Video Denoising Based on a Spatiotemporal Gaussian Scale Mixture Model

Gijesh Varghese and Zhou Wang, *Member, IEEE*

***Abstract*—We propose a video denoising algorithm based on a spatiotemporal Gaussian scale mixture model in the wavelet transform domain. This model simultaneously captures the local correlations between the wavelet coefficients of natural video sequences across both space and time. Such correlations are further strengthened with a motion compensation process, for which a Fourier domain noise-robust cross correlation algorithm is proposed for motion estimation. Bayesian least square estimation is used to recover the original video signal from the noisy observation. Experimental results show that the performance of the proposed approach is competitive when compared with state-of-the-art video denoising algorithms based on both peak signal-to-noise-ratio and structural similarity evaluations.**

***Index Terms*—Bayesian estimation, cross correlation (CC), Gaussian scale mixture (GSM), image restoration, motion estimation, statistical image modeling, video denoising.**

I. Introduction

VIDEO SIGNALS are often contaminated by noise during acquisition and transmission. Removing/reducing noise in video signals (or video denoising) is highly desirable, as it can enhance perceived image quality, increase compression effectiveness, facilitate transmission bandwidth reduction, and improve the accuracy of the possible subsequent processes such as feature extraction, object detection, motion tracking and pattern classification.

Video denoising algorithms may be roughly classified based on two different criteria: whether they are implemented in the spatial domain or transform domain and whether motion information is directly incorporated. Spatial domain denoising is usually done with weighted averaging within local 2-D or 3-D windows, where the weights can be either fixed or adapted based on the local image content. A review of spatial domain filtering methods can be found in [1]. Transform domain methods first decorrelate the noisy signal using a linear transform (e.g., a wavelet transform), and then attempt to recover the transform coefficients of the original signal (e.g., by soft/hard thresholding [2] or Bayesian estimation [3]), followed by an inverse transform that brings the signal back to the spatial domain.

The high degree of correlation between adjacent frames is a "blessing in disguise" for signal restoration. On the one hand, since additional information is available from nearby frames, a better estimate of the original signal is expected. On the other hand, the process is complicated by the presence of motion between frames. Motion estimation itself is a complex problem and it is further complicated by the presence of noise. In [1], performance of spatial domain motion compensated filters was evaluated. A multiresolution motion estimation scheme was proposed in [4] when the signal is corrupted with noise. In [5], a recursive filter is applied on wavelet transform coefficients along an estimated motion trajectory, where the filter taps are adaptively chosen based on the "reliability" of the motion vectors. Motion information or temporal correlations may also be incorporated by employing an advanced or adapted transform [6], [7] or by using an advanced statistical model that reflects the joint distributions of wavelet coefficients over space and time [8]–[10]. Some wavelet domain algorithms [5], [11] used robust motion indices that represent motion in an indirect way. Recently, a series of successful nonlocal patch-based methods emerged [12]–[17], where motion information is incorporated implicitly by adaptively clustering similar 2-D or 3-D patches. It was also shown that imposing sparseness prior models would further improve the performance of these algorithms [16], [17].

In recent years, there has been a growing interest in studying statistical models of natural images, which provide useful prior knowledge about natural images and play important roles in the design of Bayesian signal denoising algorithms [18]. While great effort has been made to study statistical models of static natural images [19], [20], much less has been done for natural video signals. In [21], the spatiotemporal Fourier power spectra of natural image sequences were investigated. In [22], independent component analysis was applied to local 3-D blocks extracted from natural image sequences and the components optimized for independence are filters localized in space and time, spatially oriented, and directionally selective. Similar shapes of linear components were also obtained by optimizing sparseness via a matching pursuit algorithm [23]. In [24], it was observed that natural video sequences exhibit strong statistical prior of temporal motion smoothness, which can be captured by temporal local phase correlations in the

Manuscript received August 31, 2009; revised December 27, 2009 and March 30, 2010. Date of publication May 27, 2010; date of current version July 16, 2010. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada in the form of Discovery and Strategic Grants, and by the Ontario Ministry of Research and Innovation in the form of an Early Researcher Award. This paper was recommended by Associate Editor X. Li.

G. Varghese is with Maxim Integrated Products, Inc., Sunnyvale, CA 94086 USA (e-mail: gijesh.varghese@ieee.org).

Z. Wang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zhouwang@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2010.2051366

complex wavelet transform domain. The application of statistical models for signal denoising has also been extended from 2-D to 3-D. In [8], a video denoising algorithm was proposed by extending the hiddern Markov tree model [25], [26] from 2-D to 3-D, which provides a simple but effective way to describe the inter-scale statistical correlations of wavelet coefficients. In [9], a statistical model for wavelet coefficients of video signals was proposed, based on which maximum *a posteriori* estimation and temporal filtering were used for video denoising.

