To appear in Proceedings of the SPIE: Human Vision and Image Processing 1998

Effects of Temporal Jitter on Video Quality: Assessment Using Psychophysical Methods

Yuan-Chi Chang [a], *Thom Carney* [b,c], *Stanley A. Klein* [b,c], *David G. Messerschmitt* [a] *and Avideh Zakhor* [a]

[a] Department of Electrical Engineering and Computer Sciences, UC Berkeley
[b] School of Optometry, UC Berkeley
[c] Neurometrics Institute
Berkeley, California 94720, USA

ABSTRACT

The conventional synchronous model of digital video, in which video is reconstructed synchronously at the decoder on a frame-by-frame basis, assumes its transport is delay-jitter-free. This assumption is inappropriate for modern integrated service packet networks such as the Internet for network delay jitter varies widely. Furthermore, multi-frame buffering is not a viable solution in interactive applications such as video conferencing. We have proposed a "delay cognizant" model of video coding (DCVC) that segments an incoming video into two video flows with different delay attributes. The DCVC decoder operates in an asynchronous reconstruction mode that attempts to maintain image quality in the presence of network delay jitter. Our goal is to maximize the allowable delay of one flow relative to that of the other with minimal effect on image quality since an increase in the delay offset reflects more tolerance to transmission delay jitter. Subjective quality evaluations indicated for highly compressed sequences, differences in video quality of reconstructed sequences with large delay offsets as compared with zero delay offset are small. Moreover, in some cases asynchronously reconstructed video sequences look better than the zero delay case. DCVC is a promising solution to transport delay jitter in low-bandwidth video conferencing with minimal impact on video quality.

Keywords: delay jitter, temporal masking, video compression

1 INTRODUCTION

In a global networking infrastructure, stringent delay requirement of interactive video applications such as collaborative video-conferencing poses challenges to the conventional, frame-by-frame, synchronous video rendering scheme. As the propagation delay alone, lower-bounded by the speed of light, can easily exceed 100 msec, significant delay jitter associated with packet networks further introduces artificially added delay due to the current scheme, which requires all video data to be present at the frame rendering moment. A straightforward solution that requires low network delay jitter for all video data can be very costly and inefficient in network resource (switching buffers and bandwidth) usage. We thus proposed a new video coding method to minimize the amount of video data that requires low delay jitter and relax the jitter constraint for the rest of the data. The new method will reduce its network resource usage, which directly reflects the connection cost, due to the minimized low-delay-jitter traffic.

This *delay cognizant* method, which we named delay cognizant video coding (DCVC), poses new research challenges to both video coding and human vision studies since we must abolish the aforementioned synchronous video rendering scheme and establish an *asynchronous* rendering scheme. The abolishment of synchronous rendering is necessary because relaxing the jitter constraint of some video data causes video data packets from the same frame to arrive at the receiving end at different instants (not within one frame display time). Under the old scheme, the packets must be realigned by adding artificial delay at the receiver, which increases the critical end-to-end delay. The increase in delay is unacceptable for interactive applications.

The asynchronous rendering scheme avoids the packet alignment and displays the video data as soon as the packets arrive at the receiver. As a result, video information from the same frame appears in different frames at the receiver. The breakup of rendering regularities can lead to video quality degradation. To minimize the negative impact on visual perception, the video coding algorithm must identify the *delay-critical* information so that the critical data can receive preferential treatment in the network with low delay jitter. However, to the best of our knowledge, there are no definition on what delay-critical is from the perspective of human vision nor established guidelines in segmenting those data from a video stream. The results reported in this paper represent an effort to address these issues.

The focus of the paper is to explore the effects of temporal jitter (delay) on video quality in this asynchronous video rendering scheme. Through psychophysical methods, video quality was assessed to evaluate the potential of both the general scheme and the specific video coding algorithm. The coding algorithm was based on informal

