TEMPORAL FILTERING OF WAVELET-COMPRESSED MOTION IMAGERY

Mark A. Robertson

Air Force Research Laboratory/IFEC, Rome, NY 13441-4505

ABSTRACT

Temporal filtering of motion imagery can alleviate the effects of noise and artifacts in the data by incorporating observations of the imagery data from several distinct frames. If the noise that is expected to occur in the data is well-modeled by independent and identically distributed (IID) Gaussian noise, then straightforward algorithms can be designed to filter along motion trajectories in an optimal fashion. This paper addresses the restoration of motion imagery that has been compressed by scalar quantization of the data's two- or three-dimensional discrete wavelet transform coefficients. Noise due to compression in such situations is neither independent nor identically distributed, and thus simple filters designed for the IID case are suboptimal. This paper shows how proper statistical characterization of the compression error can be used in temporal filtering to improve the visual quality of the compressed motion imagery.

1. INTRODUCTION

As lossy imagery compression techniques continue to utilize discrete wavelet transforms (DWT), algorithms that process the data need to become more aware of peculiarities in the compression that may affect processing performance. This paper considers the specific application of motion-compensated temporal filtering, by which noise in motion imagery is alleviated by filtering along motion trajectories. For video frames corrupted by well-behaved IID Gaussian noise, optimal temporal filters are straightforward to design and implement. However, by performing lossy compression on the data one intentionally introduces an error in the reconstructed imagery in exchange for decreased bandwidth. It is well-established that these errors are not independent nor identically distributed, and thus different models for temporal filtering (and processing in general) must be derived. Section 2 derives such a model for the compression error introduced by scalar quantization of 2D or 3D DWT coefficients, as might be performed in Motion JPEG2000 [1] or 3D SPIHT [2]. Section 3 introduces a framework for temporal filtering in compressed video, providing various experimental results for data compressed by 2D and 3D wavelet transforms. Experimental results for both synthetic and real video sequences clearly demonstrate that proper noise modeling as described in this paper can provide significant objective and subjective quality improvements relative to filtering with the simpler but common assumption of IID Gaussian noise. Section 4 concludes the paper.

2. QUANTIZATION NOISE

Without loss of generality, we represent the pixel-domain image data by the length-N vector $\mathbf{z}$, which is formed by stacking the columns from either a two-dimensional image or the images of a three-dimensional imagery volume. The multiresolution DWT being employed is represented by the matrix $\mathbf{H}$, of size $N \times N$, which can be either the two- or three-dimensional DWT, depending on the situation. The wavelet coefficients are given as $\mathbf{y} = \mathbf{Hz}$, which are quantized to $\mathbf{y}_q = \mathbf{Hz} + \mathbf{e_y}$, where $\mathbf{e_y}$ represents the error due to quantization of the wavelet coefficients. Only scalar quantization is considered here. Upon application of the inverse wavelet transform, the reconstructed image becomes $\mathbf{z}_q = \mathbf{z} + \mathbf{e_z}$, where the spatial domain quantization error is $\mathbf{e_z} = \mathbf{H}^{-1}\mathbf{e_y}$, where $\mathbf{H}^{-1}$ is the inverse DWT.

The original image data $\mathbf{z}$ has a covariance matrix of $\mathbf{K_z}$, which results in a covariance matrix for the wavelet coefficients of $\mathbf{K_y} = \mathbf{HK_zH}^t$. For natural images, many wavelet transforms will provide transform coefficients that are approximately uncorrelated. Here it is assumed that the coefficients are approximately uncorrelated, allowing the simplifying approximation that $\mathbf{K_y}$ is diagonal, which leads to the approximation that the covariance matrix of $\mathbf{e_y}$, $\mathbf{K_{e_y}}$, is approximately diagonal, and consists of the wavelet domain quantization error variances for each coefficient. Given $\mathbf{K_{e_y}}$, the covariance of the quantization error in the spatial domain can then be found as

$$\begin{aligned} \mathbf{K_{e_z}} &= \mathbf{H}^{-1}\mathbf{K_{e_y}}\mathbf{H}^{-t} \\ &= \mathbf{H}^{-1}\left[\mathbf{H}^{-1}\mathbf{K_{e_y}}\right]^t. \end{aligned} \tag{1}$$

It is argued here, and there is significant precedence for doing so [3], that a Gaussian probability distribution function provides a good description of the quantization error in the pixel domain. The primary justification is due to the basis image summation that forms a reconstructed pixel error—the quantization error for a single pixel will consist of the sum of wavelet-domain quantization errors for each basis image that overlaps with the pixel. Thus we model the spatial-domain compression error as zero-mean Gaussian with covariance $\mathbf{K_{e_z}}$, $\mathcal{N}(\mathbf{0}, \mathbf{K_{e_z}})$, where the covariance matrix may be expanded according to (1).

