https://doi.org/10.2352/ISSN.2169-2629.2017.32.27


# Temporal Drift Correction of Residues for Perceptually Based Video Compression

***Mark Q. Shaw †, Jan P. Allebach ‡ and Edward J. Delp ‡***
***†HP Inc, Boise, Idaho 83714, USA***
***‡School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA***

## Abstract

*In this paper we investigate a method for correcting a encoding artifact due to the periodic nature of video coding. This paper builds upon earlier work [1–3] by incorporating temporal drift correction feedback in the encoder to prevent visible artifacts seen when an I-frame reset occurs. The method has been implemented in the H.264 JM 18.0 reference encoder, and has been shown to significantly improve the perceived quality of the video quality when compared to the encoded video without temporal drift correction.*

## Introduction

Video compression that takes advantage of predictive coding algorithms are sensitive to error propagation due to the fact that the predictive accuracy of future frames are partially dependent on the perceptibility of the change from the frames previously encoded. There has been extensive work within the video compression field to design new techniques to limit the severity of error propagation [4, 5]. Reibman and Bottou [5] define drift as "the propagation of errors due to partial reception of less important enhancement layer information" and propose and alternative structure for scalability that uses information previously generated by the encoder to prevent drift. Arnold, *et al* [6] proposed a two-loop scalable encoding architecture with improved coding performance designed to provide a scalable efficient drift free encoder. Taubman and Zakhor [7] demonstrated a multirate spatially scalable video coders that use 3D sub-band coding, Domanski, *et al* [8] developed a spatio-temporally scalable coder that produces two bitstreams. One with a base layer with reduced spatial and temporal resolution, and the other layer used to transmit significant differences. Prades-Nebot, *et al* [9] studied rate control strategies for fully fine-grained scalable video coders, proposing a rate control algorithm based on the rate distortion characteristics of the encoded bitstream to prevent large jumps in quality. Furthermore, a variety of works [10–12] have focused on the temporal drift control applied to a variety of coding applications. The term error propagation is typically reserved for the case in which the encoder and decoder are out of sync and decoding errors result due to data loss in the transmission of the signal. Other work by Gong, *et al* [13–15] describe another phenomenon called "temporal pulsing", this is an issue in the encoder and is a result of a sudden change that happens when the end of a Group of Pictures (GOP) occurs. Due to the periodic nature of the GOP, the artifact appears as a temporal pulsing in the decoded sequence.

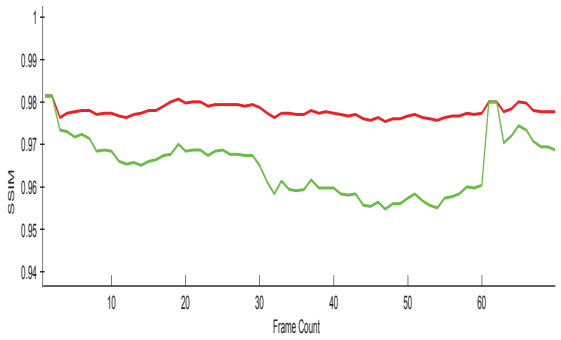



***Figure 1.*** *SSIM video comparison of the City video sequence. The red line represents the frame by frame structural similarity between the reference video stream and the decoded video stream created with the JM 18 reference encoder baseline. The green line represents the frame by frame structural similarity between the reference video stream and the decoded video stream created using the work described in Shaw, et al [3]. The I frame reset can be clearly seen at frame 60, whereby the green line jumps back up to meet the red line.*

## The basics of drift

As was described in the work of Shaw, *et al* [3], the video coding process uses a set frame sequence (GOP) recursively in the processing of the video stream. Each GOP is bounded by an I-frame, which serves as a reset to encoding errors that may have accumulated during the processing of the previous GOP. Although the idea of resetting the errors to zero at each GOP boundary is good in principle, it can result in visual artifacts at the GOP boundaries if the total error accumulation is significant. In such an event, the I-frame reset will likely result in a visible step. This can be seen in Fig. 1, where the I-frame reset occurs at frame 60. We can approximate the error in the video sequence using a simple model

$$R = R_b + R_e + R_d \quad (1)$$

where $R$ is the original sequence being encoded, $R_b$ is the baseline encoding, $R_e$ is the reconstruction error of the video sequence when compared to the original sequence and $R_d$ is the error due to temporal drift.

