AN ESTIMATION-THEORETIC APPROACH TO VIDEO DENOISEING

Jingning Han, Timothy Kopp, and Yaowu Xu

Google Inc.
1600 Amphitheatre Parkway, Mountain View, CA 94043
Emails: {jingning,tkopp,yaowu}@google.com

ABSTRACT

A novel denoising scheme is proposed to fully exploit the spatio-temporal correlations of the video signal for efficient enhancement. Unlike conventional pixel domain approaches that directly connect motion compensated reference pixels and spatially neighboring pixels to build statistical models for noise filtering, this work first removes spatial correlations by applying transformations to both pixel blocks and performs estimation in the frequency domain. It is premised on the realization that the precise nature of temporal dependencies, which is entirely masked in the pixel domain by the statistics of the dominant low frequency components, emerges after signal decomposition and varies considerably across the spectrum. We derive an optimal non-linear estimator that accounts for both motion compensated reference and the noisy observations to resemble the original video signal per transform coefficient. It departs from other transform domain approaches that employ linear filters over a sizable reference set to reduce the uncertainty due to the random noise term. Instead it jointly exploits this precise statistical property appeared in the transform domain and the noise probability model in an estimation-theoretic framework that works on a compact support region. Experimental results provide evidence for substantial denoising performance improvement.

Index Terms— Video denoising, motion compensation, estimation theory, discrete cosine transform

1. INTRODUCTION

Video noise introduced during signal acquisition impairs image quality. Its randomness nature very often negatively impacts the compression efficiency of the subsequent video codec. Video denoising has been a challenging and actively studied area for decades [1]. There is a considerable volume of prior research focused on various issues including robust spatial/temporal reference search [2]-[5] and adaptive filtering [6]-[11].

We are concerned with developing an efficient denoiser for real-time communication as part of the WebRTC coding engine. It requires the video processing unit including both compression codec and pre/post processing to be able to handle a 720p sequence at 30 fps on a single thread general purpose processor. A fairly common approach for real-time communication is to integrate the denoiser into the video codec, which allows the block motion information to be used by both denoiser and compressor, hence the system only requires a single motion estimation process per block, thereby substantially reducing the overall computational load [12]. The underlying assumption is that the video signal dominates motion search even in the presence of acquisition noise and hence the resulting motion trajectories largely reflect the true motion flow, which generally holds true for typical video conferencing scenarios particularly at 480p and above frame size. Such practical constraints rise the major challenge that we will address here on how to exploit the *compact* support region, i.e., motion compensated reference and spatially nearby pixels, for efficient noise reduction.

Classic pixel domain video denoisers typically extend various linear filtering approaches to image denoising to further incorporate the temporal reference pixels [1], thereby exploiting the additional temporal coherence in video signal. For example, the spatially adaptive local linear minimum mean squared error (LMMSE) filter [13] has been generalized to a spatio-temporal LMMSE estimator in [7]. The 2-D Kalman filter for image restoration [14] is extended to include the motion compensated pixels in a 3-D Kalman filter framework [11]. However, building a precise linear model that jointly describes both the spatial and temporal correlations with respect to the pixel of interest remains undiscovered. Moreover, there exist difficulties in calculating the Kalman gain matrix that accounts for motion compensated reference pixels, in the context of the advanced (and more sophisticated) sub-pixel motion reference scheme. A notable adaptive weighted averaging (AWA) filter approach is proposed in [6], where the weight of each surrounding pixel (both temporal and spatial) decreases with its distance to the observed pixel value.

The wavelet techniques transform the image frames and apply certain shrinkage constraints to remove the noise component [8, 9]. Recent research work propose to employ a multi-stage motion estimation in the wavelet domain for robust motion compensated referencing [15], and to adaptively select the linear filter weights applied to the wavelet transform coefficients according to the reliability of motion vectors [5]. The wavelet transforms often require a rather large support region, i.e., multiple image frames, which may largely limit the cache performance of general computing platform and are not quite suitable for real-time scenarios. A block transform coefficient estimation approach based on spatio-temporal Gaussian scale mixture model is proposed in [3], where the motion estimation is conducted over multiple reference frames including future frames to construct the motion trajectory per block. A even more comprehensive reference search is proposed in [4] to locate non-local reference blocks in the current frame, as well as prior and future frames. When certain repetition exists in the image content, it is possible to create a sizable reference block set which statistically reduces the uncertainty due to additive noise in the form of Wiener filter. The performance can potentially be further enhanced by employing a redundant representation dictionary, which is trained and maintained to maximize the sparseness of video signal in the transform domain [16].