Individual wavelet domain quantization error variances that compose $\mathbf{K_{e_y}}$ determine the error in the pixel domain. For high-rate situations, the quantization step sizes used for compression are small enough that a uniformly distributed random variable accurately models the quantization error in the wavelet domain; Watson et al. [4] and Karray et al. [3] also use this model. In such cases, the diagonal components of $\mathbf{K_{e_y}}$ are composed of terms $\frac{1}{12}\Delta_i^2$, where Δ_i is the quantization step size for the i^{th} wavelet coefficient. For lower-rate situations, analysis is more complicated. Such situations are not considered here, and the interested reader is referred to other work [5] for an analogous analysis of the low-rate case when using the discrete cosine transform (DCT).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

3. TEMPORAL FILTERING

This section demonstrates a temporal-filtering method of restoring motion imagery compressed by scalar quantization of either 2D or 3D wavelet coefficients. Filtering along the time domain allows the incorporation of information from surrounding frames for the restoration of the current frame. Such ideas have been used in the super-resolution literature, where the goal is to improve the spatial resolution by using information from surrounding frames; here, however, the goal is to improve the quality of a frame without changing resolution. Some examples of temporal filtering in this manner can be found in the literature [6, 7]. The section begins by formulating a general framework for temporal filtering, followed by several subsections of experimental results. The first subsection provides results for synthetically-generated video whose individual frames have been compressed by using the 2D DWT. Subsection 3.2 extends these results to the case of compression with a 3D DWT. Subsection 3.3 considers the case of actual video, showing that benefits of using the proposed quantization noise model apply for real video as well as the synthetic sequences used in the previous subsections.

Suppose the original noise-free image at time k is $\mathbf{z}^k$. Since a temporal filter uses pixel information from frames at different time instants, one must first describe the relationships between the images at different times. One common method of relating images at two time instants k and l is to model them such that [8]

$$\mathbf{z}^l = \mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} + \mathbf{e}^{k,l}, \tag{2}$$

where the matrix $\mathbf{A}_{k,l}$ forms a prediction of $\mathbf{z}^l$ given $\mathbf{z}^k$, $\mathbf{e}^{k,l}$ is the error in such a prediction, and $\boldsymbol{\mu}^{k,l}$ is a "mean" term that accounts for pixels in $\mathbf{z}^l$ that are unobservable from $\mathbf{z}^k$. Note that when $l = k$, $\mathbf{A}_{k,k} = \mathbf{I}$, $\mathbf{e}^{k,k} = \mathbf{0}$, and $\boldsymbol{\mu}^{k,k} = \mathbf{0}$. When the original images are compressed, the model that relates the observation at time l to the original image at time k must be modified according to the quantization error,

$$\mathbf{z}_q^l = \mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} + \mathbf{e}^{k,l} + \mathbf{e}_{\mathbf{z}}^l, \tag{3}$$

where the additional noise term is the same quantization error term introduced in Section 2. For notational convenience, the two noise terms can be combined into a single noise term,

$$\mathbf{z}_q^l = \mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} + \mathbf{n}^{k,l}. \tag{4}$$

The error is assumed to be normally distributed, with mean $\mathbf{0}$ and covariance $\mathbf{K}_{k,l}$, which implies that the probability distribution function of $\mathbf{z}_q^l$ given $\mathbf{z}^k$ is $\mathcal{N}(\mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l}, \mathbf{K}_{k,l})$,

$$p(\mathbf{z}_q^l|\mathbf{z}^k) = \frac{1}{(2\pi)^{N/2}|\mathbf{K}_{k,l}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} - \mathbf{z}_q^l)^t \mathbf{K}_{k,l}^{-1}(\mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} - \mathbf{z}_q^l)\Big\}. \tag{5}$$

With the simplifying approximation that $\mathbf{z}_q^l|\mathbf{z}^k$ and $\mathbf{z}_q^m|\mathbf{z}^k$ are independent for $m \neq l$, the joint distribution of observations of compressed images at times $k-n, \ldots, k+n$ is

$$p(\mathbf{z}_q^l|\mathbf{z}^k, l = k-n, \ldots, k+n) = \prod_{l=k-n}^{k+n} p(\mathbf{z}_q^l|\mathbf{z}^k). \tag{6}$$

