{p}_1 + \mathbf{p}_2}{2} \quad \mathbf{p}_2 \quad \frac{\mathbf{p}_2 + \mathbf{p}_3}{2} \quad \ldots \quad \mathbf{p}_8 \quad \mathbf{p}_8 \right)_{8\times 16} . \quad (6)$$

$\mathbf{Z}$ can also be expressed as:

$$\mathbf{Z} = \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i\mathbf{e}_j^T , \quad (7)$$

$$\text{where} \quad \mathbf{e}_j = \left(b_{j1} \quad \frac{b_{j1} + b_{j2}}{2} \quad b_{j2} \quad \ldots \quad b_{j8} \quad b_{j8} \right)^T \quad (8)$$

We define $\mathbf{e}_j$ as an extended basis vector for reconstruction purpose. The distortion between the original and the received and reconstructed pixels is:

$$\mathcal{E}_r = \| \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i\mathbf{e}_j^T - \mathbf{X} \|^2 . \quad (9)$$

To minimize $\mathcal{E}_r$ with respect to $\mathbf{C}$, we first linearize each matrix into a vector by raster-scan order, *i.e.*, following the first row by the second row in a matrix, and so on. The following notations are defined after linearization:

$$\begin{aligned} \vec{\mathbf{u}} &= (C_{ij})_{(8\times 8)} \\ \vec{\mathbf{v}}_{8(i-1)+j} &= \mathbf{b}_i\mathbf{e}_j^T{}_{(8\times 16)} \qquad i, j = 1, 2, \ldots, 8 \\ \vec{\mathbf{w}} &= (X_{ij})_{(8\times 16)} . \end{aligned}$$

We further define matrix $\mathbf{V}$ as:

$$\mathbf{V} = (\vec{\mathbf{v}}_1 \quad \vec{\mathbf{v}}_2 \quad \vec{\mathbf{v}}_3 \quad \ldots \quad \vec{\mathbf{v}}_{64}) . \quad (10)$$

Then (9) can be rewritten as follows:

$$\mathcal{E}_r = \| \mathbf{V}\vec{\mathbf{u}} - \vec{\mathbf{w}} \|^2, \tag{11}$$

where $\mathbf{V}$ is a 128×64 matrix, $\vec{\mathbf{u}}$, a 64×1 vector, and $\vec{\mathbf{w}}$, a 128×1 vector. Since the linear system of equations $\mathbf{V}\vec{\mathbf{u}} = \vec{\mathbf{w}}$ is an over-determined one, there exists at least one least-square solution $\vec{\mathbf{u}}$ that minimizes (11) according to the theory of linear algebra [1]. Specifically, the solution $\vec{\mathbf{u}}$ with the smallest length $|\vec{\mathbf{u}}|^2$ can be found by first performing SVD decomposition of matrix $\mathbf{V}$:

$$\mathbf{V} = \mathbf{S}\,[diag(w_j)]\,\mathbf{D}^t, \qquad j = 1, 2, \ldots, 64, \tag{12}$$

where $\mathbf{S}$ is a 128×64 column-orthogonal matrix, $[diag(w_j)]$, a 64×64 diagonal matrix with positive or zero elements (singular values), and $\mathbf{D}$, a 64×64 orthogonal matrix. Then the least-square solution can be expressed as:

$$\vec{\mathbf{u}} = \mathbf{D}\,[diag(1/w_j)]\,\mathbf{S}^T\,\vec{\mathbf{w}}. \tag{13}$$

In the above diagonal matrix $[diag(1/w_j)]$, the element $1/w_j$ is replaced by zero if w_j is zero. Therefore, ORB-DCT is a product of three matrices: $\mathbf{D}\,[diag(1/w_j)]\,\mathbf{S}^T$.

To derive the ORB-DCT transform for $\mathbf{C}_2$, simply replace $\mathbf{e}_j, j = 1, 2, \ldots, 8$, in (8) by the following:

$$\mathbf{e}_j = \begin{pmatrix} b_{j1} & b_{j1} & \frac{b_{j1}+b_{j2}}{2} & b_{j2} & \ldots & \frac{b_{j7}+b_{j8}}{2} & b_{j8} \end{pmatrix}^T.$$

The rest of the steps are similar.