For the purpose of this research we are calling this accumulated error ($R_d$) "drift". Whereby the accumulation of small errors, each of which are imperceptible, upon an I frame reset result in a visually noticeable change in the decoded video stream. The

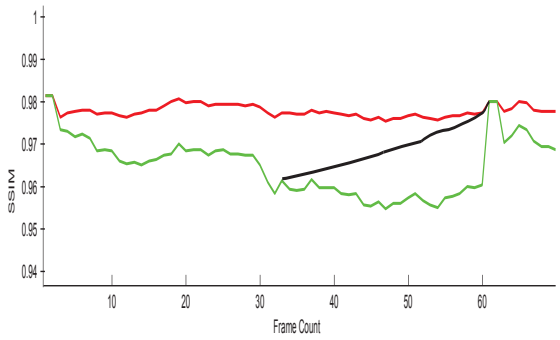



***Figure 2.*** *Idealized representation of how drift correction would work. The black line represents the drift correction's impact on the SSIM measure*

sudden change from the frame with accumulated error to the reference frame can result in a visible "jump" in the sequence during playback. In the literature this has also been called "temporal pulsing" due to the periodic temporal effect that it has on the decoded video sequence [13–15].

## Residue Weighting Correction

To correct for this, the accumulation of errors must be constrained such that the perception of the difference is minimized at a GOP boundary. Ideally, the residuals of the dynamic compression method would approach that of the baseline compression method. An example of this is shown in Fig. 2.

If one is to think of the video encoding workflow discussed in [3], the correction factor can be implemented as a constraint on the modified residues. As the encoder approaches the GOP boundary, the amount of residue loss becomes more and more constrained. Such that at the GOP boundary, the dynamic compression method is not allowed to modify the residues at all.

A very important attribute to the drift correction is knowing when to start applying the correction. It is clear from Fig. 2 that we do not want to constrain the residues for the first half of the GOP. Starting the drift correction too early will result in a loss of the compression gains, and starting the drift correction too late will not provide the effect that we desire.

Therefore, a drift compensation factor ($\beta$) is calculated during the encoding of each frame that achieves the goal shown in Fig. 3 whereby the compensation factor is zero for the first half of the GOP, and then increases linearly to a value of 1 as the end of the GOP is approached. The drift correction factor is calculated using

$$\beta_n = 2\left(\frac{\text{current frame position} - \text{last I frame position}}{\text{next I frame position} - \text{last I frame position}}\right) - 1 \tag{2}$$

where $n$ represents the current frame position, and the *next* and *last* frames are the respective frame numbers for the current GOP. Once the drift correction factor has been computed for the frame under inspection, the weighted residuals are then modified to include the drift correction. In order to accommodate this change, the computation of the modified residues is shown in Fig. 4.