The temporal filter described in this section finds the maximum likelihood (ML) estimate of a single frame $\mathbf{z}^k$ given degraded observations of the sequence at time instants within $\pm n$ of k, i.e., $k-n, \ldots, k+n$. The maximum likelihood estimator chooses an estimate for $\mathbf{z}^k$ that maximizes the likelihood term in (6). Since maximizing a function is equivalent to minimizing the negative of its natural logarithm, the ML estimate can be written as

$$\hat{\mathbf{z}}^k = \arg\min_{\mathbf{z}^k} \sum_{l=k-n}^{k+n} (\mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} - \mathbf{z}_q^l)^t \mathbf{K}_{k,l}^{-1}(\mathbf{A}_{k,l}\mathbf{z}^k + \boldsymbol{\mu}^{k,l} - \mathbf{z}_q^l). \tag{7}$$

Elements of the mean term $\boldsymbol{\mu}^{k,l}$ are chosen as

$$\boldsymbol{\mu}^{k,l} = \begin{cases} 0 & \text{for observable pixels} \\ \mathbf{z}_q^l & \text{for unobservable pixels} \end{cases}. \tag{8}$$

For pixels at time l that do not correspond to any pixels at time k, the corresponding rows of $\mathbf{A}_{k,l}$ consist entirely of zeros. To classify a pixel at time l as having or not having corresponding pixels at time k, a simple threshold can be applied to the difference $|\mathbf{z}_q^l - \mathbf{A}_{k,l}\mathbf{z}_q^k|$ after motion estimation; knowledge of pixels appearing or disappearing at image edges due to camera motion can also be used. More sophisticated algorithms could be used to estimate both $\mathbf{A}_{k,l}$ and $\boldsymbol{\mu}^{k,l}$ simultaneously, although they are not explored here.

The various temporal filtering schemes presented in this section differ in their dimensionality (two- or three-dimensional DWT) as well as their choice of covariance matrix $\mathbf{K}_{k,l}$, but the general problem setup is that of (7). Equation (7) is solved iteratively using a conjugate gradient optimization algorithm.

The following subsections discuss various methods and experiments of temporally filtering wavelet-compressed motion imagery.

3.1. Compression with 2D DWT

Consider the case where each frame of an image sequence is compressed by use of a 2D DWT, independently of the other frames. This subsection considers the simplified case where both the horizontal and vertical motion between frames is an integer number of pixels. The simplest of temporal filters assumes Gaussian noise with covariance matrix $\mathbf{K}_{k,l}$ that is diagonal with elements σ^2 for each $l = k-n, \ldots, k+n$. Such a filter equally weights all observations of each frame at each time instant, and can be loosely interpreted as a box filter in the temporal dimension along the motion trajectories described by the $\mathbf{A}_{k,l}$ matrices. The other case considered here is to use Gaussian noise as derived previously with $\mathbf{K}_{k,l} = \mathbf{K}_{\mathbf{e}_{\mathbf{z}}^l}$, which cannot be implemented by simple averaging, but is instead implemented with a conjugate gradient method. Both of these noise models for $\mathbf{K}_{k,l}$ consider only the noise introduced by compression.

The initial experiment given in this subsection is designed to isolate the effect of the quantization noise model. The input sequence consists of five frames, where each frame is taken from a single source image and globally translated with integer shift; thus the only difference between these five images is that they are global translates of each other, where the global motion parameters are known. In a very simple sense, these five images form an artificially-constructed "image sequence." Obviously, sequences with such a property will almost never occur in actual video; however, the point of this subsection is to isolate the effect of the quantization noise model, and the method discussed here eliminates

Sequence name	Compressed PSNR, dB	IID PSNR, dB	$\mathbf{K}_{\mathbf{e}_\mathbf{z}^l}$ PSNR, dB
barb	32.86 ± 0.02	36.00	37.15
boat	30.49 ± 0.03	32.65	33.71
mandrill	32.29 ± 0.01	35.52	36.83
peppers	32.50 ± 0.03	33.81	34.37
test-pattern	31.10 ± 0.06	33.26	34.24
test-pattern	32.84 ± 0.06	35.10	36.14
test-pattern	34.73 ± 0.07	37.10	38.09

Table 1. PSNR values for restoration using various quantization noise models. The range of PSNR's under "Compressed" refers to the PSNR's for the five compressed input frames.

the influence of other motion-related error such as $\mathbf{e}^{k,l}$ or uncertainty in the matrices $\mathbf{A}_{k,l}$.

The integer shifts that relate the five frames are (0, 0), (1, 0), (0, 1), (−1, 0), and (0, −1). The compressed images are formed by quantization of four-level DWT decompositions of the images, where the quantization is performed as described in [9]. The wavelet filters here and elsewhere in this paper are the 9/7 Daubechies biorthogonal DWT [10]. The frames used for compression are the central 384×384 portions of the original 512×512 image, which allows the shifted frames to be extracted without missing pixels at the image borders.