2.2. ORB-DCT for Inter-Coded Blocks

For inter-coded blocks, output $\mathbf{Y}_1$ after inverse DCT, as shown in (3), is the residual block after motion prediction. Denote its corresponding reference block as:

$$\mathbf{R} = \begin{pmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \ldots & \mathbf{r}_8 \end{pmatrix}_{8\times 8}. \tag{14}$$

Then the interpolated data $\mathbf{Z}$ is the sum of two terms after motion compensation:

$$\begin{aligned} \mathbf{Z} &= \begin{pmatrix} \mathbf{p}_1 & \frac{\mathbf{p}_1+\mathbf{p}_2}{2} & \mathbf{p}_2 & \frac{\mathbf{p}_2+\mathbf{p}_3}{2} & \ldots & \mathbf{p}_8 & \mathbf{p}_8 \end{pmatrix} \\ &+ \begin{pmatrix} \mathbf{r}_1 & \frac{\mathbf{r}_1+\mathbf{r}_2}{2} & \mathbf{r}_2 & \frac{\mathbf{r}_2+\mathbf{r}_3}{2} & \ldots & \mathbf{r}_8 & \mathbf{r}_8 \end{pmatrix} \\ &= \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i\mathbf{e}_j^T + \mathbf{R}'. \end{aligned} \tag{15}$$

Substituting the above equation into (9) results in the reconstruction error for inter-coded blocks.

$$\mathcal{E}_r = \| \sum_{i=1}^{8}\sum_{j=1}^{8} C_{ij}\mathbf{b}_i\mathbf{e}_j^T - (\mathbf{X} - \mathbf{R}') \|^2. \tag{16}$$

To derive ORB-DCT in this case, we note that only vector $\vec{\mathbf{w}}$ is different as compared to the case of intra-coded blocks. From (13), it is obvious that the transform itself does not depend on $\vec{\mathbf{w}}$; therefore, ORB-DCT retains the same form.

In short, a uniform transform of ORB-DCT exists for both intra- and inter-coded blocks. For intra-coded blocks, it is applied to an original block $\mathbf{X}$ to produce transform coefficients $\mathbf{C}_i, i = 1, 2$; whereas for inter-coded blocks, it is applied to interpolated motion-predicted blocks $(\mathbf{X} - \mathbf{R}')$.

Like DCT, ORB-DCT is also a row-column-separable transform. To compute a transform coefficient of ORB-DCT by a row-column approach, it takes 40 floating-point multiplications and 37 floating-point additions. In the future, we plan to study fast implementations of ORB-DCT, similar to what was done in deriving fast DCT.

2.3. Handling Longer Burst Lengths

In the above derivations, video frames are assumed to be partitioned into two descriptions. However, from the traffic study we have conducted on loss characteristics, we have found that two descriptions may not always be sufficient to conceal losses for transcontinental connections [4, 5]. In those connections, we may need four descriptions that are constructed by a combination of 2-way interleaving along both horizontal and vertical directions, in a way similar to that described in [5].

3. EXPERIMENTAL RESULTS

We have evaluated our proposed ORB-DCT in two scenarios: a synthetic scenario under controlled losses and real Internet tests. Our experiments were done using two video sequences in CIF (352×288) YUV format: *missa* (Miss America) consisting of 150 frames and *football* consisting of 90 frames. *Missa* represents a typical video conferencing sequence with slow head-and-shoulder movements, whereas *football* features a fast-motion movie.

We measure the reconstruction quality by the *peak signal-to-noise ratio* (PSNR). In the following, we only show the PSNR of the Y component (the dominant component in human perception).

We have made two modifications to the H.263 codec from Tenet RD (*http://www.nta.no/brukere/DVC/*) in our experiments. a) In the encoder, we use ORB-DCT in the transformation stage instead of DCT. b) In the decoder, we feed reconstructed frames back to the motion-compensation module for better decoding of future dependent frames when some frames within a description are lost but can be reconstructed using other descriptions.

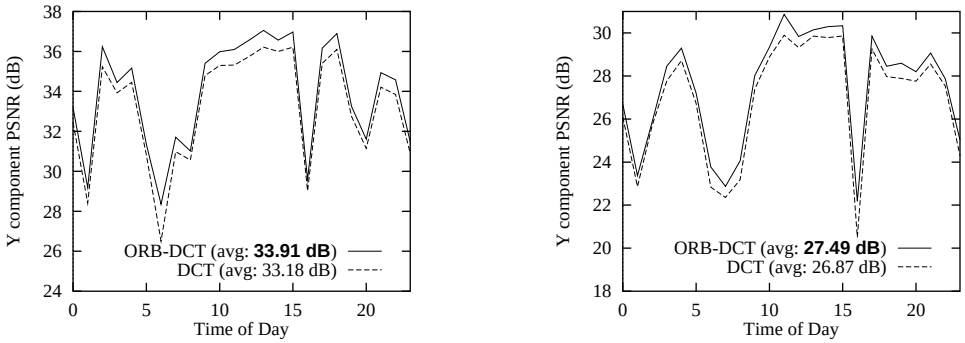
3.1. Quality Comparisons under Controlled Losses

In the following, we compare our proposed ORB-DCT with the original DCT, assuming that video data is divided into two descriptions along horizontal directions. Results along the vertical direction are similar and are not shown.

Table 1 compares the results for two cases. First, we assume that only the odd-numbered stream of the two descriptions is received. (Results of reconstruction from the even-numbered stream are similar.) Improvements due to the new transform for both sequences are around 0.4 dB. Second, we assume that both descriptions are received. In this case, ORB-DCT introduces a negligible degradation (no greater than 0.07 dB) as compared to DCT. The PSNR values are calculated at the same bit rate ($20\% - 30\%$ more than single-description coding) for the two transforms, both applied to two interleaved streams.