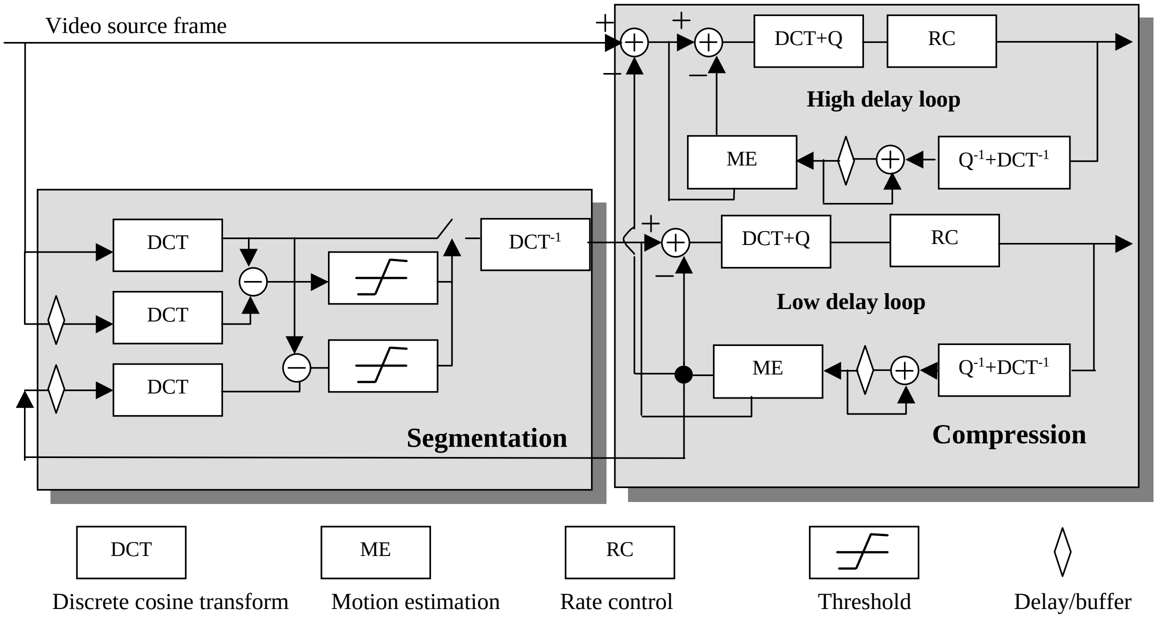

Figure 1 DCVC encoder block diagram.

testing and authors' knowledge of human visual function. Although we used the same coding algorithm for all video sequences, we assume with improved models of human vision we could devise a better algorithm to achieve optimal delay cognizant coding. One of our goals in this, and future studies is to determine parameters critical for DCVC.

In the past, we developed several different video coding algorithms [1]-[4] for DCVC with the goal of minimizing the delay-critical data flow while maintaining video quality. Techniques applied in segmenting those data from a video stream include comparisons in pixel and frequency domain as well as motion estimation. Segmented video data must be compressed for transmission efficiency and segmenting techniques, which lead to low compression performance, are avoided. Based on past experience, the current coding design utilizes segmentation of blocks in the frequency domain and compression of blocks by motion compensation and transform coding. The video encoder and decoder architectures are reported in the next section.

In Section 2, we will describe the DCVC algorithm, which encodes and decodes test video sequences used in our psychophysical experiments. The section is followed by the experimented methods, data collection and analysis in Section 3. Section 4 summarizes the paper.

2 ADAPTING TO TEMPROAL JITTER WITH DELAY COGNIZANT VIDEO CODING

As we have pointed out, in order to minimize video quality degradation, a DCVC algorithm must identify the *delay-critical* information and segment it from the rest of the video stream. The segmented data is grouped into and carried by network *flows* with different delay requirements. Although the number of flows is arbitrary, we assume that there are only two flows, the low- and high-delay flows. The low-delay flow carries the delay-critical data while the delay-tolerant data takes the high-delay flow. Our goal is to maximize the allowable delay of one flow relative to that of the other, since an increase in the delay offset reflects more tolerance to transmission delay jitter.[1]

The DCVC encoder shown in Figure 1 is divided into two stages: segmentation and compression. As the encoder segments the video information into low- and high-delay flows, it attempts to achieve the following two objectives simultaneously:

- Minimize total traffic, while maximizing that portion in the high-delay flow and minimizing that portion in the low-delay flow.

[1] The authors acknowledge that although network flows (connections) with delay guarantees are only experimental today, the next generation Internet Protocol (IPv6) has incorporated a 28-bit flow label into its packet header structure to support future deployment.