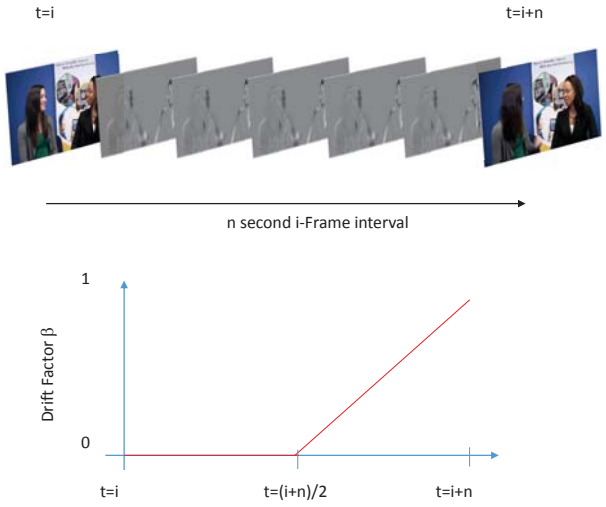



***Figure 3.*** *Illustration of how the drift compensation factor $\beta$ varies between the two I frames within a GOP*

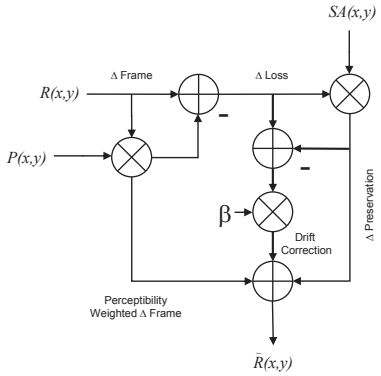



***Figure 4.*** *Modified residue preprocessing workflow to incorporate drift correction. The correction factor $\beta$ is computed on a frame by frame basis to amount of attenuation required for that frame.*

$$\begin{aligned}
\Delta_{\text{loss}}(x,y) &= R(x,y)\Big[1 - P(x,y)\Big], \\
\Delta_{\text{weighted}}(x,y) &= R(x,y)P(x,y), \\
\Delta_{\text{pres}}(x,y) &= \Delta_{\text{loss}}(x,y)\Big[1 - SA(x,y)\Big]^{\rho}, \\
\Delta_{\text{drift}}(x,y) &= \beta_n\Big[\Delta_{\text{loss}}(x,y) - \Delta_{\text{pres}}(x,y)\Big], \\
\tilde{R}(x,y) &= \Delta_{\text{drift}}(x,y) + \Delta_{\text{weighted}}(x,y) + \\
&\quad \Delta_{\text{pres}}(x,y).
\end{aligned} \tag{3}$$

This can be seen in Fig. 4. Here $\Delta_{\text{weighted}}(x,y)$, $\Delta_{\text{loss}}(x,y)$, and $\Delta_{\text{pres}}(x,y)$ represent, respectively, the perceptibility weighted residual, the residual signal loss that would be incurred if one were to just use the perceptibility weighted residual, and the residual that needs to be preserved, given the information from the spatial activity map. The power law operation in

(3) with exponent $\rho$ was chosen to allow control of the relative weight assigned to low and high values of variance in the Lightness channel. The value chosen for this work is $\rho = 2.2$.

The location of the drift correction block is shown in Fig. 5, we have outlined the region to clearly identify the changes described in (3). Fig. 6 shows the result of using the drift correction algorithm to the video quality metric SSIM, when applied to the same video sequence as was shown in Fig. 1. The improvement can be clearly seen starting half way between the two I frames within the GOP.

The magnitude of the differences between the residues with and without drift correction for the *City* and *Kristen and Sara* video sequences are shown in Fig. 7. The frame extracted is one frame before the I-frame reset, to show the maximum impact of the drift correction. It can be seen in both sequences that the differences are significant.

## Quantitative and Qualitative Experimental Results

The impact on visual image quality are shown in Table 1. One can see that the image quality metrics improve slightly, but as is expected, the conditional drift correction has an impact on compression gains for each of the video sequences. Table 2 shows the compression gain for a fixed dynamic tone map function both with and without drift correction.

In order to assess the visual quality of the video sequences when using the proposed method, we performed a subjective quality evaluation. We used the Double-Stimulus Impairment Scale

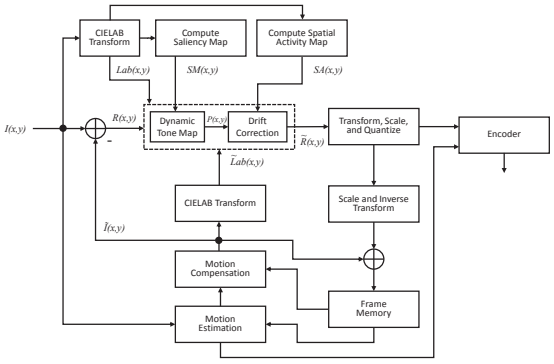



***Figure 5.*** *Modified algorithm workflow diagram showing the addition of drift correction.*

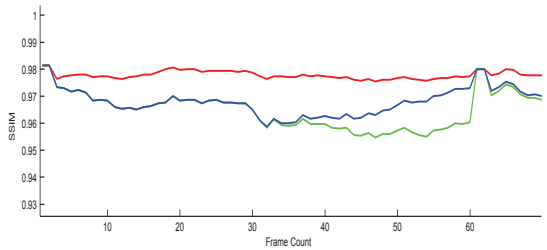



***Figure 6.*** *SSIM result of drift correction algorithm using the same video sequence. The blue line represents the drift corrected video sequence, where the SSIM metric can be clearly seen to approach that of the baseline (red).*

