THEORETICAL CONSIDERATION OF GLOBAL MOTION TEMPORAL FILTERING

Andreas Krutz, Alexander Glantz, and Thomas Sikora

Communication Systems Group
Technische Universitaet Berlin
Berlin, Germany

ABSTRACT

A widely used technique to reduce the noise variance of a signal is a temporal overlapping of several noisy versions of it. It will be shown that the same idea can be applied for video sequences. Several versions of the current frame can be aligned using motion compensation that adjacent frames represent a noisy version of the current frame. In a first theoretical calculation of this concept combining the temporal overlapping of several noisy versions of the same signal and a rate-distortion equation, it has been shown that a theoretical bit rate reduction of $\frac{1}{2}\log_2(N)$ is possible. In this work, the concept will be advanced to be closer to practice by adding a model for the motion estimation error. It will be shown that the derived theoretical equation confirms the practice and models the behavior of a video encoding environment using parametric motion compensated temporal filtering very well.

1. INTRODUCTION

Temporal filtering is a widely used method in video processing techniques such as video coding, video enhancement, and superresolution. It is well known that noise reduction using overlapping of a number of noisy versions of the same signal is very powerful. Due to the fact that consecutive frames of a video sequence are highly correlated, it is possible to use this technique to reduce noise in a video sequence. For that, adjacent frames of the current frame to be filtered can be taken into account to be multiple versions of the same frame.

Higher-order motion compensation as shown in e.g. [1], [2] brings improved coding efficiency for certain test sequences. The idea of combining these techniques including a superresolution approach with a common hybrid coding scheme has been outlined in [3]. This promising technique also includes a theoretical model. Here, higher-order motion compensation was applied to improve the prediction efficiency of a hybrid encoder loop. It has also been shown theoretically that the use of the proposed technique is reasonable. Inspired by this work and further theoretical approaches in image processing, e.g. [4], [5], and [6], the motivation of this paper is to show a new parametric motion temporal filtering technique including a fundamental theoretical background. It has been shown in [7] that the use of higher-order motion parameters can lead to a significant improvement depending on the application. Here, a theoretical model is developed to prove the efficiency of parametric motion temporal filtering for deblocking. It is shown that, based on some assumptions, having this temporal filtering approach inside a coding system, the quantization noise (blocking artifacts) can be significantly reduced. For that, the idea of noise reduction using temporal overlapping is connected with a rate-distortion environment shown in [8]. Additionally, the parametric motion estimation error that occurs when the image stack is built is also taken into account.

The paper is organized as follows. Section II decribes the derivation of the theoretical model of parametric motion temporal filtering. Section III shows experimental results where the model is compared with practical results using the coding environment proposed in [8]. It will be shown that the theoretical equation is very close to the real world. The last section summarizes the paper.

2. DERIVATION OF A THEORETICAL MODEL

It is assumed that a number of distorted versions Y of an original image X are available. We consider the kth pixel value $y_k(m,n)$ of the kth version which is the sum of the original pixel $x(m,n)$ and a value from the white noise signal $n_k(m,n)$:

$$y_k(m,n) = x(m,n) + n_k(m,n) \tag{1}$$

We calculate the mean value using each candidate of pixel $y_k(m,n)$:

$$y(m,n) = \frac{1}{N}\sum_{k=1}^{N} y_k(m,n) = x(m,n) + \underbrace{\frac{1}{N}\sum_{k=1}^{N} n_k(m,n)}_{r(m,n)}. \tag{2}$$

White noise is assumed with the variance σ_n^2 and the autocorrelation matrix:

$$R_{nn} = \begin{pmatrix} \sigma_n^2 & 0 & \cdots \\ 0 & \sigma_n^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (3)$$

We now show that the variance of the noise is reduced by the factor N (number of overlapping signals). The mean noise signal is $r(m, n)$. Its variance can be calculated as:

$$\sigma_r^2 = E[R^2(m,n)] = \frac{1}{N^2}\sum_{i=1}^{N}\sigma_n^2 = \frac{\sigma_n^2}{N} \quad (4)$$