The real-time constraints, however, preclude the use of the above approaches that involve multiple runs of motion estimation and require reference block from future frames. The available support region very often simply comprises of a motion compensated block from previously denoised frame and the spatially neighboring pixels of the current frame. Therefore, we will primarily focus on devising a video denoiser with compact support region in this work. It is premised on the realization that the true temporal correlations emerge only in the transform domain and significantly vary across the frequency spectrum [17]. Unlike the conventional methods that build linear filters to estimate the original signal either in pixel [7, 6] or transform domain [3, 4], the proposed approach exploits this precise statistical property and the noise model in an estimation-theoretic framework that produces a non-linear optimal estimate, in the mean squared error sense, per transform coefficient. It is experimentally shown that this estimation-theoretic denoising framework substantially outperforms other competitors in the setting of compact support region. We note that the proposed approach can be readily extended to incorporate multiple blocks along the motion trajectory, provided a sizable reference set, which is beyond the scope of this work.

2. RELATED WORK

2.1. Adaptive Weight Filtering

We find the adaptive weighted averaging (AWA) filter [6] work fairly well in the context of compact support region, which we briefly revisit here.

Let $f(m, n; k)$ and $u(m, n; k)$ denote the true pixel value and the additive noise at position (m, n) in frame k, respectively. The noisy observation is denoted by

$$g(m, n; k) = f(m, n; k) + u(m, n; k).$$

The AWA estimate $\hat{f}(m, n; k)$ is defined as

$$\tilde{f}(m, n; k) = \sum_{(i,j;l)\in S_{m,n;k}} w(i, j; l)g(i, j; l), \qquad (1)$$

where

$$w(i, j; l) = \frac{K(m, n; k)}{1 + max\{2\sigma^2, (g(m, n; k) - g(i, j; l))^2\}} \qquad (2)$$

are the weights associated with pixels in the support $S_{m,n;k}$, σ^2 is the noise variance, and $K(m, n; k)$ is a normalization factor

$$K(m, n; k) \quad = \quad (\sum_{(i,j;l)\in S_{m,n;k}} \frac{1}{1 + max\{2\sigma^2, (g(m, n; k) - g(i, j; l))^2\}})^{-1}.$$

2.2. Wiener Filter

The Wiener filter is frequently employed to estimate the coefficients by many transform domain approaches [4]. Let $x_n(i, j; p)$ be the actual transform coefficient of frequency index (i, j) in block p of frame n. For exposition simplicity, we denote it as x_n hereafter. Similarly, we denote the noise term as v_n and the observation $\hat{x}_n = x_n + v_n$. Given the noisy observations, the Wiener filter approach estimates x_n as

$$\tilde{x}_{n,wie} = \frac{s^2}{s^2 + \sigma^2}(\hat{x}_n - \bar{x}_n) + \bar{x}_n, \qquad (3)$$

where s^2 and σ^2 are the variances of the signal x_n and noise terms, respectively, and $\bar{x}_n$ denotes the mean value of x_n. Clearly, the efficacy of Wiener filter approach (3) critically relies on the fact that there is a sufficiently large set of reference samples x_n to allow an accurate estimate of its first and second order moments, which is not valid in the compact support region setting.

3. ESTIMATION-THEORETIC DENOISING FRAMEWORK

3.1. Statistical Model

Following the notations in Sec. 2.2, we use x_n to denote a coefficient at a specific frequency in the current block, and x_{n-1} the corresponding reference block coefficient at the same frequency, located in the previous frame. The evolution of transform coefficients along the motion trajectory is modeled by the auto-regressive (AR) process

$$x_n = \rho x_{n-1} + z_n, \qquad (4)$$

where ρ is the correlation coefficient of consecutive samples, and the driving innovation variables denoted by $\{z_n\}$ are independent and identically distributed with probability density function $p_Z(z)$.

Much study has been devoted to establishing the probability distribution $p_Z(z_n)$ of the innovation term z_n, e.g., [18]-[22]. It is commonly recognized that this density is well approximated by a zero-mean Laplacian distribution:

$$p_Z(z_n) = \frac{\lambda}{2} e^{-\lambda |z_n|}, \tag{5}$$

whose statistical characteristics are determined by the model parameter λ. The maximum likelihood estimate of λ, given outcomes $z_0, \cdots, z_{N-1}$ of N independent draws of the random variable Z, is

$$\lambda_{ML} = \frac{N}{\sum_{i=0}^{N-1} |z_i|}. \tag{6}$$

Ideally, one would need to obtain the innovations of each motion trajectory, per frequency, from the original video signal, and substitute in (6) to estimate the corresponding Laplacian parameter. For simplicity, we obtain these parameters from an offline training set and arbitrarily use them for all test sequences. We also set the correlation coefficient ρ to unit. A comprehensive yet sophisticated treatment to maintain spatio-temporal adaptive innovation probability model can be found in [23].