In this paper, we propose a new video denoising method based on a spatiotemporal model of motion compensated wavelet coefficients. Our paper was motivated by the success of the Gaussian scale mixture (GSM) model in static image denoising [18]. GSM was originally proposed in the statistics literature [27] and has been applied to the modeling of static natural images [28]. It was found to be a convenient and effective model to account for the nonGaussian marginal distributions of wavelet coefficients, as well as the variance-dependencies between neighboring wavelet coefficients. Since there exist strong correlations between wavelet coefficients across adjacent video frames, it is expected that grouping wavelet coefficients along temporal directions based on the GSM framework would help strengthen the statistical regularities of the coefficients and thus, improve the performance of Bayesian signal denoising algorithms. It is also important to be aware that video signals are more than simple 3-D extensions of 2-D static image signals, where the major distinction is the capability of representing motion information [29]. If the motion between frames is properly compensated, then the temporal correlations between wavelet coefficients can be enhanced. Effective motion compensation relies on reliable motion estimation. In the literature, a large number of local motion estimation methods have been proposed, which are mostly based on block matching (e.g., [30]), optical flow [31], [32] or phase disparity [33] evaluations. To avoid local nonlinear motion compensation processes that affect the validity of the Gaussian noise model assumed in GSM denoising, we opt to use global motion estimation methods based on cross correlation (CC) [34]–[37]. The challenge here is that the motion information must be estimated from noisy video signals rather than the noise-free original signals. To address this issue, another important aspect of our paper was the development and incorporation of a novel noise-robust motion estimation algorithm, which has not been carefully examined in previous video denoising algorithms.

II. Background of GSM Image Denoising

A random vector $\mathbf{x}$ is a GSM if it can be expressed as the product of two independent components

$$\mathbf{x} = \sqrt{z}\mathbf{u} \tag{1}$$

where $\mathbf{u}$ is a zero-mean Gaussian vector with covariance matrix $\mathbf{C_u}$, and z is called a mixing multiplier. The density of $\mathbf{x}$ is then given by

$$p_x(\mathbf{x}) = \int \frac{1}{[2\pi]^{N/2}|z\mathbf{C_u}|^{1/2}} \exp\left(-\frac{\mathbf{x}^T(z\mathbf{C_u})^{-1}\mathbf{x}}{2}\right) p_z(z)dz \tag{2}$$

where N is the size of the vector $\mathbf{x}$ and $p_z(z)$ is the mixing density. When applying it for image modeling, $\mathbf{x}$ is typically composed of a center wavelet coefficient, x_c, together with a set of coefficients located near x_c in the same wavelet subband or nearby subbands across scale and/or orientation. Among various choices of the prior density distribution $p_z(z)$, the one that was found to give superior image denoising performance is the noninformative Jeffrey's prior [18], [38] given by

$$p_z(z) \propto \frac{1}{z}\,. \tag{3}$$

The key feature of this prior is its amplitude scale invariance, which means that the inference procedure is invariant with respect to changes in the measurement units (or the scale of amplitude) [38], [39].

Assume that the original image is contaminated by additive, independent white Gaussian noise, then in the wavelet transform domain, a noisy neighborhood coefficient vector $\mathbf{y}$ can be modeled as

$$\mathbf{y} = \mathbf{x} + \mathbf{w} = \sqrt{z}\mathbf{u} + \mathbf{w} \tag{4}$$

where $\mathbf{w}$ is a noise coefficient vector with covariance matrix $\mathbf{C_w}$. The group of neighboring coefficients constitutes a sliding window that moves across the wavelet subband. At each step, only the center coefficient, x_c, of the window is estimated (i.e., denoised). Consequently, the objective here is converted to estimating x_c of $\mathbf{x}$, given the noisy observation $\mathbf{y}$.

It is not difficult to show that the Bayes least square (BLS) estimator (which minimizes the expected value of the squared estimation error given the noisy observation $\mathbf{y}$) is the conditional mean, which can be computed by [18]

$$E\{x_c|\mathbf{y}\} = \int_0^\infty E\{x_c|\mathbf{y}, z\}p(z|\mathbf{y})dz\,. \tag{5}$$

The right hand side of (5) is a 1-D integral over z, where two components, $E\{x_c|\mathbf{y}, z\}$ and $p(z|\mathbf{y})$, need to be computed for each given z. The first component $E\{x_c|\mathbf{y}, z\}$ is linear based on the facts that $\mathbf{w}$ is Gaussian and $\mathbf{x}$ is also Gaussian when conditioned on z. In particular, we have

$$E\{\mathbf{x}|\mathbf{y}, z\} = z\mathbf{C_u}(z\mathbf{C_u} + \mathbf{C_w})^{-1}\mathbf{y} \tag{6}$$

where $\mathbf{C_u}$ can be estimated from the observed noisy covariance matrix by $\mathbf{C_u} = \mathbf{C_y} - \mathbf{C_w}$. Equation (6) gives a full estimate of the original coefficient vector $\mathbf{x}$ for any given z. For our purpose here, only the center coefficient x_c is of our interest, which leads to significant simplifications of the computation [18]. The second component in the integral in (5), i.e., the posterior density $p(z|\mathbf{y})$, can also be estimated by Bayes' rule: $p(z|\mathbf{y}) \propto p(\mathbf{y}|z)p_z(z)$, where $p_z(z)$ can be calculated using