Thus, the variance of the noise has been reduced by the factor N. We now turn to our deblocking problem in a video. Assume that it is possible to observe N representations of y, i.e. corresponding quantized pixels from N frames of a video sequence. Averaging these quantized pixels y in temporal direction reduces the error variance by

$$E[e_f^2] = \frac{E[e^2]}{N}, \quad (5)$$

with N being the number of noisy versions x. This assumes that the versions of e in temporal direction are also not correlated. Our goal is to see whether it is possible to code all versions of x along the temporal direction with reduced bits/sample when afterwards temporal filtering is applied. Before showing that, we have to consider the generation of a version of e in temporal direction. That means we have to align a number of frames for filtering one reference image. This is conducted using short-term and long-term motion estimation. We now find a way to model this. It is still assumed that we have the 2-dimensional Gaussian distributed memoryless signal $x = x_n$. We now try to estimate the current sample using the previous one. This is done with an additive operation, that is in an optimal estimation:

$$x_n(x,y) = x_{n-1}(x + t_x, y + t_y), \quad (6)$$

where t is the optimal estimation parameter. In practice, this operation does not match exactly due to the way the estimation is calculated, right motion model, interpolation operation for sub-pixel case, etc. Therefore, an estimation error appears, which can be written for our theoretical case:

$$x_{n-1}(x+t_x+\Delta_x, x+t_y+\Delta_y) = x_n(x+\Delta_x, y+\Delta_y), \quad (7)$$

where Δ_x, Δ_y is the estimation error. Thus, the resulting error signal is:

$$e_n(x,y) = x_n(x,y) - x_n(x+\Delta_x, y+\Delta_y). \quad (8)$$

The term $x_n(x+\Delta_x, y+\Delta_y)$ can be approximated using the first Taylor expansion:

$$x_n(x+\Delta_x, y+\Delta_y) \approx x_n(x) + \nabla x_n^T \cdot \begin{pmatrix} \Delta_x \\ \Delta_y \end{pmatrix}. \quad (9)$$

With this, the error signal is shown in (10).

$$e_n(x,y) = -\frac{\partial x_n(x,y)}{\partial x}\cdot\Delta_x - \frac{\partial x_n(x,y)}{\partial y}\cdot\Delta_y \quad (10)$$

We can now calculate the error variance using the expected value

$$\sigma_{e_n}^2 = E\left[\left(-\frac{\partial x_n(x,y)}{\partial x}\Delta_x - \frac{\partial x_n(x,y)}{\partial y}\Delta_y\right)^2\right] \quad (11)$$

with the assumption that $\frac{\partial x_n(x,y)}{\partial x}$, Δ_x, $\frac{\partial x_n(x,y)}{\partial y}$, and Δ_y are uncorrelated and statistically independent:

$$\begin{aligned} \sigma_{e_n}^2 \;=\; & E[\Delta_x^2]\cdot E\left[\left(\frac{\partial x_n(x,y)}{\partial x}\right)^2\right] + \\ & E[\Delta_y^2]\cdot E\left[\left(\frac{\partial x_n(x,y)}{\partial y}\right)^2\right] \quad (12) \end{aligned}$$

Now, we approximate the derivative of x_n with the first numerical derivative in both directions:

$$\begin{aligned} \frac{\partial x_n(x,y)}{\partial x} &\approx x_n(x,y) - x_n(x-1,y) \\ \frac{\partial x_n(x,y)}{\partial y} &\approx x_n(x,y) - x_n(x,y-1), \quad (13) \end{aligned}$$

using this and the assumption $E[\Delta_x^2] = E[\Delta_y^2] = E[\Delta^2]$ we can plug the approximation in (12) and results in (14), which is the prediction error variance due to the estimation of one signal from a previous version. We assume that the autocorrelation function (ACF) in x- and y-direction can be approximated as a first order autoregressive process ($AR(1)$) with a correlation factor α between zero and one.