- Maximize the allowable delay offset (the difference between the maximum allowed delay of the high-delay flow and that of the low-delay flow), subject to the constraint that data from the high-delay flow still improves video quality at the time of its arrival.

At the segmentation stage, the encoder decides, for each block, which flow (low- or high-delay) should be assigned to carry the video data in the block. The decision criterion we use is conditional block replenishment. The block diagram of the segmentation stage is shown in Figure 1. A video source frame is first divided into 8 by 8 pixel blocks. For each block, its discrete cosine transform (DCT) is computed to obtain the transform coefficients. Each coefficient is tested against the following two conditions:

Condition 1:

$$\left|P_{i,j,n,t} - P_{i,j,n,t-1}\right| < V_{i,j}$$

Condition 2:

$$\left|P_{i,j,n,t} - P_{i,j,n,update}\right| < V_{i,j}$$

where $P_{i,j,n,t}$ is the (i, j)th coefficient for block n in frame t; $V_{i,j}$ is a fixed preset threshold for the (i, j)th coefficient; $P_{i,j,n,update}$ is the value from the last update block. Note that at this stage, the transform coefficients are not quantized and are stored in full precision. The first (second) condition corresponds to the upper (lower) threshold block in the diagram. If not all coefficients satisfy both conditions, which means the block has changed significantly, this block is declared a low-delay block, causing it to be fed into the low-delay loop. The segmentation stage extracts the low-delay blocks and puts them in an image plane. These frequency domain blocks are then inverse transformed back to the pixel domain before compression. The high-delay information is the difference between the original image and the image plane formed by low-delay blocks. An example in Figure 2 is given to show the original image, the low-delay image plane and the high-delay image plane. In this specific frame, approximately an equal number of blocks go to each flow. In the general case, the ratio depends on the motion content of the video.

The percentage of blocks in a frame carried by the low-delay flow depends on the DCT thresholds. Lowering the thresholds will increase the number of low-delay blocks while raising the thresholds has the reverse effect. Aggressive segmentation by setting high threshold values hurts the overall performance of the algorithm since visual artifacts appear as soon as the delay offset exceeds zero. On the other hand, too many low-delay blocks defeat the purpose of DCVC since the amount of traffic requiring low delay jitter is will be comparable to the output of a traditional, non-DCVC encoder.

Segmenting a video frame at the granularity of blocks is not always the best approach. Since blocks have artificially designated boundaries, choosing a suitable set of DCT thresholds needs to take into consideration possible pixel value variations in a block. This creates inefficiency when only a small portion of the block has changed but the whole block has to be sent to the low-delay flow. The most effective approach we found so far is to segment at the granularity of pixels. With a similar decision criterion, conditional pixel replenishment can extract most delay-critical information. The difficulty in this approach, however, is the compression of the addresses of selected pixels. The addressing overheads turned out to be the dominant contributor in bandwidth usage. The 8x8 block size is thus a tradeoff between effective segmentation and efficient compression.

The image planes at the output of the segmentation stage typically exhibit significant temporal redundancy and are thus differentially coded to reduce the bandwidth required. Motion estimation (ME) with block DCT [5][6] is used to remove the redundancy in the low-delay image plane. The high-delay image plane is obtained by subtracting the anticipated decoded low-delay image from the original video. This allows quantization errors in the low-delay ME loop (lower loop in the figure) to be passed as a part of the input to the high-delay ME loop (upper loop).

Quantized transform coefficients are run-level encoded to further remove their statistical dependency. Run-level codes are typically generated through a set of training sequences. Our experiment video sequences are standard ITU-T H.263[7] test clips for low bit-rate (< 64 kbps for 15fps, QCIF size video) coding. We thus decided to use the default setting of H.263 standard [7] in quantization