3.2. Tests on the Internet

We compare in this section the reconstruction quality of both transforms – DCT and ORB-DCT – based on tests on the Internet.



a) *missa* sequence b) *football* sequence

Figure 2: Reconstruction qualities over a 24-hour period for the Champaign-China connection.

Table 1: Comparison of reconstruction quality of video sequences transformed by DCT and ORB-DCT using two descriptions. Results are shown for two cases: a) the odd-numbered description is received, and b) both descriptions are received.

Video Sequence	PSNR (dB)			
	a) odd received		b) both received	
	DCT	ORB-DCT	DCT	ORB-DCT
Missa	36.20	**36.61**	**36.74**	36.70
football	29.43	**29.82**	**30.16**	30.09

For a fair comparison of both algorithms under the same traffic conditions, we did trace-driven simulations by applying the same trace of packets lost in real Internet transmissions on sequences transformed by DCT and ORB-DCT.

In collecting traffic traces, we sent 512-byte packets periodically from our home site in Champaign to a remote echo server in China (`public.qd.sd.cn`) at a rate of 20 packets/sec. The transmission rate was chosen in such a way that it did not impose an excessive network load. From the packets echoed back, we recorded the sequence numbers and sending and arrival times and determined packet losses based on the sequence numbers recorded. The packet-loss rate estimated was likely to be pessimistic since each packet traversed a round trip. We set the jitter-buffer size to be comparable to the standard deviation of packet inter-arrival times so that any packet that arrived later than its scheduled arrival time plus the jitter time was considered lost.

Our experiments to apply traffic traces consist of a sender process and a receiver process. The sender process was in charge of compressing and packetizing video frames, and mapping packet losses to GOB losses of each frame. The number of descriptions (2 or 4) was set periodically every 0.5 sec at the sender according to feedback information on GOB losses of frames from the receiver. In our simulations, we assume that the receiver collected GOB-loss information every 0.5 sec before sending the information to the sender, and that the network delay was constant at 0.5 sec. The receiver process read from the file of GOB losses saved by the sender, discarded the corresponding GOBs lost during transmission, decompressed the remaining coded streams, and deinterleaved them. For every GOB of each frame, any missing information due to packet losses was reconstructed by average interpolation using adjacent pixels. The reconstructed frame was sent back to the decoder as a reference for future inter-coded frames. If the entire GOB was lost, it was reconstructed by copying the corresponding GOB from the last received frame.

Figure 2 compares the reconstruction quality over a 24-hour period for the Champaign-China connection (collected on Nov. 19, 1999). Loss rates of this connection range from 10% to 45% in most cases. The new transform ORB-DCT yields better playback quality at all times. For the *missa* sequence, the average PSNR is 33.91 dB using ORB-DCT and 33.18 dB using DCT. For the *football* sequence, the average is 27.49 dB for ORB-DCT and 26.87 dB for DCT.

It is interesting to note that under real loss situations, the gain of the new transform for both the *missa* and *football* sequences are higher than the synthetic scenario shown in Table 1. This is not surprising because in real tests, we always fed the reconstructed frames that were lost back to the motion-compensation loop, and the improvement of the reconstruction quality due to the new transform accrued as the video was played. In contrast, in synthetic scenarios, feedbacks were not possible since one or more streams were completely lost.

In short, our results suggest that our proposed ORB-DCT transform works well in a lossy transmission environment, such as the Internet.

4. REFERENCES

[1] L. Rade and B. Westergren. *Mathematics Handbook for Science and Engineering*. Studentlitteratur Birkhauser, 1995.

[2] S. D. Servetto, K. Ramchandran, and V. A. Nahrstedt. Multiple-description wavelet based image coding. In *Proc. IEEE Int'l Conf. on Image Processing*, October 1998.

[3] V. A. Vaishampayan. Design of multiple description scalar quantizer. *IEEE Trans. on Information Theory*, 39(3):821-834, May 1993.

[4] B. W. Wah and D. Lin. Transformation-based reconstruction for real-time voice transmissions over the internet. *IEEE Trans. on Multimedia*, 1(4):342–351, December 1999.

[5] B. W. Wah and X. Su. Streaming video with transformation-based error concealment and reconstruction. In *Proc. Int'l Conf. on Multimedia Computing and Systems*, volume 1, pages 238–243. IEEE, June 1999.

[6] Y. Wang, M. T. Orchard, and A. R. Reibman. Multiple description image coding for noisy channels by pairing transform coefficients. In *Proc. IEEE First Workshop Multimedia Signal Processing*, pages 419-424, June 1997.

[7] Y. Wang and Q. Zhu. Error control and concealment for video communications: a review. *Proceedings of the IEEE*, 86(5):974-997, May 1998.