$$\begin{aligned} \sigma_{e_n}^2 &= \sigma_\Delta^2\cdot\left\{E\left[\left(\frac{\partial x_n(x,y)}{\partial x}\right)^2\right] + \right. \\ &\qquad \left. E\left[\left(\frac{\partial x_n(x,y)}{\partial y}\right)^2\right]\right\} \\ &= \sigma_\Delta^2\cdot\{\sigma_x^2 - 2\,\underbrace{E[x_n(x,y)x_n(x-1,y)]}_{ACF(AR(1))=\sigma_x^2\cdot\alpha_1^{|1|}} + \\ &\qquad \sigma_x^2 + \sigma_x^2 - 2\,\underbrace{E[x_n(x,y)x_n(x,y-1)]}_{ACF(AR(1))=\sigma_x^2\cdot\alpha_2^{|1|}} + \sigma_x^2\} \\ &= \sigma_\Delta^2\sigma_x^2\cdot(4 - 2(\alpha_1+\alpha_2) \\ &= 2\sigma_\Delta^2\sigma_x^2(2-\alpha_1-\alpha_2) \quad (14) \end{aligned}$$

Knowing this, we can derive a rate-distortion equation for reducing the noise using temporal filtering with the constraint of the estimation error variance. For our Gaussian distributed memoryless signal x_n, the D-R-function is:

$$\sigma^2_{e_{x_q}} = 2^{-2R} \cdot \sigma^2_x. \tag{15}$$

In the aligning process for our temporal filtering, we calculate short-term parameters for the estimation between consecutive signals. Every aligned signal which represent a “new” version of the signal to be filtered is generated by applying long-term motion parameters according to the reference signal. The long-term motion parameters are calculated using accumulative multiplication of the short-term parameters. We assume that the model for the short-term estimation errors between two consecutive frames derived in (14) can serve for every estimation step. If these errors are now accumulated because of building the long-term motion parameters, the overall error caused by the motion estimation process increases. To model the long-term motion compensation error, we basically sum the errors which occur due to short-term motion estimation as derived in (14). It is highly emphasized that this assumption is made to simplify the theoretical modeling at this stage, because calculating the long-term motion parameters and the blending process to build the filtered version of the current image is a very sophisticated process. To design a more accurate model for that process is an issue for further work. However, it will be shown later that this assumption approximates the real behavior of the video codec using global temporal filtering very well.

Thus, we consider two error components of our model for temporal noise reduction. The temporally overlapped quantization error represented by its variance $\sigma^2_{e_q}$ and the prediction error variance due to the motion estimation $\sigma^2_{e_m}$:

$$\begin{aligned} \sigma^2_{e_q} &= 2^{-2R}\frac{\sigma^2_x}{N} \\ \sigma^2_{e_m} &= N \cdot 2\sigma^2_\Delta \sigma^2_x (2-\alpha_1-\alpha_2) \end{aligned} \tag{16}$$

We assume that the final error variance is built by the sum of the two components shown above. Thus, the D-R-function of our model for the temporal noise reduction with (16) is:

$$\sigma^2_{e_{tf}} = 2^{-2R}\frac{\sigma^2_x}{N} + N \cdot 2\sigma^2_\Delta \sigma^2_x (2-\alpha_1-\alpha_2). \tag{17}$$

Now, it is of interest how possible bit rate savings are carried out from this theoretical D-R-function. For that, the distortion values of (15) and (17) are set equal. The bit rate of the general quantization error shall be R_1 and the bit rate using temporal noise reduction shall be R_2. An equation of the bit rate R_2 can now be derived as shown in (18).