(a) Baseline reference frame for *City*

(b) Baseline reference frame for *Kristen and Sara*

(c) Encoding error magnitude without drift correction

(d) Encoding error magnitude without drift correction

(e) Encoding error magnitude with drift correction

(f) Encoding error magnitude with drift correction

***Figure 7.*** *Comparison of the encoding errors both with (e and f) and without (c and d) conditional drift correction. The original frames (a and b) are shown for reference. The frames under inspection are one frame before the I-frame reset. It can be seen that the magnitude of the errors has been significantly reduced.*

**Impact on video quality metric VSSIM when using the drift correction with the dynamic tone map function**

| Sequence | H.264 JM 18 Baseline | Without Drift Correction | With Drift Correction |
|---|---|---|---|
| City | 0.9389 | 0.9015 | 0.9084 |
| Kristen And Sara | 0.9472 | 0.9393 | 0.9405 |

**Impact on compression gains of HD sequences when using the drift correction with the dynamic tone map function**

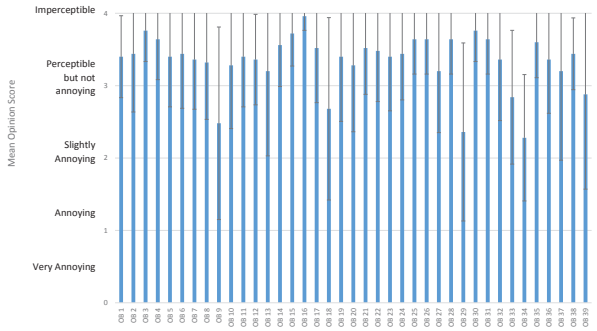
| Sequence | Dynamic Tone Map without Drift Correction | Dynamic Tone Map with Drift Correction |
|---|---|---|
| Basketball Drive | 29.33 | 24.92 |
| Park Scene | 43.71 | 34.37 |
| Cactus | 41.61 | 34.68 |
| Tennis | 23.28 | 20.09 |
| BQ Terrace | 103.30 | 84.04 |
| Kimono1 | 44.50 | 35.57 |
| City | 101.30 | 73.59 |
| Johnny | 39.87 | 31.34 |
| Mobile Calendar | 95.53 | 73.04 |
| Four People | 34.15 | 26.59 |
| Kristen And Sara | 39.64 | 31.35 |
| Average | 54.20 | 42.69 |

(DSIS) method described in Rec. ITU-R BT.500 [16], in which subjects view the two stimuli displayed sequentially with a delay between them. A total of 39 subjects were asked to evaluate the

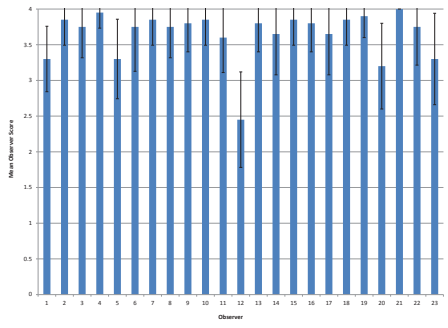
subjective quality of the modified sequences relative to the reference encoding using the following five ratings: Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying, and Very annoying. The subjective quality evaluation was performed in a room with dim back-light to simulate a home viewing environment.

The experiment was performed in two separate locations, one using a 21 inch HP Compaq LCD LA2206x display with a resolution of $1920 \times 1080$ pixels, and the other with a 24 inch HP DreamColor LP2480zx display with a resolution of $1920 \times 1200$ pixels, at a 60 Hz refresh rate. The PLUGE reference target was used to verify that the display contrast and brightness were adequate to render the full tonal range, while maintaining a gamma of 2.2 and a $D_{65}$ white point [16]. One area in which the experiment deviated from the standard was that of viewing distance. Observers were seated 60cm from the display, rather than the ITU recommendation of almost 2 meters. It was felt that a viewing distance of 2 meters was not a natural viewing distance.