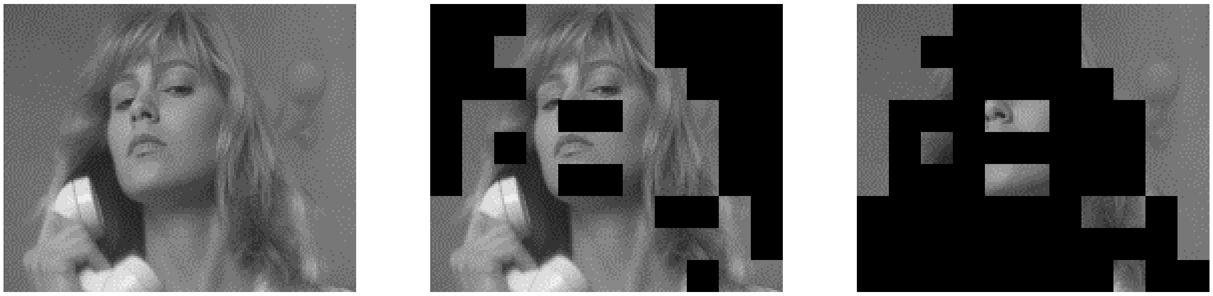

Figure 2 Left: original video frame; center: low-delay image plane; right: high-delay image plane.

levels and run-level codes.

Rate control modules shown in the encoder block diagram are disabled in encoding the test sequences used in the psychophysical experiments to simplify the control variables.

The DCVC decoder follows a simple set of rules to display received blocks. Compressed bit streams from both flows are tagged with frame numbers as temporal references. The decoder maintains one table for each flow, in which each entry stores the temporal reference of the received block at the coordinates. The tables are initialized to zero and replace blocks from earlier frames with those from later frames. By comparing $TR_{n,L}$, temporal reference of the *n*th block from the low-delay flow, and $TR_{n,H}$, temporal reference of the *n*th block from the high-delay flow, the decoder makes the following decision:

- $TR_{n,L} > TR_{n,H}$, display the block from the low-delay flow;
- $TR_{n,L} = TR_{n,H}$, display the sum of two blocks;
- $TR_{n,L} < TR_{n,H}$, display the block from the high-delay flow.

3 PSYCHOPHYSICAL AND HUMAN VISION MODEL BASED QUALITY ASSESSMENT OF DCVC VIDEO

3.1 Video fidelity

Casual observation by the authors indicated that a few frames of delay using DCVC might be indiscriminable from the zero delay case. Therefore we first evaluated the effects of temporal jitter on video fidelity. Specifically, we are interested in knowing when compared with the original, jitter free video rendering, if sequences with some temporal jitter (nonzero delay offset between the two flows) are visually discriminable.

The video sequences used in the experiments were composed of the luminance components of standard H.263 test clips: Suzie, Salesman, and Mother-daughter (see section 3.2.1 for details). Both compressed and uncompressed sequences were used. A standard self-paced psychophysical method of adjustment was used with 3 delay offsets (0, 1, and 2 frames). Each run consisted of 100-150 trials with correct response feedback provided after each trial.

The subjects included vision science professionals as well as naïve observers, who have never worked in related areas. We found their judgements on fidelity to be unanimous: for both uncompressed and compressed video sequences, asynchronous rendering does not preserve video fidelity, delays of one frame can be detected. This result is surprising since aggressive segmentation ought to create discriminable visual differences.

Interestingly, we noted that while delays can be detected, the video quality of compressed sequences did not necessarily degrade. In fact for some compressed sequences and some observers, the quality increased with relatively long delays. This effect appears to be related to a reduction of mosquito noise when delay is introduced. This observation prompted us to examine video quality in more detail using DCVC sequences. Our findings are described in the following section.

3.2 Video quality

From the fidelity experiments, we concluded that for both uncompressed and compressed video sequences, with nonzero delay jitter, are discriminable from their counterparts with zero delay jitter. In this set of the experiments, we quantify the impact of temporal jitter on video quality using psychophysical and computational methods.

3.2.1 Methods

The seven raw video sequences used in the experiments are standard H.263 test clips: Carphone, Claire, Foreman, Miss America, Mother-daughter, Salesman, and Suzie. Their image sizes are in the 4:2:0 QCIF format (176 by 144). To simplify our studies, in both fidelity and quality experiments, only the luminance component of the video is extracted and used. Each sequence is 2.5 seconds long (75 frames) and is presented on a Sony Trinitron monitor at 60 Hz (two scans / frame). Matlab with the PC-MatVis psychophysical testing and stimulus presentation extension[2] was used to control the experiment .