$$\begin{aligned} \sigma^2_{e_{x_q}} &\overset{!}{=} \sigma^2_{e_{tf}} \\ 2^{-2R_1}\sigma^2_x &= 2^{-2R_2}\frac{\sigma^2_x}{N} + N \cdot 2\sigma^2_\Delta \sigma^2_x (2-\alpha_1-\alpha_2) \\ 2^{-2R_2}\frac{1}{N} &= 2^{-2R_1} - N \cdot 2\sigma^2_\Delta (2-\alpha_1-\alpha_2) \\ R_2 &= -\frac{1}{2}\Bigg\{ ld\{2^{-2R_1} - \\ &\qquad N2\sigma^2_\Delta(2-\alpha_1-\alpha_2)\} + ld(N)\Bigg\} \end{aligned} \tag{18}$$

For a meaningful interpretation of (18), limits are calculated where (18) is valid considering the real coding and filtering algorithm. First, a lower limit for R_1 is derived. The equation in (18) makes only sense if the term $2^{-2R_1} - N2\sigma^2_\Delta(2-\alpha_1-\alpha_2)$ is greater than zero. This leads to:

$$\begin{aligned} 0 &< 2^{-2R_1} - N2\sigma^2_\Delta(2-\alpha_1-\alpha_2) \\ 2^{-2R_1} &> N2\sigma^2_\Delta(2-\alpha_1-\alpha_2) \\ -2R_1 &> ld(N2\sigma^2_\Delta(2-\alpha_1-\alpha_2)) \\ R_1 &> -\frac{1}{2}ld(N2\sigma^2_\Delta(2-\alpha_1-\alpha_2)) \end{aligned} \tag{19}$$

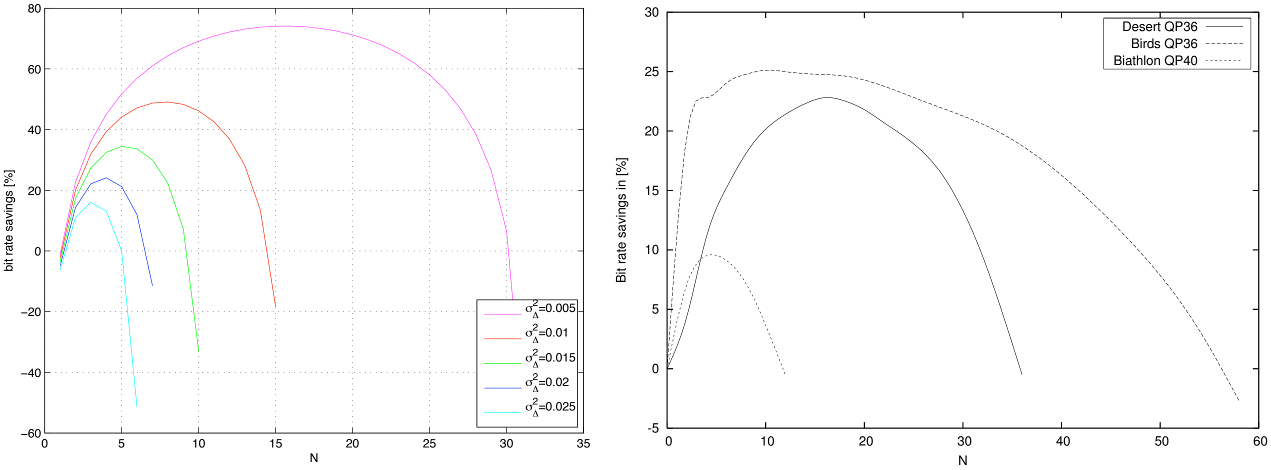
The next consideration is that the number of frames for temporal filtering N is greater or equal to 1. This means that $\text{ld}(N) \geqslant 0$. The upper limit of N can be derived from (18). The upper limit for N and the error variance of the pixel difference due to motion estimation σ^2_Δ, respectively:

$$\begin{aligned} 0 &\leqslant 2^{-2R_1} - N2\sigma^2_\Delta(2-\alpha_1-\alpha_2) \\ 2^{-2R_1} &\geqslant N2\sigma^2_\Delta(2-\alpha_1-\alpha_2) \\ N &\leqslant \frac{2^{-2R_1}}{2\sigma^2_\Delta(2-\alpha_1-\alpha_2)} \end{aligned} \tag{20}$$