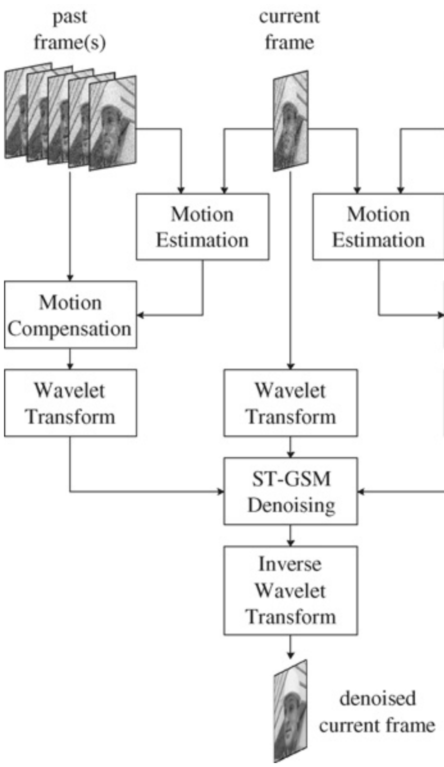


Fig. 1. Diagram of the proposed ST-GSM video denoising algorithm.

(3). Given (4) and the facts that both $\mathbf{u}$ and $\mathbf{w}$ are zero-mean Gaussian, it is straightforward to derive that $p(\mathbf{y}|z)$ can be computed as a Gaussian function of zero-mean and covariance $\mathbf{C_y} = z\mathbf{C_u} + \mathbf{C_w}$ [18].

III. Spatiotemporal GSM Video Denoising

Fig. 1 illustrates the diagram of the proposed video denoising system. The denoising of the current frame involves not only the frame itself, but also a set of adjacent past and future frames. Motion estimation is performed between the current frame and the past/future frames. Details on the motion estimation algorithm will be given in Section IV. The estimation results are used for global motion compensation. Wavelet transform is then applied to the current frame as well as the motion compensated past and future frames. The wavelet transform is a linear multiresolution analysis tool that decomposes an image signal into multiple subbands, each with a different characteristic scale and orientation. An excellent description can be found in [40]. Next, wavelet coefficient vectors are formed from a spatiotemporal neighborhood, and an spatiotemporal Gaussian scale mixture (ST-GSM) denoising method similar to the BLS estimator discussed in Section II is employed. Finally, an inverse wavelet transform is applied to the denoised wavelet coefficients to create a denoised current frame.

At the core of the proposed video denoising system shown in Fig. 1 is the ST-GSM denoising algorithm, where the key is to include temporal neighborhoods in the wavelet coefficient vector. This is illustrated in Fig. 2, where a coefficient vector of length $N_1 \times N_2 \times N_f$ is formed by a set of spatial neighboring wavelet coefficients from the same scale and the same orientation but across a number of past and future frames. Here, N_1 and N_2 denote the spatial dimensions and N_f is the total number of frames involved (including past, current and future frames). In the illustration given in Fig. 2, $N_1 = N_2 = N_f = 3$, but our experiments show that a larger number of N_f often leads to better denoising results, and we empirically set $N_f = 9$ in all of our denoising experiments.

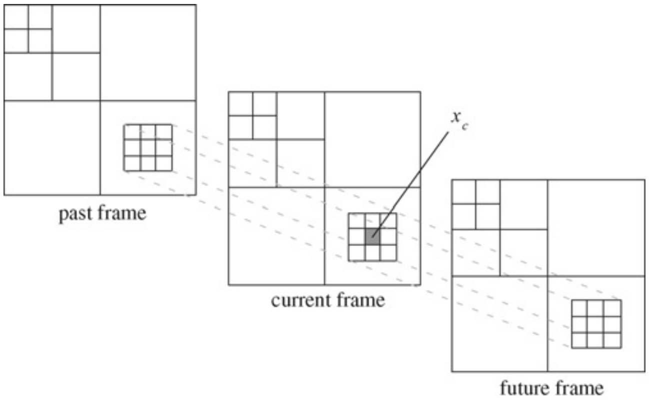


Fig. 2. Formation of 3×3×3 wavelet coefficient neighborhood over three video frames.