Among the encoded sequences, the number of low-delay blocks is between 10 to 20 percent of the total. The actual percentage is content dependent. For nonzero delay jitter sequences, we applied the same amount of delay, in the units of frame display time (1/30th of a second), to the video data in the high-delay flow.

Unlike the fidelity experiments, this set of experiments focuses on evaluating the quality of compressed video only. In particular, we are interested in the compression-introduced masking effect on the nonzero delay jitter sequences. From the fidelity experiments, we know that introducing delay jitter introduces visual artifacts in uncompressed sequences. It is also known that lossy compression generates quantization noise. Furthermore, we observed that the noise contributed by compression seems to be stronger and dominates the perception of the overall video quality. Our goal in these experiments is to

[2] MatVis information see: http://members.aol.com/neuromet/index.html

compare the relative effects of delay jitter and lossy compression on video quality.

Compression level: The first frame (the only I frame) of all four sequences (four stimulus conditions) contains identical information. The amount of compression-introduced noise in subsequent frames is controlled by the quantization level (QL). All 64 DCT coefficients of a nonintra-coded block are quantized with the same step size. Increasing the quantization level decreases the video quality nonlinearly and vice versa.

The test sequences are generated with four stimulus conditions. In stimulus conditions 1 and 2, QL is set to 10; in stimulus conditions 3 and 4, QL is set to 12 to compress Salesman, Mother-daughter, and Miss America while the other four are compressed with QL equal to 13. Depending on the video content, a decrease of QL from level 12 to 10 increases the compressed bit rate by 20 to 50 percent.

Delay level: Jitter free video rendering applies to conditions 1 and 3. A delay offset of 12 frames (~400 milliseconds) between the low- and high-delay flows is applied to stimulus conditions 2 and 4.

The procedure of evaluating video sequence quality involves the following 3 steps.

1) All four stimulus conditions are presented simultaneously, two across and two down on the screen. Stimulus condition position is chosen randomly. The 2.5-sec sequence presentation is repeated 10 times (additional viewing time is available as desired by the subject). The subject is asked to rank order the 4 sequences for quality using there own subjective criteria for quality.

2) Next, each of the four stimulus conditions are presented separately in random order for a total of 20 trials, 5 for each condition. Each stimulus presentation lasts 2.5 seconds (two repeats). After each stimulus presentation the subject is asked to rate the image quality from 0 to 9. Subjects were not told that only four levels of image quality were being presented. They were told that the four levels that were observed in step one bracketed the range of levels presented in this phase of the experiments.

3) Finally, step one above is repeated using the same stimulus sequences. Again, subjects were not informed that pictures in step 1 and 3 are in the same locations.

The same 3 steps above are performed for all seven sequences and are performed twice. It takes about an hour to finish the experiment.

3.2.2 Psychophysical experiment results

We collected data from 11 college students. Subjects were naïve as to the goals and experimental manipulations of the study.

The most often received comment from the test subjects is the difficulty in rating the quality in Step 2. With highly compressed sequences, different patterns of noise appear in different parts of the image and are varying over time. Therefore, when Step 2 was done in preliminary studies without Step 1, subjects tended to shift their rating criteria during the experiment and generate “inconsistent” results, such as a wide distribution of ratings.

We also found that in Step 1 presenting four stimuli simultaneously helped in reducing the inconsistency. The longer viewing time gave subjects an opportunity to study the stimuli and establish stable criteria.

We summarize the data from Step 1 and 3 in tables shown Tables 1 through 7. The numbers in the table represent the frequencies of stimuli being rated as the best, 2nd best, 3rd best and worst. The four stimulus conditions are denoted as R_xD_x, where x can be H or L. $R_{H(L)}$ means the stimulus was encoded at the higher (lower) bit rate. $D_{H(L)}$ means the stimulus has a delay offset of 12 (0) frames.

As we expected, higher bit rates delivered better quality. This is observed through the fact that the first two columns are most often rated as the best or the 2nd best among four stimuli.

What is more interesting, however, is that within the same compression rate, the sequence with the large delay offset is favored over the sequence rendered delay jitter free. In the first two columns, we compared the frequencies of the long delay column (the 2nd) being rated as the best of four versus the frequencies of the no delay column (the 1st) being the best. In most cases, the former is greater than the latter. We then compared the 3rd and 4th columns, which represent sequences compressed at a lower quality. They cannot compete for quality with the high bit rate sequences and thus are often rated as either the 3rd best of the worst of all four stimuli. Nevertheless, by looking at the frequencies of being the 3rd best, we reached the same conclusion that for lower bit rate sequences, sequences with long delay also look better.

