VIDEO CODING USING GLOBAL MOTION TEMPORAL FILTERING

Alexander Glantz, Andreas Krutz, Martin Haller, and Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Berlin, Germany

ABSTRACT

Recent deblocking techniques are based on spatial filtering. We present a new deblocking technique based on temporal filtering of spatially aligned frames. This approach is used in an H.264/AVC coding environment. The algorithm estimates the ideal amount of frames used for temporal filtering at the encoder side. In that way it is assured that the receiver is presented with the best possible visual quality in terms of structural similarity. Theoretical consideration of the problem proves the concept of the new approach. Experimental evaluation shows that the new temporal deblocking filter significantly improves visual quality and reduces bit rate compared to common H.264/AVC deblocking by up to 18%.

***Index Terms*—** H.264/AVC, video coding, deblocking, temporal filtering, quality assessment

1. INTRODUCTION

Research on video compression techniques has first emerged in the 1960s. Since then, algorithms have evolved leading to the latest state-of-the-art codec H.264/AVC [1] which aims at compressing high-quality video contents at low bit rates. Nevertheless, the importance of ongoing research in that area has been outlined before in [2].

One of the still existing problems in video coding today are distortions, i.e. blocking artifacts that strongly affect the perceived video quality at the receiver. There are mainly two reasons for these artifacts. One of them is the quantization of transform coefficients. Depending on the coarseness of quantization this can cause visually disturbing edges between block boundaries. The second source is the motion compensated prediction. Here, blocks are predicted from temporally neighboring frames that already have been locally decoded and therefore contain discontinuities at block boundaries. These are often copied into the interior of the prediction signal.

Research on deblocking filters has been vast. The H.264/AVC standard itself defines a deblocking filter that is based on the work by List et al. [3]. This algorithm first tries to distinguish between different kinds of discontinuities using boundary analysis. Here, it is assumed that depending on the kind of neighboring blocks, i.e. intra or inter coded, boundaries are more or less severe. The second step is spatial filtering of horizontal and vertical edges. Although subjective quality and the prediction signal could be improved significantly, blocking artifacts are still visible in the decoded video at low bit rates. Nearly all other techniques focus on spatial deblocking. An example can be found in [4].

In this paper we present a coding approach that uses a new deblocking algorithm based on temporal filtering of spatially aligned frames. For a given reference frame, the algorithm takes into account a set of distorted neighboring frames, i.e. frames containing blocking artifacts. The global motion between the reference frame and its neighbors is estimated using a pixel-based gradient descent approach and compensated. This produces a successively growing image stack of spatially aligned distorted frames that are blended together in every step to form a deblocked representation of the reference frame. We show that using this technique we can on one hand reduce blocking artifacts significantly and on the other hand decrease the bit rate needed for transmission. The temporal deblocking is performed both at the encoder and decoder. At the encoder side this is merely done to assess the best possible visual quality by estimating the ideal amount of neighboring frames used for blending. This amount is then transmitted to the receiver so that the temporal deblocking filter at the decoder produces the best quality for the viewer.

It is well known that the peak signal-to-noise ratio (PSNR) used so far for measuring image or video quality is not well matched to perceived visual quality. On the other hand, subjective tests are often not feasible because they are very time-consuming. In this work we use the structural similarity index (SSIM) as visual quality assessment metric. SSIM is based on the work by Wang et al. [5] and the results are very convincing. The main difference of SSIM compared to PSNR is that it does not estimate perceived errors to quantify image degradations but considers them as perceived changes in structural information variation.

This paper is organized as follows. Section 2 discusses the approach presented in this work theoretically. Section 3 describes the proposed video codec. In Section 4 we present experimental results and Section 5 summarizes the paper.