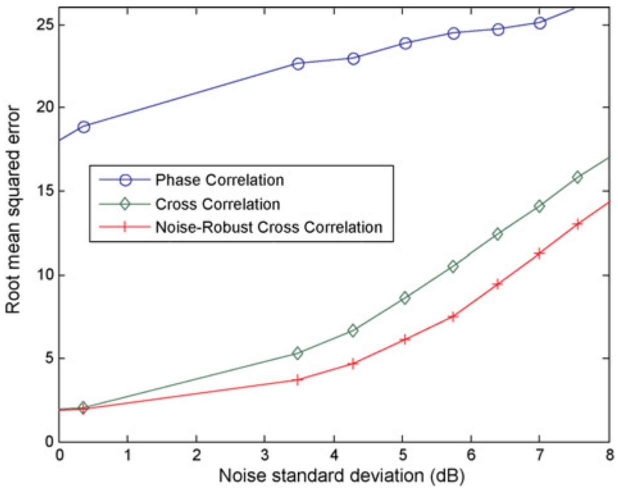


Fig. 3. RMS error comparison of phase correlation, cross correlation and noise-robust cross correlation-based motion estimation algorithms as a function of noise level (normalized noise standard deviation in decibel).

There are a number of practical issues in our implementation of the proposed ST-GSM denoising algorithm.

1) We choose an 8-orientation, 4-scale steerable pyramid [41] for wavelet decomposition. The steerable pyramid is an overcomplete wavelet transform that avoids aliasing in subbands.
2) After global motion compensation, blank regions are created at the image boundaries. We fill them by mirror replication of the shifted frame.
3) The BLS denoising of coefficient x_c in (5) involves a 1D integration over z. We implement this (using default parameters as in [18]) by sampling z in logarithm domain by 13 points in the range of -20.5 to 3.5.
4) For convenience, the first frames in a video sequence are denoised using the available past frames only. For

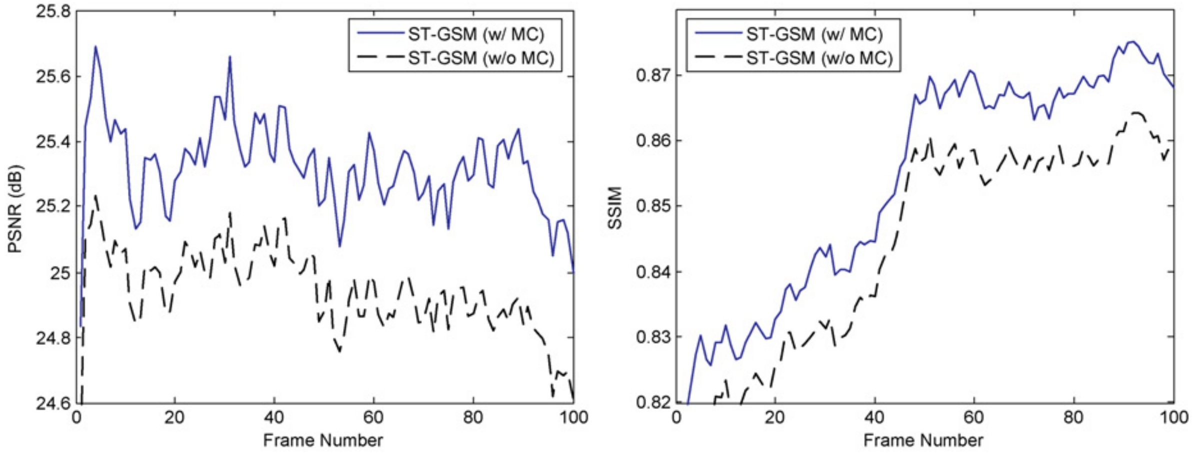


Fig. 4. PSNR and SSIM comparisons of ST-GSM denoising with and without motion estimation/motion compensation. (a) PSNR plot of denoised *Garden* sequence corrupted by noise with σ = 30, (b) SSIM plot of denoised *Garden* sequence corrupted by noise with σ = 30.