Table 1 summarizes results for the test images using both of the quantization noise models. As can be seen from Table 1, using the derived quantization error covariance matrix produces PSNR results that are significantly better than those produced when assuming IID compression error.

3.2. Compression with 3D DWT

Here, we repeat the experiment of Subsection 3.1 by replacing the two-dimensional DWT by the three-dimensional DWT, where three levels of decomposition are applied in the temporal dimension. Again, two noise models are compared: Using the full $\mathbf{K}_{\mathbf{e}_\mathbf{z}^l}$, and using an IID model. The length of the transform in the temporal direction is sixteen; the image sequence is synthesized by shifts applied to a prototype image of (0, 0), (1, 0), (1, 1), and continuing in an outward counter-clockwise spiral to the final shift of (−1, 2). All of the sixteen frames in this group of pictures are filtered to produce the estimate of the original image.

Quantitative results for the experiment are shown in Table 2. Significant PSNR improvements are evident relative to the received noisy images, but this is to be expected since the filtering is able to make use of sixteen noisy versions of the exact same original frame; in real-life situations, rarely will such fortunate circumstances arise. The important conclusions to be drawn from the results are not in regard to PSNR improvement relative to the noisy images, but rather the difference in PSNR improvements: again there is significant gain in using the full theoretic covariance matrix $\mathbf{K}_{\mathbf{e}_\mathbf{z}^l}$ relative to the IID assumption.

3.3. Experiment with Real Video

While the previous two subsections dealt with synthetically-generated video, such situations are not necessarily indicative of the results one would obtain with real video sequences. This subsection considers temporal filtering of actual motion imagery that has been compressed using the

Sequence name	PSNR, dB		
	Quantized	Gaussian IID	Gaussian $\mathbf{K}_{\mathbf{e}_\mathbf{z}^l}$
barb	36.27	40.19	42.38
boat	34.01	36.69	39.17
mandrill	31.30	35.34	38.50
peppers	35.35	36.77	37.90
test-pattern	36.24	39.15	41.14

Table 2. Restoration results for compression that uses a three-dimensional DWT. The PSNR listed under "Quantized" is the highest of the PSNR's for the 16 input images

2D DWT. We will consider two cases: When the observation error is dominated by quantization noise, and hence the motion-compensation error $\mathbf{e}^{k,l}$ can be neglected; and when $\mathbf{e}^{k,l}$ is included in the formulation. For each of the two cases, two models are used for the quantization error—IID Gaussian noise with covariance matrix of $\sigma^2\mathbf{I}$, and Gaussian noise with covariance $\mathbf{K}_{\mathbf{e}_\mathbf{z}^l}$.

In previous subsections, the motion between input frames was known exactly because of the artificial construction of the sequences. With real video, the motion must be estimated. Here we consider video sequences that contain frames that differ by a global transformation, e.g., stationary scenes captured by a moving camera that is located at a far distance from the actual scene, as might be expected from aerial surveillance video. Stationary scenes are not a requirement, but they make motion estimation simpler; with more sophisticated motion estimation, the algorithm described here could be applied equally well. To register frames of the input imagery, an affine motion model is employed. The six parameters of the affine model that relate the two frames are estimated iteratively within a coarse-to-fine multiresolution pyramid. Note that while previous sections assumed integer pixel motions, the more general model here allows for floating-point pixel motions; $\mathbf{A}_{k,l}$ matrices are constructed based on bilinear interpolation for these fractional pixel motions.

Results are shown for five frames of the *stickers* sequence; the author acquired this uncompressed sequence using a Pixelink PL-A641 monochrome camera. While the sequence is certainly not the same as aerial surveillance video, it does share certain qualities—a large global-motion component, as well as significant detail at fine resolutions. The writing on the stickers will serve as a sort of resolution chart for comparison of the restoration algorithms. For this example, the compressed versions of these five frames will be used to reconstruct the central frame.

Results are first presented under the assumption that the quantization error dominates the overall error $\mathbf{n}^{k,l}$. As in previous subsections, the compression is simulated by scalar quantization of the coefficients from a four-level DWT decomposition. Figure 1 presents results for the central image when the frames are compressed to approximately 32.3 dB.

In Figure 1, both the quantization noise model and the IID noise model show considerable improvement over the original compressed image, which shows the potential advantages of temporal filtering for this type of compressed sequence. The visual improvement of the quantization noise model relative to the IID noise model is evidenced by an overall sharper reconstruction, with more apparent contrast. The PSNR is also slightly higher for the restoration using the quantization noise model.