3.2. Estimation-Theoretic Denoiser

Let $\hat{x}_n = x_n + v_n$ denote the noisy observation of x_n, where v_n is the additive noise with probability density function $p_V(v_n)$, and $\tilde{x}_{n-1}$ denote the corresponding motion compensated reference from previously denoised reference frame. The optimal estimate of x_n given $\hat{x}_n$ and $\tilde{x}_{n-1}$ is

$$\tilde{x}_{n,et} = E\{x_n | \hat{x}_n, \tilde{x}_{n-1}\} = \int x_n f(x_n | \hat{x}_n, \tilde{x}_{n-1}) dx_n, \tag{7}$$

the expectation over a joint conditional probability density function $f(x_n|\hat{x}_n, \tilde{x}_{n-1})$[1].

Applying Bayesian rule, one can obtain

$$\begin{aligned} f(x_n|\hat{x}_n, \tilde{x}_{n-1}) &= \frac{f(x_n, \hat{x}_n, \tilde{x}_{n-1})}{f(\hat{x}_n, \tilde{x}_{n-1})} \\ &= \frac{f(\hat{x}_n, \tilde{x}_{n-1}|x_n) f(x_n)}{f(\hat{x}_n, \tilde{x}_{n-1})}. \end{aligned} \tag{8}$$

According to Markov property of the AR process and assuming $\tilde{x}_{n-1} \approx x_{n-1}$, $f(\hat{x}_n, \tilde{x}_{n-1}|x_n)$ can be decomposed by

$$f(\hat{x}_n, \tilde{x}_{n-1}|x_n) \approx f(\hat{x}_n|x_n) \cdot f(\tilde{x}_{n-1}|x_n), \tag{9}$$

i.e., the observation $\hat{x}_n$ and motion compensated reference $\tilde{x}_{n-1}$ are independent conditioned on the true sample x_n. Substituting (9) into (8) gives

$$\begin{aligned} f(x_n|\hat{x}_n, \tilde{x}_{n-1}) &= \frac{f(\hat{x}_n, \tilde{x}_{n-1}|x_n) f(x_n)}{f(\hat{x}_n, \tilde{x}_{n-1})} \\ &\approx \frac{f(\hat{x}_n|x_n) f(\tilde{x}_{n-1}|x_n) f(x_n)}{f(\hat{x}_n, \tilde{x}_{n-1})} \\ &= \frac{f(\hat{x}_n|x_n) f(x_n|\tilde{x}_{n-1}) f(\tilde{x}_{n-1})}{f(\hat{x}_n, \tilde{x}_{n-1})} \\ &= \frac{f(\hat{x}_n|x_n) f(x_n|\tilde{x}_{n-1})}{f(\hat{x}_n|\tilde{x}_{n-1})}, \end{aligned} \tag{10}$$

where the last two steps follow Bayesian rule.

Note that

$$f(\hat{x}_n|x_n) = P_V(x_n + v_n|x_n), \tag{11}$$

is the probability density function of noise term centered at x_n. Assume $\tilde{x}_{n-1} \approx x_{n-1}$, we have

$$f(x_n|\tilde{x}_{n-1}) \approx P_Z(\tilde{x}_{n-1} + z_n|\tilde{x}_{n-1}), \tag{12}$$

i.e., the probability density function of the innovation term shifted to be centered at $\tilde{x}_{n-1}$. Clearly the joint probability density function (10) can be well approximated by the product of two marginal probability density functions, where the denominator $f(\hat{x}_n|\tilde{x}_{n-1})$ is a normalization factor and is independent of variable x_n. The expectation (7) can hence be translated into

$$\begin{aligned} \tilde{x}_{n,et} &= E\{x_n|\hat{x}_n, \tilde{x}_{n-1}\} \\ &= \frac{\int x_n P_V(\hat{x}_n - x_n) P_Z(x_n - \tilde{x}_{n-1}) dx_n}{\int P_V(\hat{x}_n - x_n) P_Z(x_n - \tilde{x}_{n-1}) dx_n}. \end{aligned} \tag{13}$$

We evaluate the performance in the context of additive zero-mean Gaussian model for preliminary results next. It is noteworthy that the proposed estimation-theoretic denoising approach is generally applicable to a large variety of noise models, provided the probability characteristics.

When the motion compensated reference is much distant from the observed current block, i.e., the temporal correlations are less pronounced, we apply a subsequent spatial denoiser in light of Sec. 2.1 in our implementation.