TABLE I

PSNR AND SSIM [47] COMPARISONS OF VIDEO DENOISING ALGORITHMS FOR 6 VIDEO SEQUENCES AT 5 NOISE LEVELS

Video Sequence	*Foreman*					*Salesman*					*Miss America*				
Noise std (σ)	10	15	20	50	100	10	15	20	50	100	10	15	20	50	100
PSNR Results (dB)															
Wiener2D	33.14	30.46	28.55	23.09	16.39	31.97	29.51	27.80	21.31	13.09	34.51	31.64	29.56	21.76	12.84
Wiener3D	29.54	29.26	28.87	25.35	15.28	29.59	29.30	28.88	22.92	13.50	36.95	35.60	34.06	23.39	13.27
WRSTF [5]	35.48	33.37	31.82	NA	NA	35.54	33.56	32.00	NA	NA	37.82	36.17	34.79	NA	NA
SEQWT [10]	NA	NA	NA	NA	NA	32.86	30.59	29.02	NA	NA	NA	NA	NA	NA	NA
3DWTF [6]	NA	NA	NA	NA	NA	34.96	33.33	32.03	NA	NA	NA	NA	NA	NA	NA
IFSM [9]	34.13	31.98	30.50	25.74	20.98	34.22	31.85	30.22	25.40	20.78	37.52	35.41	33.86	29.79	22.49
3DSWDCT [7]	36.17	34.46	33.07	27.33	21.46	36.98	35.12	33.75	**28.82**	**22.30**	38.87	37.72	36.74	32.88	**23.43**
VBM3D [14]	**37.17**	**35.57**	**34.36**	**28.95**	21.36	**38.33**	**36.60**	**35.12**	28.49	21.39	40.29	39.30	**38.54**	**33.39**	22.81
2DGSM [18]	35.05	33.10	31.70	26.41	20.65	33.80	31.73	30.28	24.95	20.32	38.52	37.14	36.14	30.49	22.16
ST-GSM	36.74	34.98	33.72	27.80	**21.54**	38.04	36.03	34.62	26.87	20.87	**40.58**	**39.40**	38.50	31.62	22.55
SSIM Results															
Wiener2D	0.860	0.774	0.694	0.425	0.164	0.859	0.778	0.704	0.340	0.074	0.818	0.704	0.602	0.214	0.044
Wiener3D	0.865	0.839	0.808	0.582	0.103	0.839	0.818	0.785	0.425	0.066	0.908	0.868	0.808	0.286	0.040
WRSTF [5]	0.914	0.877	0.841	NA	NA	0.932	0.901	0.868	NA	NA	0.908	0.877	0.846	NA	NA
SEQWT [10]	NA	NA	NA	NA	NA	0.900	0.846	0.796	NA	NA	NA	NA	NA	NA	NA
3DWTF [6]	NA	NA	NA	NA	NA	0.923	0.903	0.882	NA	NA	NA	NA	NA	NA	NA
IFSM [9]	0.886	0.836	0.793	0.667	0.666	0.904	0.851	0.801	0.609	0.488	0.904	0.857	0.812	0.780	0.804
3DSWDCT [7]	0.932	0.907	0.884	0.769	0.708	0.955	0.930	0.905	**0.772**	**0.611**	0.946	0.928	0.909	0.850	**0.829**
VBM3D [14]	0.935	**0.917**	**0.903**	**0.837**	0.699	**0.960**	**0.945**	**0.925**	0.771	0.538	0.947	0.939	0.933	**0.905**	0.789
2DGSM [18]	0.916	0.889	0.867	0.780	0.672	0.909	0.865	0.825	0.611	0.464	0.936	0.922	0.913	0.874	0.809
ST-GSM	**0.937**	**0.917**	0.901	0.820	**0.715**	**0.960**	0.941	0.923	0.727	0.496	**0.952**	**0.943**	**0.936**	0.892	0.823
Video Sequence	*Tennis*					*Garden*					*Football*				
Noise std (σ)	10	15	20	50	100	10	15	20	50	100	10	15	20	50	100
PSNR Results (dB)															
Wiener2D	31.07	28.55	26.78	20.65	15.46	29.73	26.77	24.80	18.09	14.17	25.07	24.76	24.37	21.19	15.23
Wiener3D	22.88	22.81	22.71	21.50	15.14	18.36	18.33	18.30	17.82	13.49	21.97	21.91	21.84	20.82	14.73
WRSTF [5]	33.68	31.35	29.71	NA	NA	30.59	27.95	26.13	NA	NA	NA	NA	NA	NA	NA
SEQWT [10]	31.19	29.14	27.59	NA	NA	29.30	26.43	24.38	NA	NA	NA	NA	NA	NA	NA
3DWTF [6]	31.96	29.91	28.56	NA	NA	30.25	27.70	25.95	NA	NA	NA	NA	NA	NA	NA
IFSM [9]	32.41	30.10	28.56	23.81	20.76	30.05	27.25	25.40	20.23	16.50	31.23	28.78	27.15	22.62	19.79
3DSWDCT [7]	33.83	31.79	30.50	26.16	**22.00**	31.80	29.40	27.70	21.50	**16.95**	32.32	29.97	28.47	23.78	19.99
VBM3D [14]	**34.89**	**32.88**	**31.49**	**27.21**	21.06	**32.54**	**30.30**	**28.74**	**22.52**	16.51	**33.09**	**30.90**	**29.36**	**24.58**	**20.04**
2DGSM [18]	31.82	29.87	28.65	24.36	20.10	30.40	27.65	25.76	19.98	16.04	31.33	29.14	27.74	23.30	19.31
ST-GSM	34.05	31.97	30.59	25.85	20.40	31.48	29.08	27.49	22.23	16.78	32.11	29.87	28.36	23.58	19.66
SSIM Results															
Wiener2D	0.813	0.712	0.625	0.288	0.092	0.916	0.853	0.792	0.413	0.165	0.703	0.676	0.643	0.434	0.157
Wiener3D	0.577	0.560	0.539	0.363	0.045	0.510	0.500	0.488	0.391	0.041	0.590	0.581	0.568	0.437	0.071
WRSTF [5]	0.897	0.839	0.790	NA	NA	0.953	0.922	0.889	NA	NA	NA	NA	NA	NA	NA
SEQWT [10]	0.842	0.772	0.716	NA	NA	0.941	0.893	0.842	NA	NA	NA	NA	NA	NA	NA
3DWTF [6]	0.856	0.793	0.740	NA	NA	0.909	0.872	0.840	NA	NA	NA	NA	NA	NA	NA
IFSM [9]	0.855	0.776	0.709	0.485	0.458	0.927	0.882	0.837	0.623	0.344	0.884	0.813	0.749	0.510	0.370
3DSWDCT [7]	0.894	0.834	0.790	0.620	**0.513**	0.959	0.931	0.900	0.688	**0.396**	0.911	0.851	0.801	0.595	**0.398**
VBM3D [14]	**0.901**	**0.847**	**0.800**	0.640	0.478	**0.962**	**0.940**	**0.916**	0.738	0.336	**0.923**	**0.874**	**0.822**	**0.601**	**0.398**
2DGSM [18]	0.831	0.758	0.711	0.577	0.456	0.939	0.899	0.857	0.611	0.296	0.871	0.798	0.744	0.552	0.346
ST-GSM	0.894	0.841	0.797	**0.642**	0.464	0.950	0.925	0.900	**0.747**	0.363	0.913	0.865	0.820	0.597	0.364