Our analysis on the rating scores collected in Step 2 also showed a similar result. The correlation between the quality rating and the compressed bit rate is always positive. The correlation between the rating and the delay is near zero or positive, depending on specific video sequence. Details of the analysis will be provided in future publications.

We are able to identify at least one condition in which delayed video looks better. For its qualitative illustration, please see Figure 3. This condition happens when an original, uncoded block is varying mildly in time. Under high compression, the compressed block carries quantization noise. What we observed is that the jitter free rendering allows the received video closely follow the variation of the original block. The noise accompanying with the variation thus becomes time varying, as shown in the second row of Figure 3. For delayed video, however, the encoder sends the second to the fourth block to the high delay flow, which will arrive at the receiver after 400 msec. In the meaning time, the decoder simply keeps showing the first block received as shown in the third row of Figure 3. The noise seen by our subjects is thus static. Our experiment results indicate the static noise is preferred.

4 SUMMARY

In this paper we evaluated the effects of temporal jitter on video quality using psychophysical methods. As we argued, relaxing the delay jitter (delay) requirements of video traffic lessens the demand on bandwidth usage of modern packet networks. If the amount of delay-critical data can be effectively reduced without significantly degrading the quality of received video, the network connection cost can be significantly reduced.

As the experiment results indicate, introducing temporal jitter in video rendering without degrading video quality is achievable with the specific DCVC algorithm design. For highly compressed sequences, video quality actually improves with delay. A possible cause is that long delays postpone the rendering of some mosquito noise introduced in lossy compression. With further investigation, this result may lead to the improvement of compression algorithm design.

The goal of DCVC, however, is not to improve compression algorithms but rather it focuses on how to segment video data into multiple delay flows with minimal impact on video quality. An aggressive segmentation should eventually lead to quality degradation. New guidelines and HVS models need to be developed to incorporate the understanding of the temporal jitter effects on video. These findings will then lead to a better design of DCVC.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Murat Kunt and Dr. Christian J. van den Branden Lambrecht of the Signal Processing Lab at Swiss Federal Institute of Technology in Lausanne, Switzerland, for providing the Moving Pictures Quality Metric (MPQM) system. This research was supported by the University of California MICRO program, grants from Harris, Rockwell, LG Electronics, National, Plessey, Philips, and an IBM Fellowship. The research was also supported by grant from the Air Force Office of Scientific Research (F49620-95) to the Neurometrics Institute.

REFERENCES

[1] J. Reason, L. C. Yun, A. Lao, D. G. Messerschmitt, “Asynchronous video: coordinated video coding and transport for heterogeneous networks with wireless access,” *Mobile Wireless Information Systems*, Kluwer Academic Press, 1995.