4. EXPERIMENTAL RESULTS

We implemented the proposed approach (13) as well as the AWA filter approach [6] in the framework of VP9 [24] real-time coding mode. Regular motion estimation was conducted at $1/8$ sub-pixel accuracy level over a single reference frame, with respect to the *unfiltered* version of current frame. The resulting motion vectors were used by the denoiser to fetch corresponding motion compensated reference block from previously denoised and *uncoded* frame. We note that this framework implicitly assumes that the actual signal dominates motion search process. The motion alignment accuracy may be enhanced by employing advanced multi-stage

[1] We use streamlined $f(.)$ to denote probability density function hereafter when there is no risk of ambiguity.

wavelet domain motion estimation approach [5, 15], hence further improving the proposed estimation-theoretic denoiser performance. We focus on the estimation efficiency in this work and let both schemes use the same regular block motion estimation process.

For simplicity, we used the synthetic zero-mean white Gaussian noise to demonstrate the performance of our proposed approach, while we emphasized that the scheme (13) would readily handle other noise types provided the probability density models. The noise terms were added to the original video sequences, which were fed into the video codec. We compared the denoised *uncoded* version to the original video sequence in terms of PSNR at various noise variance levels. The results were shown in Fig.1-3. Clearly when strong temporal correlation existed in the video signal, the propose estimation-theoretic denoiser significantly outperformed the AWA spatio-temporal filtering approach (Fig. 1 and 2). For sequences with less temporal redundancy, both schemes subsumed to spatial denoiser more frequently and the performance gains due to estimation-theoretic denoiser were less pronounced as seen in Fig.3.

5. CONCLUSIONS

We derive an estimation-theoretic denoiser that jointly exploits the precise statistical properties of video signal and the noise probability model in the transform domain to optimally estimate the original signal per transform coefficient. The approach is built on a compact support region and can be integrated into a video compression codec for efficient computing. It is experimentally shown that the proposed estimation-theoretic denoiser substantially outperforms the conventional schemes that allow a similar compact support region.

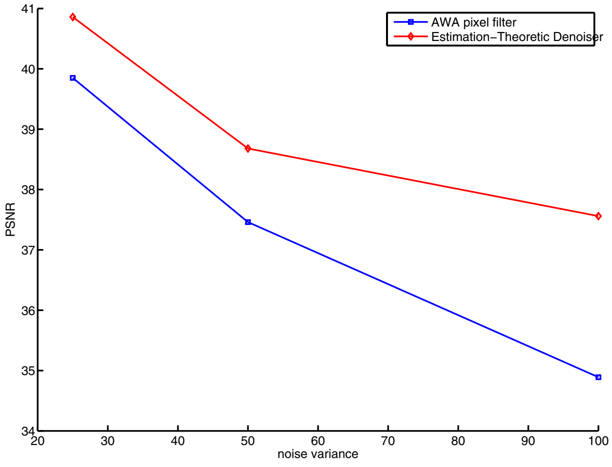

Fig. 1. Performance comparison of AWA filter and the proposed estimation-theoretic denoiser. The test sequence is $vidyo1$ at $720p$ resolution.

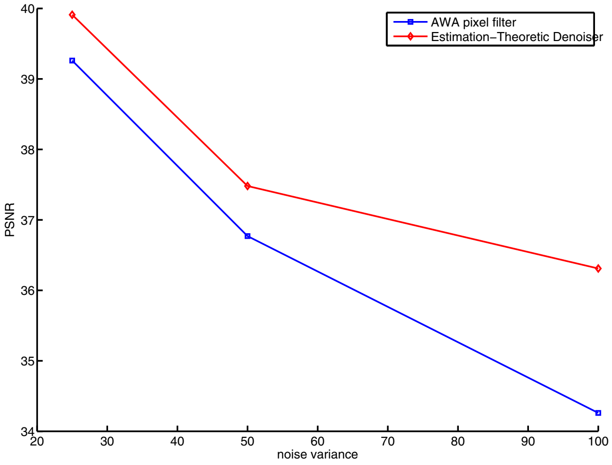

Fig. 2. Performance comparison of AWA filter and the proposed estimation-theoretic denoiser. The test sequence is $vidyo3$ at $720p$ resolution.