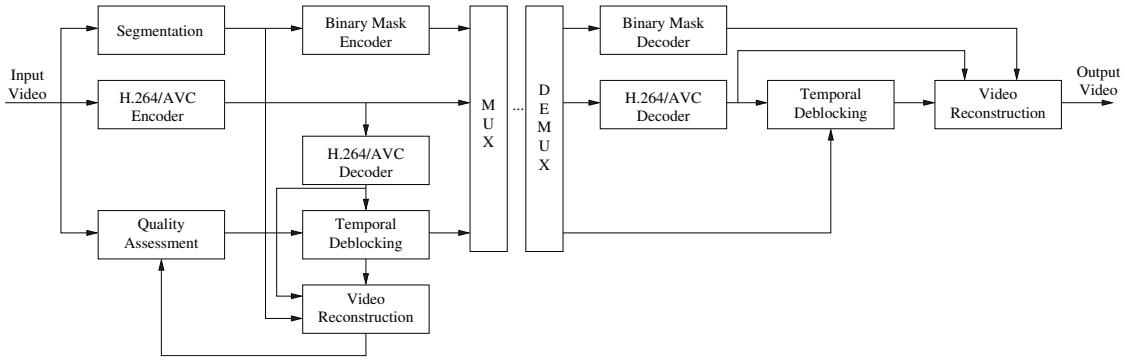


Fig. 1. Global motion temporal filtering within a video coding environment

2. THEORETICAL CONSIDERATION

2.1. Temporal superposition for noise reduction

We assume, that a quantized signal y_k is the sum of an original signal x and quantization noise e_k:

$$y_k = x + e_k, \tag{1}$$

y_k, x, and e_k being sample points of stationary Gaussian distributed zero mean random processes, with $E[xe_k] = 0$, i.e. x and e_k being statistically independent.

Assume, it is is possible to observe N representations of y_k, each being quantized with a different sample point e_k. Averaging the quantized pixel values in temporal direction results in an enhanced representation of y:

$$y_{avg} = \frac{1}{N}\sum_{k=1}^{N} y_k = x + \underbrace{\frac{1}{N}\sum_{k=1}^{N} e_k}_{=e_f}, \tag{2}$$

with e_f being the mean, i.e. the filtered, quantization noise signal. The resulting variance of e_f can be calculated using the variance $\sigma^2_{e_{nf}}$ of the non-filtered signal e_k:

$$\sigma^2_{e_f} = E[e_f^2] = \frac{1}{N^2}\sum_{k=1}^{N}\sigma^2_{e_{nf}} = \frac{\sigma^2_{e_{nf}}}{N}, \tag{3}$$

It can be seen, that the variance of the distortion affecting the original signal has been reduced by factor N.

2.2. Rate-distortion performance for noise reduction

Rate-distortion theory states, that, given a stationary Gaussian distributed source X, the distortion-rate function

$$D(R) = 2^{-2R}\sigma_x^2 \tag{4}$$

describes the minimal possible distortion D, i.e. the mean squared error (MSE) between original and reconstructed signal, for a given rate R.

To see whether it is possible to encode a filtered signal with reduced bits per sample compared to encoding the non-filtered representation, we request same quality after reconstruction, measured using SSIM. Given, that $\sigma_y^2 = \sigma_x^2 + \sigma_e^2$ and $\mathrm{cov}(x, y) = \sigma_x^2$, $\mathrm{SSIM}(x, y)$ reduces to

$$\mathrm{SSIM}_{nf}(x, y_k) = \frac{2\,\mathrm{cov}(x, y_k) + C_2}{\sigma_x^2 + \sigma_{y_k}^2 + C_2} = \frac{2\sigma_x^2 + C_2}{2\sigma_x^2 + \sigma_{e_{nf}}^2 + C_2}, \tag{5}$$

$$\mathrm{SSIM}_{f}(x, y_{avg}) = \frac{2\,\mathrm{cov}(x, y_{avg}) + C_2}{\sigma_x^2 + \sigma_{y_{avg}}^2 + C_2} = \frac{2\sigma_x^2 + C_2}{2\sigma_x^2 + \frac{\sigma_{e_{nf}}^2}{N} + C_2}, \tag{6}$$

with C_2 being a constant factor. Since we requested, that $\mathrm{SSIM}_{nf} = \mathrm{SSIM}_f$, and therefore $D_f(R_f) = D_{nf}(R_{nf})$, it is

$$\begin{aligned} 2^{-2R_{nf}}\sigma^2_{e_{nf}} &= 2^{-2R_f}\sigma^2_{e_f} \\ 2^{-2R_{nf}}\sigma^2_{e_{nf}} &= 2^{-2R_f}\frac{\sigma^2_{e_{nf}}}{N} \\ 2^{-2R_{nf}} &= \frac{2^{-2R_f}}{N} \\ R_f &= R_{nf} - \frac{1}{2}\log_2(N) \end{aligned} \tag{7}$$

Assuming same SSIM, a bit rate reduction of $\frac{1}{2}\log_2(N)$ bits per sample is achieved. This means, it is possible to encode a filtered version of signal x with a bit rate reduction that directly depends on the amount N of distorted representations used for the process of filtering. Or in other words, given the same bit rate, the filtering approach results in an improved quality.