example, when denoising the third frame, only 2 past frames and 4 future frames are involved. Similar strategy is employed in denoising the last frames.

IV. Noise-Robust Motion Estimation

One of the challenges in the implementation of the above algorithm is to estimate motion in the presence of noise. Here, we propose a simple but reliable noise-robust CC method for global motion estimation at integer pixel precision. The limitation of using global motion estimation is that it cannot account for rotation, zooming and local motion. However, local motion compensation processes (such as those based on block matching and optical flow) are spatially adaptive nonlinear operators, which significantly change noise statistics and affect the adequacy of the Gaussain noise model assumed in GSM denoising. As a result, the BLS-GSM denoising estimator becomes invalid. Therefore, we restrict our motion estimation to be global in our current implementation of ST-GSM.

Let $f_1(\mathbf{v})$ and $f_2(\mathbf{v})$ represent two image frames, where $\mathbf{v}$ is a spatial integer index vector for the underlying 2-D rectangular image lattice. A classical approach to estimating a global motion vector between the two frames is the cross correlation method [34]–[37], which is based on the observation that when $f_2(\mathbf{v})$ is a shifted version of $f_1(\mathbf{v})$, the position of the peak in the CC function between $f_1(\mathbf{v})$ and $f_2(\mathbf{u})$ corresponds to the motion vector. Despite the simplicity of the idea, the computation of the CC function is often costly. An equivalent but more efficient approach is to use the Fourier transform method: Let $F(\boldsymbol{\omega}) = \mathscr{F}\{f(\mathbf{v})\}$ represents the 2-D Fourier transform of an image frame, where $\mathscr{F}$ denotes the Fourier transform operator. Then, the CC function can be computed as

$$k_{cc}(\mathbf{v}) = \mathscr{F}^{-1}\left\{Y(\boldsymbol{\omega})\right\} \tag{7}$$

where $Y(\boldsymbol{\omega}) = F_1(\boldsymbol{\omega})F_2^*(\boldsymbol{\omega})$. The estimated motion vector is given by

$$\mathbf{v}_{opt} = \underset{\mathbf{v}}{\operatorname{argmax}}\, k_{cc}(\mathbf{v})\,. \tag{8}$$

An interesting variation of this approach is the phase correlation (PC) method [42]–[44], where the Fourier spectrum is normalized in the frequency domain to have unit energy across all frequencies. The phase correlation function is given by

$$k_{pc}(\mathbf{v}) = \mathscr{F}^{-1}\left\{\frac{Y(\boldsymbol{\omega})}{|Y(\boldsymbol{\omega})|}\right\}. \tag{9}$$

To have a close look, let us assume that $f_2(\mathbf{v})$ is simply a shifted version of $f_1(\mathbf{v})$, i.e., $f_2(\mathbf{v}) = f_1(\mathbf{v} - \Delta\mathbf{v})$. Based on the shifting property of the Fourier transform, we have $F_2(\boldsymbol{\omega}) = F_1(\boldsymbol{\omega})\exp\{-j\boldsymbol{\omega}^T\Delta\mathbf{v}\}$ and $Y(\boldsymbol{\omega}) = |F_1(\boldsymbol{\omega})|^2\exp\{j\boldsymbol{\omega}^T\Delta\mathbf{v}\}$, and thus

$$k_{pc}(\mathbf{v}) = \mathscr{F}^{-1}\left\{\exp\{j\boldsymbol{\omega}^T\Delta\mathbf{u}\}\right\} = \delta(\mathbf{v} + \Delta\mathbf{v}) \tag{10}$$

which creates an impulse at the true motion vector position and is zero everywhere else.