To better account for the motion compensation error,

(a) (b) (c)

Fig. 1. Zoomed close-ups of restoration results for real video data. The five input images have PSNR values in the range of 32.2 dB to 32.4 dB. (a) Unprocessed; (b) filtered according to $\mathbf{K}_{\mathbf{e_z}}$; and (c) filtered according to IID. PSNR values for the full 640×480 frames are (a) 32.23 dB; (b) 34.91 dB; and (c) 34.58 dB.

one can explicitly model the term $\mathbf{e}^{k,l}$ such that the overall noise term would be $\mathbf{n}^{k,l} = \mathbf{e}_{\mathbf{z}}^{l} + \mathbf{e}^{k,l}$. The motion compensation noise term $\mathbf{e}^{k,l}$ has often been modeled as IID-Gaussian distributed [8] with variance terms $\lambda^{k,l}$; note that $\lambda^{k,l}$ is zero for $k = l$. With such an assumption, the overall noise term for the IID quantization noise case would have covariance matrix

$$\begin{aligned} \mathbf{K}_{k,l} &= \sigma^2\mathbf{I} + \lambda^{k,l}\mathbf{I} \\ &= \widehat{\lambda}^{k,l}\mathbf{I}, \end{aligned} \tag{9}$$

which for practical purposes is nearly equivalent to the IID quantization noise model used previously. For the non-IID quantization noise model, the covariance matrix becomes

$$\mathbf{K}_{k,l} = \mathbf{H}^{-1}\mathbf{K}_{\mathbf{e}_{\mathbf{y}}^{l}}\mathbf{H}^{-t} + \lambda^{k,l}\mathbf{I}. \tag{10}$$

The restoration algorithms require the inversion of the above covariance matrix, but since only the product of the matrix inverse with some input vector is needed—and not the actual explicit inverse matrix—iterative methods can be employed when using (10). Explicitly modeling the motion compensation error $\mathbf{e}^{k,l}$ within the overall noise $\mathbf{n}^{k,l}$ leads to improvements in the restorations of both IID and non-IID quantization noise models, with the performance of the non-IID model being superior. Results are not provided here due to space limitations.

4. CONCLUSION

This paper has discussed temporal filtering as a method of restoration for motion imagery compressed by quantization in the wavelet-transform domain. It was shown that proper modeling of the compression error can lead to significant improvements, in terms of both objective and subjective quality, compared to the common assumption of IID noise. Experimental results included temporal filtering for synthetic video compressed with 2D and 3D wavelet transforms, as well as for real video sequences compressed with a 2D wavelet transform. The results of this work are applicable to numerous scenarios, including Motion JPEG2000 [1] and 3D SPIHT [2]. Other restoration applications that make use of temporal information, such as super-resolution, could also benefit from the noise modeling framework introduced here. Further details on the work reported in this paper can be found in a more lengthy technical report [11].

5. REFERENCES

[1] ISO/IEC 15444-3, "Information technology—JPEG 2000 image coding system—Part 3: Motion JPEG 2000," 2002.

[2] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. on CSVT*, vol. 10, no. 8, pp. 1374–1387, Dec. 2000.

[3] L. Karray, P. Duhamel, and O. Rioul, "Image coding with an L^∞ norm and confidence interval criteria," *IEEE Trans. on Image Proc.*, vol. 7, no. 6, pp. 621–631, May 1998.

[4] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. on Image Proc.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[5] M. A. Robertson and R. L. Stevenson, "DCT quantization noise in compressed images," in *Int. Conf. on Image Proc.*, Oct. 2001, vol. 1, pp. 185–188.

[6] Y. Yang, M. Choi, and N. Galatsanos, "New results on multichannel regularized recovery of compressed video," in *Int. Conf. on Image Proc.*, Oct. 4–7 1998, vol. 1, pp. 391–395.

[7] C.-J. Tsai, P. Karunaratne, N. Galatsanos, and A. Katsaggelos, "A compressed video enhancement algorithm," in *Int. Conf. on Image Proc.*, Oct. 23–27 1999.

[8] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. on Image Proc.*, vol. 5, no. 6, pp. 996–1011, June 1996.

[9] J. W. Woods and T. Naveen, "A filter based bit allocation scheme for subband compression of HDTV," *IEEE Trans. on Image Proc.*, vol. 1, no. 3, pp. 436–440, July 1992.

[10] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. on Image Proc.*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[11] M. A. Robertson, "Restoration of wavelet-compressed images and motion imagery," AFRL Technical Report AFRL-IF-RS-TR-2004-5, 2004, available from http://www.dtic.mil.