A real coding environment imposes a set of constraints, that limits the applicability of this theoretical consideration, e.g. the statistics of video data differing from the theoretical assumption or global motion estimation error affecting the filtering process. Nevertheless, the basic idea of temporal superposition for quantization noise reduction still leads to quality

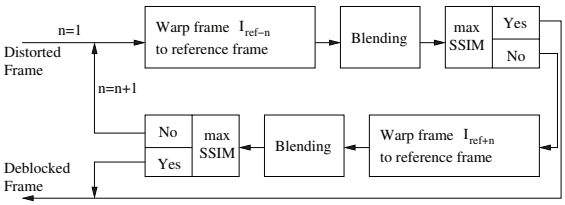


Fig. 2. Temporal deblocking for frame I_{ref}

improvement. This can be seen in the next section, where this technique is used in a coding environment.

3. GLOBAL MOTION TEMPORAL FILTERING

Figure 1 shows a block diagram of the temporal deblocking approach included in an H.264/AVC coding environment. Herein, frames of a given video sequence are normally coded as one would do using H.264/AVC without using its standardized deblocking filter [3].

At the encoder side, frames are decoded and stored in a buffer that is used for the temporal deblocking filter. This filter has been described before in [6]. Figure 2 shows a block diagram of the filter. For a given reference frame I_{ref} the temporal neighborhood, consisting of a set of distorted frames, is taken into account for deblocking. The algorithm successively warps temporally neighboring frames into the coordinate system of the reference frame. To this end, the global motion – e.g. an 8-parameter perspective motion model – has to be estimated which is done using a hierarchical gradient descent approach based on the Gauss-Newton method. Thus, a growing image stack of spatially aligned frames is generated which is blended together to build a single denoised representation I'_{ref} of the reference frame. In every blending step the quality of the current representation I'_{ref} is compared to the original frame $I_{orig,ref}$ using SSIM as mentioned in the Introduction. The representation with the highest SSIM value is taken as the final deblocked frame. The number of frames N used for deblocking is differentially coded using signed mapping Exp-Golomb codes and sent as side information to the receiver. Temporal deblocking at the encoder only takes place to assess best quality at the receiver and therefore to measure the ideal amount of frames to be used.

Foreground segments are defined as segments in the scene that move other than the global motion. These segments vanish successively from the deblocked frames the more temporal neighbors are used for generating them. The segments are coded using H.264/AVC with standardized deblocking filter. The segment mask has to transmitted as side information to the receiver. Automatic segmentation takes place in a preprocessing step. Since it is only necessary to ensure a correct binary mask, we will not further define the way it is generated. In our work we used the algorithm previously published in [7] which is an anisotropic diffusion-based background subtraction technique.

(a) H.264/AVC deblocking filter (38.17 kbit/s, PSNR = 38.00, SSIM = 0.72)

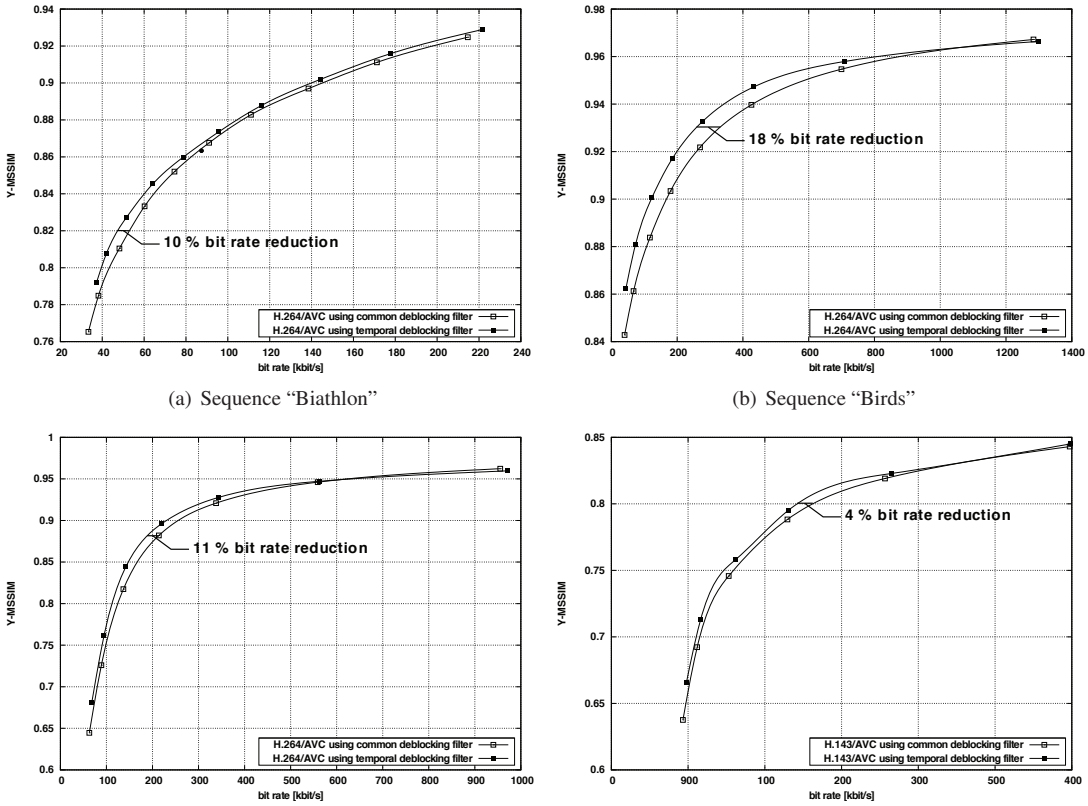
(b) Temporal deblocking filter (37.19 kbit/s, PSNR = 38.01, SSIM = 0.73)