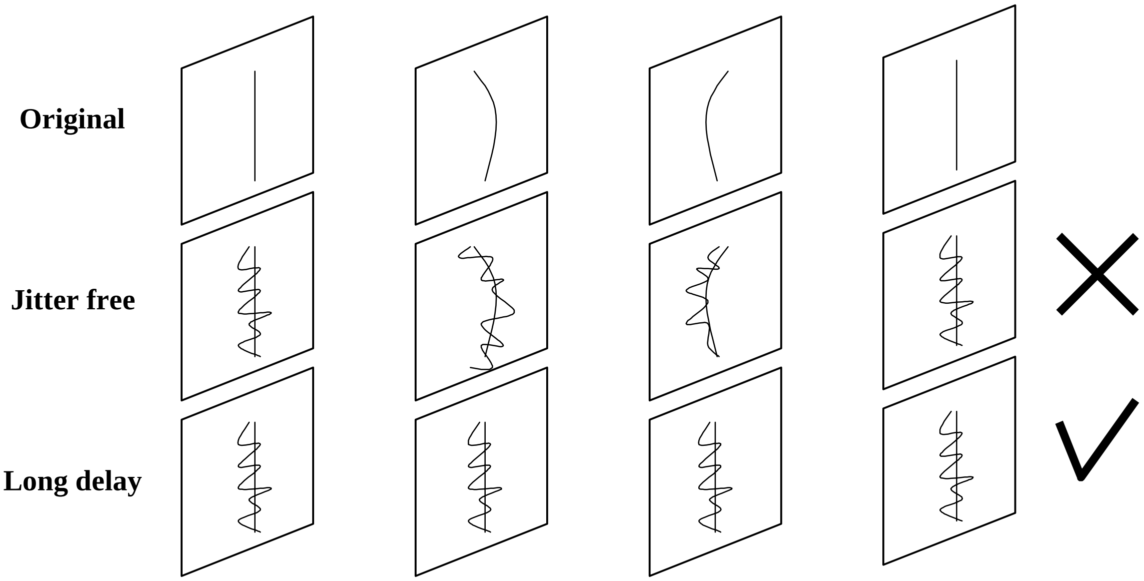

Figure 3 A qualitative illustration of the condition when delayed video looks better.

[2] Y. C. Chang and D. G. Messerschmitt, “Delay cognizant video coding,” *Proceedings of Intl. Conf. on Networking and Multimedia*, Kaohsiung, Taiwan, pp. 110-117, 1996.

[3] Y. C. Chang and D. G. Messerschmitt, “Segmentation and compression of video for delay-flow multimedia networks,” submitted to *1998 IEEE Intl. Conf. On Acoustics, Speech and Signal Processing.*

[4] Y. C. Chang and D. G. Messerschmitt, “Improving network video quality with delay cognizant video coding,” submitted to *1998 IEEE Intl. Conf. On Image Processing*.

[5] R. J. Clarke, “Digital compression of still images and video,” *Academic Press*, 1995.

[6] J. L. Mitchell, W. B. Pennebaker, C. E. Fogg and D. J. LeGall, *MPEG video compression standard*, Chapman & Hall, 1997.

[7] “Video coding for low-bit rate communications: draft recommendation ITU-T H.263,” International Telecommunications Union - Telecommunication Standardization Sector, May 1996.

[8] C. Lambrecht and O. Verscheure, “Perceptual quality measure using a spatio-temporal model of the human visual system,” Proceedings of the SPIE vol. 2668, San Jose, CA, pp.450-61, 1996.

[9] C. Lambrecht, “A working spatio-temporal model of the human visual system for image restoration and quality assessment applications,” *Proceedings of IEEE Int. Conf. On Acoustics, Speech, and Signal Processing,* Atlanta, GA, pp. 2291-4, 1996.

[10] S. A. Klein and D. M. Levi, “Hyperacuity threshold of one second: theoretical prediction and empirical validation,” *J. Opt. Soc. A.*, vol. 1, pp.1170-90, 1985.

Table 1 Ranking histogram of the Carphone sequence

Carphone	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	12	30	0	2
2nd Best	23	11	2	8
3rd Best	9	3	8	24
Worst	0	0	34	10

Table 2 Ranking histogram of the Claire sequence

Claire	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	23	18	1	2
2nd Best	17	20	4	3
3rd Best	3	4	17	20
Worst	1	2	22	19

Table 3 Ranking histogram of the Foreman sequence

Foreman	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	17	26	0	1
2nd Best	24	17	2	1
3rd Best	3	1	18	22
Worst	0	0	24	20

Table 4 Ranking histogram of the Ms. America sequence

Miss Am.	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	19	24	0	1
2nd Best	14	16	5	9
3rd Best	7	4	18	15
Worst	4	0	21	19

Table 5 Ranking histogram of the Mother sequence

Mother	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	14	28	0	2
2nd Best	23	12	3	6
3rd Best	7	3	17	17
Worst	0	1	24	19

Table 6 Ranking histogram of the Salesman sequence

Salesman	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	20	19	0	5
2nd Best	16	15	5	8
3rd Best	6	1	13	15
Worst	2	0	26	16

Table 7 Ranking histogram of the Suzie sequence

Suzie	$R_H D_L$	$R_H D_H$	$R_L D_L$	$R_L D_H$
Best	20	19	3	2
2nd Best	17	15	4	8
3rd Best	4	7	15	18
Worst	3	3	22	16