TABLE II

Average PSNR and SSIM [47] Performance of Video Denoising Algorithms at 5 Noise Levels

Noise std (σ)	10	15	20	50	100
PSNR Results (dB)					
Wiener2D	30.92	28.62	26.98	21.02	14.53
Wiener3D	26.55	26.20	25.78	21.97	14.24
IFSM [9]	33.26	30.90	29.28	24.60	20.22
3DSWDCT [7]	35.00	33.08	31.71	26.75	21.02
VBM3D [14]	36.05	34.26	32.94	27.52	20.53
2DGSM [18]	33.49	31.44	30.05	24.92	19.76
ST-GSM	35.50	33.56	32.21	26.33	20.30
SSIM Results					
Wiener2D	0.828	0.750	0.677	0.352	0.116
Wiener3D	0.715	0.694	0.666	0.414	0.061
IFSM [9]	0.893	0.836	0.784	0.612	0.522
3DSWDCT [7]	0.933	0.897	0.865	0.716	0.576
VBM3D [14]	0.938	0.910	0.883	0.749	0.540
2DGSM [18]	0.900	0.855	0.820	0.668	0.507
ST-GSM	0.934	0.905	0.880	0.738	0.538

Both CC and PC methods were designed with the assumption that there is no noise in the images. With the presence of noise, their performance degrades. Our noise-robust cross correlation (NRCC) function is defined as

$$k_{\mathrm{nrcc}}(\mathbf{v}) = \mathscr{F}^{-1}\left\{Y(\boldsymbol{\omega})\left(1 - \frac{|N(\boldsymbol{\omega})|^2}{|Y(\boldsymbol{\omega})|}\right)\right\} \tag{11}$$

where $|N(\boldsymbol{\omega})|^2$ is the noise power spectrum (in the case of white noise, $|N(\boldsymbol{\omega})|^2$ is a constant). To better understand this, it is useful to formulate the three approaches (PC, CC, and NRCC) using a unified framework. In particular, each method can be viewed a specific weighting scheme in the Fourier domain

$$k(\mathbf{v}) = \mathscr{F}^{-1}\left\{W(\boldsymbol{\omega})\exp\{j\boldsymbol{\omega}^T\Delta\mathbf{v}\}\right\} \tag{12}$$

where the differences lie in the definition of the weighting function $W(\boldsymbol{\omega})$

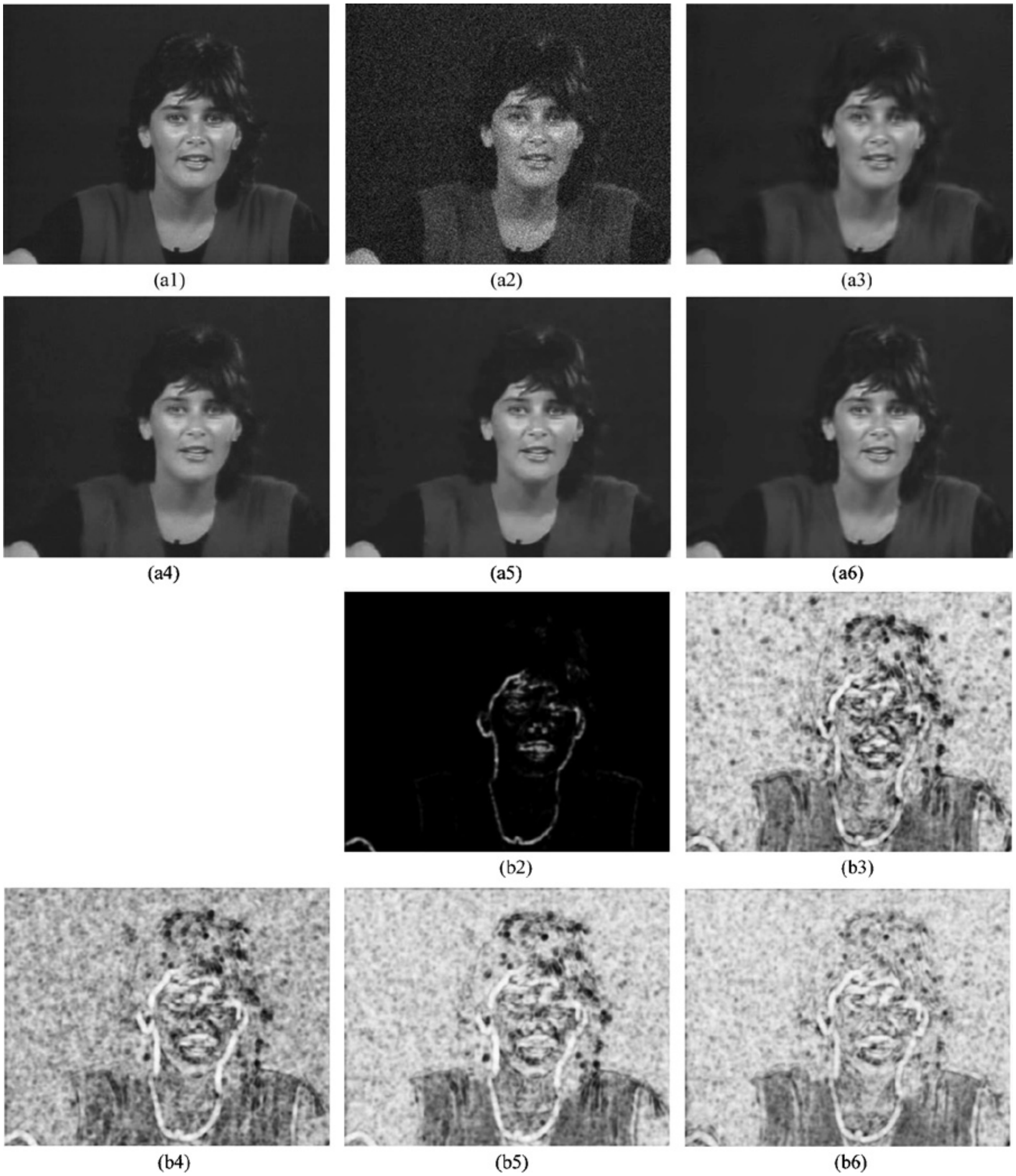
$$\begin{aligned} W_{pc}(\boldsymbol{\omega}) &\equiv 1 \\ W_{cc}(\boldsymbol{\omega}) &= |F_1(\boldsymbol{\omega})|^2 \\ W_{nrcc}(\boldsymbol{\omega}) &= |F_1(\boldsymbol{\omega})|^2 - |N(\boldsymbol{\omega})|^2. \end{aligned} \tag{13}$$