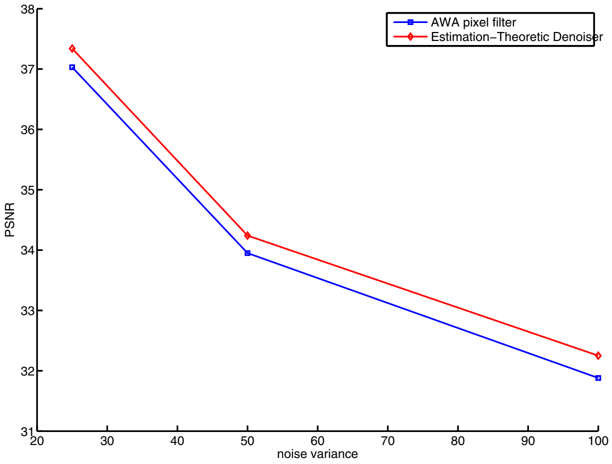

Fig. 3. Performance comparison of AWA filter and the proposed estimation-theoretic denoiser. The test sequence is $pedestrian\_area$ at $1080p$ resolution.

6. REFERENCES

[1] J. C. Brailean, R. P. Kleihorst, S. Efstratiadis, A. K. Katsaggelos, and R. L. Lagendijk, “Noise reduction filters for dynamic image sequences: a review,” *Proc. IEEE*, vol. 83, pp. 1272–1292, 1995.

[2] C. Liu and W. T. Freeman, “A high-quality video denoising algorithm based on reliable motion estimation,” *European Conference on Computer Vision*, pp. 706–719, 2010.

[3] G. Varghese and Z. Wang, “Video denosing based on a spatiotemporal Gaussian scale mixutre model,” *IEEE*

Trans. on Circuits and Systems for Video Technology, vol. 20, no. 7, pp. 1032–1040, 2010.

[4] M. Maggioni, G. Boracchi, A. Foi, and K. Egizarian, "Video denoising, deblocking and enhancement through separable 4-D nonlocal spatiotemporal transforms," *IEEE Trans. on Image Processing*, vol. 17, no. 8, pp. 3952–3966, 2012.

[5] V. Zlokolica, A. Pizurica, and W. Philips, "Wavelet domain video denoising based on reliability measures," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 8, pp. 993–1007, 2006.

[6] M. K. Ozkan, M. I. Sezan, and A. M. Tekalp, "Adaptive motion-compensated filtering of noisy image sequences," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 3, no. 4, pp. 277–290, 1993.

[7] R. Samy, "An adaptive image sequence filtering scheme based on motion detection," *SPIE*, vol. 596, pp. 135–144, 1985.

[8] D. L. Donoho, "Denoising by soft-thresholding," *IEEE Trans. on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] S. Rahman, M. Ahmad, and M. Swamy, "Video denoising based on inter-frame statistical modeling of wavelet coefficients," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 2, pp. 187–198, 2007.

[10] J. Boulanger, C. Kervrann, and P. Bouthemy, "Space-time adaptation for patch-based image sequence restoration," *IEEE Trans. on Pattern Analysis and machine intelligence*, vol. 29, no. 6, pp. 1096–1102, 2007.

[11] J. Kim and J. W. Woods, "Spatio-temporal adaptive 3-D Kalman filter for video," *IEEE Trans. on Image Processing*, vol. 6, no. 3, pp. 414–424, 1997.

[12] L. Guo, O. C. Au, M. Ma, and P. Wong, "Integration of recursive temporal LMMSE denoising filter into video codec," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 20, no. 2, pp. 236–249, 2010.

[13] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 165–168, 1980.

[14] J. W. Woods and C. H. Rademan, "Kalman filtering in two dimensions," *IEEE Trans. on Information Theory*, vol. 16, no. 4, pp. 473–482, 1977.

[15] F. Jin, P. Fieguth, and L. Winger, "Wavelet video denoising with regularized multiresolution motion estimation," *EURASIP Journal on Advances in Signal Processing*, pp. 1–11, 2006.

[16] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. on Image Processing*, vol. 18, no. 1, pp. 27–35, 2009.

[17] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic approach to delayed decoding of predictively encoded video sequences," *IEEE Trans. on Image Processing*, vol. 22, no. 3, pp. 1175–1185, 2013.

[18] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, "Statistical analysis of the 2D-DCT coefficients of the differential signal for images," *Signal Processing: Image Communication*, vol. 4, pp. 477–488, Nov. 1992.

[19] G. J. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. Information Theory*, vol. 42, no. 5, pp. 1365–1374, Sep. 1996.

[20] H.-M. Hang and J.-J. Chen, "Source model for transform video coder and its application - Part I: fundamental theory," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 7, no. 2, pp. 287–298, Apr. 1997.

[21] E. Lam and J. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Processing*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.

[22] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[23] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding with spatially adaptive spectral correlations," *Proc. IEEE Multimedia Signal Processing Workshop*, pp. 1–6, 2011.

[24] WebM Project, *www.webmproject.org*, 2015.