The decoded reference sequence was obtained by running the JM 18.0 H.264/AVC reference encoder with the baseline configuration described in [3]. Each of the decoded test sequences was obtained by running the encoder using one of the modified tone maps. The decoded sequences were displayed sequentially with the user interface. Each of the sequences was 10 seconds in length. For each sequence, the reference was compared in two ways - firstly, against four alternatives including the reference itself and secondly, comparing the continuously varying tone map




(a) Current Psychometric




(b) Original Psychometric

***Figure 8.*** *Comparison of the observer variance of the current experiment compared to the first psychometric experiment*

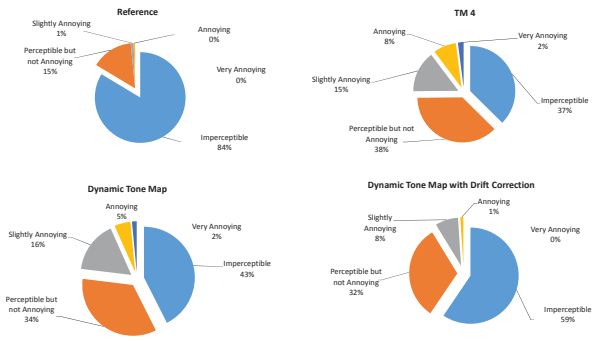



***Figure 9.*** *Observer assessments as a function of processing method, averaged over all observers and all HD video sequences averaged over the four processing methods and the five video sequences*

with drift correction against the continuously varying tonemap without drift correction. A total of five video sequences were used, resulting in a total of 25 observations per subject. As instructed in the ITU-R recommendations, the sequential pair ordering and selection of the sequence pairs was randomized to minimize observer bias and observers were trained prior to the experiment to reduce observer variance.

Fig. 8 shows the observer variance across all of the 25 sequences that each observer viewed. In this experiment we were specifically interested to understand if training the observers prior to the experiment would reduce the observer variance seen in the previous two experiments. It can be seen in Fig. 8 that the observer variances are similar, but the general trend is that the variances are larger than in the previous experiment. One of the key differences between the experiments that may be influencing the results is that the current experiment was performed using high definition sequences, whereas the original experiment used standard definition sequences.

Fig. 9 shows the impact of the processing method. As has been seen in the previous experiments, 16% of observers stated that they could see a difference even when comparing the reference to itself. One can also see that observers rated the visual quality of the continuously variable dynamic tone map with drift correction better than Tone Map 4 (TM 4) and the the dynamic tone map without drift correction.

When looking at the direct comparison between the continuously variable dynamic tone map with and without drift correction shown in Fig. 10, one can see that 50% of observers reported that they could not see a difference. The other 50% were able to see a perceptible degradation when compared to the sequence without drift control. During discussion with some of the observers after completing their experiments, some of the observers noted that for the *Tennis* and *Cactus* sequences they found it very hard to observe any differences at all. Some commented that motion in the *Tennis* sequence masked the subjects ability to see the differences. From these results we can see that the drift control does have a significant impact on the visual quality of the perceptibility of the compression artifacts, and would recommend that that drift control is incorporated into any spatial domain residue preprocessing compression scheme.

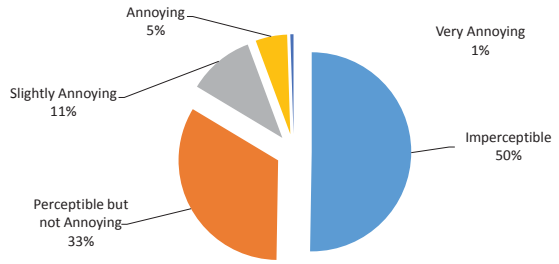



***Figure 10.*** *Observer assessment comparing the continuously varying dynamic tone map with drift correction against the dynamic tone map without drift correction. Averaged over all observers and all HD video sequences averaged over the four processing methods and the five video sequences*

## Conclusions

This work builds upon our previous works [1–3] by addressing the temporal pulsing artifact that was identified due to the periodic I-frame reset in the encoded sequence. The correction algorithm is able to achieve significantly improve the perceptual response of observers when compared to the sequence without correction.