Fig. 3. Subjective comparison of H.264/AVC deblocking with temporal deblocking, detail from sequence "Biathlon", frame 32 (PSNR and SSIM are computed for this window)

At the receiver the common H.264/AVC bitstream, the binary object mask, and the number of frames used in the temporal deblocking filter are decoded. The deblocking filter computes the denoised frames as defined above. Finally, the frames are reconstructed using the binary foreground object mask and presented to the viewer. In our approach global motion parameters are not transmitted. The decoder estimates the parameters based on the noisy decoded images.

4. EXPERIMENTAL EVALUATION

For experimental evaluation we compared the common H.264/AVC deblocking filter to the new temporal deblocking approach. For the H.264/AVC encoder we used hierarchical B-frames, CABAC entropy coding. The in-loop deblocking filter is turned off for our new approach. We used 4 test sequences: "Biathlon" (352×288, 200 frames) taken from a German televison broadcast, "Birds" (720×576, 110 frames) from the BBC documentary "Planet Earth", "Desert" (720×400, 240 frames) from "Planet Earth" as well, and "Race1" (554×336, 100 frames) from an MPEG multiview test sequence. Figure 4 shows compression efficiency in terms of rate-distortion performance. For distortion measurement we used the mean SSIM on the luminance channel (Y-MSSIM) instead of PSNR for the reasons mentioned in the Introduction. We reach bit rate savings of up to 18% ("Birds" at 0.93 Y-MSSIM) mostly in lower bit rates. This is understandable because with higher bit rate the amount of blocking artifacts drops rapidly meaning there should be no difference between the various deblocking approaches. Figure 3 shows a subjective comparison of a representative frame from sequence "Biathlon". Here, one can clearly see the significant deblocking capabilities of the approach presented in this paper. The proposed approach performs excellent deblocking while preserving edges in the images efficiently.



(a) Sequence "Biathlon"

(b) Sequence "Birds"

(c) Sequence "Desert"

(d) Sequence "Race1"

Fig. 4. Comparison of H.264/AVC deblocking with temporal deblocking in terms of rate-distortion performance

5. SUMMARY

In this paper we have presented a new deblocking technique included in a coding environment based on H.264/AVC. For deblocking, temporal filtering of spatially aligned frames is performed instead of spatial deblocking. The approach presented estimates the ideal amount of frames used for temporal filtering so that the receiver is presented with the best possible visual quality in terms of structural similarity. Bit rate savings of up to 18% are observed.

6. REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, July 2003.

[2] T. Sikora, "Trends and Perspectives in Image and Video Coding," *Proceedings of the IEEE*, vol. 93, pp. 6–17, January 2005.

[3] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 614–619, July 2003.

[4] Gwang Hoon Park, Min Woo Park, Sung-Chang Lim, Woo Sung Shim, and Yung-Lyul Lee, "Deblocking filtering for illumination compensation in multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 10, pp. 1457–1461, Oct. 2008.

[5] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, April 2004.

[6] A. Krutz, A. Glantz, M. Frater, and T. Sikora, "Local background sprite generation," in *International Workshop on Local and Non-Local Approximation in Image Processing, LNLA 2008*, Lausanne, Switzerland, Aug. 2008.

[7] A. Krutz, A. Glantz, T. Borgmann, M. Frater, and T. Sikora, "Motion-based object segmentation using local background sprites," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009.