$$\sigma^2_\Delta \leqslant \frac{2^{-2R_1}}{2N(2-\alpha_1-\alpha_2)} \tag{21}$$

3. EXPERIMENTAL EVALUATION

We would like to evaluate the theoretical model found in (17) and (18). For that, We used three test sequences: “Biathlon” (352×288, 200 frames) taken from a German televison broadcast, “Birds” (720 × 576, 110 frames) and, “Desert” (720 × 400, 240 frames) from BBC documentary “Planet Earth”. We consider theoretical rate (R) vs. number of frames for filtering (N) function (18). It is demonstrated how well our theoretical function models the behavior of the visual quality assessed video codec presented in [8]. For that, we conducted experiments to draw R-N-curves with the experimental data. Fig. 1 (a) shows the results for (18) with different motion estimation error variances σ_Δ and (b) shows the results for three test sequences at one QP. It can be seen that the experimental curves



(a) Theoretical ΔR-N-curves

(b) Practical ΔR-N-curves

Fig. 1. Comparing theoretical and experimental curves of bit rate savings (ΔR) vs. number of frames (N) with variable motion estimation error variance

approximately follow the theoretical curves. The curves have different shapes depending on the possible bit rate saving that can be achieved. As the theoretical function demonstrates, the bit rate savings increase with a decreasing motion estimation error variance. The motion estimation performs very well with the test sequences “Birds” and “Desert”, but is more difficult with “Biathlon”. Therefore, less bit rate savings are possible with “Biathlon”. Thus, we have shown with our theoretical model that the global motion temporal filter inside a video codec can bring a very good coding performance regarding bit rate reduction and visual quality enhancement.

4. CONCLUSION

We have shown an approach for theoretical modeling of parametric motion temporal filtering. The motivation was to show how this kind of filtering method can perform in a coding environment. Therefore, it was assumed that the quantization noise can be treated as white noise so that temporal overlapping of several noisy versions of the same signal can be applied for noise reduction. This was integrated in a rate-distortion function. The resulting equation has mapped the behavior of the filtering approach in a coding environment very well. It can be concluded that with this theoretical consideration the power of the parametric motion temporal filtering approach has been proved and it motivates further work e.g. implementing this idea in a hybrid encoder loop or work on further video enhancement algorithms.

5. REFERENCES

[1] A. Smolic, T. Sikora, and J.-R. Ohm, “Long-term global motion estimation and its application for sprite coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1227–1242, Dec 1998.

[2] Frederic Dufaux and Janusz Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Transactions on Image Processing*, vol. 9, pp. 497–501, Mar 2000.

[3] A. Smolic, Y. Vatis, Heiko Schwarz, and T. Wiegand, “Improved H.264/AVC coding using long-term global motion compensation,” in *IS&T/SPIE Symposium on Visual Communications and Image Processing (VCIP’04)*, San Jose, CA, USA, Jan. 2004.

[4] A. Jain, “Advances in mathematical models for image processing,” *Proceedings of the IEEE*, vol. 69, no. 5, pp. 502–534, May 1981.

[5] R. Mester, “A system-theoretical view on local motion estimation,” in *Image Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on*, 2002, pp. 201 –205.

[6] M.A. Agostini and M. Antonini, “Theoretical model of the coding error in mcwt video coders,” in *Image Processing, 2006 IEEE International Conference on*, Oct. 2006, pp. 1885 –1888.

[7] G. Dane and T.Q. Nguyen, “The effect of global motion parameter accuracies on the efficiency of video coding,” in *Image Processing, 2004. ICIP ’04. 2004 International Conference on*, Oct. 2004, vol. 5, pp. 3359 – 3362 Vol. 5.

[8] A. Glantz, A. Krutz M. Haller, and T. Sikora, “Video coding using global motion temporal filtering,” in *IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, Nov 2009.