## References

[1] M. Q. Shaw, A. Parra, and J. P. Allebach, "Improved video compression using perceptual modeling," in *IS&T/SID CIC 20 : Twentieth Color and Imaging Conference*, 2012, pp. 9–14.

[2] M. Q. Shaw, J. P. Allebach, and E. J. Delp, "Color difference weighted adaptive residual preprocessing using perceptual modeling for video compression," in *Perception Inspired Multimedia Signal Processing Techniques, The 2nd IEEE Global Conference on Signal and Information Processing*, 2014.

[3] M. Shaw, J. P. Allebach, and E. J Delp, "Color difference weighted adaptive residual preprocessing using perceptual modeling for video compression," *European Association for Signal Processing (EURASIP) journal Signal Processing: Image Communication*.

[4] Ming-Ting Sun and A.R. Reibman, *Compressed Video over Networks*, Marcel Dekker, Inc, New York, NY 10016, 2001.

[5] A.R. Reibman and L. Bottou, "Managing drift in dct-based scalable video coding," in *Data Compression Conference, 2001. Proceedings. DCC 2001.*, 2001, pp. 351–360.

[6] J.F. Arnold, M.R. Fracter, and Yaqiang Wang, "Efficient drift-free signal-to-noise ratio scalability," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 70–82, Feb 2000.

[7] D. Taubman and A. Zakhor, "Multirate 3-d subband coding of video," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 572 – 88, Sept. 1994.

[8] M. Domanski, A. Luczak, and S. Mackowiak, "Spatio-temporal scalability for mpeg video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 7, pp. 1088–1093, Oct 2000.

[9] Josep Prades-Nebot, Gregory W. Cook, and Edward J. Delp III, "Rate control for fully fine-grained scalable video coders," 2002.

[10] E. Magli, M. Grangetto, and G. Olmo, "Conditional access to h.264/avc video with drift control," Piscataway, NJ, USA, 2006//, pp. 4 pp. –.

[11] Wai-Tian Tan and Andrew Patti, "Improving error resilience of scalable h.264 (svc) via drift control," Dallas, TX, United states, 2010, pp. 2302 – 2305, Bit cost;Drift control;Error concealment;Error resilience;Low-complexity;Mitigate errors;Reference software;Scalable video;Spatial resolution;.

[12] P. Correia, P. Assuncao, and V. Silva, "Multiple description video transcoding with temporal drift control," Piscataway, NJ, USA, 2010//, pp. 558 – 61.

[13] Yanchao Gong, Shuai Wan, Kaifang Yang, Fuzheng Yang, and Li Cui, "An efficient algorithm to eliminate temporal pumping artifact in video coding with hierarchical prediction structure," *Journal of Visual Communication and Image Representation*, vol. 25, no. 7, pp. 1528 – 1542, 2014.

[14] Shuai Wan, Yanchao Gong, and Fuzheng Yang, "Perception of temporal pumping artifact in video coding with the hierarchical prediction structure," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 503–508.

[15] Yanchao Gong, Shuai Wan, Kaifang Yang, Bo Li, and Hong Ren Wu, "Perception-based quantitative definition of temporal pumping artifact," in *Digital Signal Processing (DSP), 2014 19th International Conference on*, Aug 2014, pp. 870–875.

[16] "Methodology for the subjective assessment of the quality of television pictures," *Tech. Rep., International Telecommunication Union/ITU, Radio Communication Sector*, 2002.

## Acknowledgements

The authors wish to thank X. Wang and R. Want for their assistance in conducting the psychophysical experiments for this research project.

## Author Biography

*Mark received his M.S. degree in Color Science from the Munsell Color Science Laboratory, RIT, and a PhD in Electrical and Computer Engineering from the department of Signal and Image Processing, at Purdue University under the guidance of Prof. Jan Allebach. Mark is currently working in the LES HW division of HP Inc. in Boise, Idaho as a Senior Color and Imaging Master Architect. Mark has over 15 years experience in the Color and Imaging Industry.*