The PC method assigns uniform weights to all frequencies, the CC method assigns the weights based on the total signal power (which is the sum of signal and noise power), while the NRCC method assigns the weights proportional to the signal power only (by removing the noise power part). It converges to the CC method when the images are noise-free. We tested the PC, CC, and NRCC methods at different noise levels and use root mean squared (RMS) error between the true shift and estimated shift to evaluate their performance. Fig. 3 shows our test results by estimating the shift between a 1-D signal (extracted from one row of the "Einstein" image) and a shifted version of it using the three methods. It appears that the CC and NRCC methods perform much better at all noise levels, and NRCC leads to the best performance.

TABLE III

PSNR COMPARISONS WITH LATEST VIDEO DENOISING ALGORITHMS

Sequence	Size	Input PSNR	STA [15]	K-SVD [16]	3-D-Patch [17]	ST-GSM
Salesman	176 × 144 × 449	28	35.13	37.91	**39.26**	37.93
		24	32.60	35.59	**36.35**	35.17
Miss America	176 × 144 × 108	28	39.39	40.49	**42.23**	41.43
Suzie	176 × 144 × 150	28	37.07	37.96	**38.40**	38.36
		24	35.11	35.95	**36.32**	36.21
Trevor	176 × 144 × 90	28	36.68	38.10	**38.31**	37.17
		24	34.79	**35.97**	35.81	34.68
Foreman	176 × 144 × 300	28	34.94	**37.86**	36.88	36.85
		24	32.90	**35.86**	34.55	34.37


(a1) (a2) (a3)

(a4) (a5) (a6)

(b2) (b3)

(b4) (b5) (b6)

Fig. 5. Denoising results of Frame 80 in *Miss America* sequence corrupted with noise standard deviation σ = 20. (a1)–(a6) Image frames in the original, noisy, and 2DGSM [18], IFSM [9], VBM3D [14], and ST-GSM denoised sequences. (b2)–(b6) Corresponding SSIM quality maps (brighter indicates larger SSIM value) with mean SSIM values.

V. Validation

The video denoising algorithms were tested using publicly available video sequences [45] contaminated with additive white Gaussian noise. All video sequences are in YCrCb 4:2:0 format, but only the denoising results of the luma channel are reported here for algorithm validation. Two objective criteria, namely the PSNR and the SSIM [46]–[48], were employed to provide quantitative quality evaluations of the denoising results. Specifically, PSNR is defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{L^2}{\text{MSE}} \right) \tag{14}$$

where L is the dynamic range of the image (for 8 bits/pixel images, $L = 255$) and MSE is the mean squared error between the original and distorted images. SSIM is first calculated within local windows using

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{15}$$

where **x** and **y** are the image patches extracted from the local window from the original and distorted images, respectively. μ_x, σ_x^2, and σ_{xy} are the mean, variance, and cross-correlation computed within the local window, respectively. The overall SSIM score of a video frame is computed as the average local SSIM scores. PSNR is the mostly widely used quality measure in the literature, but has been criticized for not correlating well with human visual perception [49]. SSIM is believed to be a better indicator of perceived image quality [49]. It also supplies a quality map that indicates the variations of images quality over space. The final PSNR and SSIM results for a denoised video sequence are computed as the frame average of the full sequence, after clipping the denoised pixels to the range of 0–255.

Three experiments were carried out to validate various aspects of the proposed ST-GSM algorithm. In the first experiment, we verify the usefulness of including temporal wavelet neighbors in forming the coefficient vector in GSM denoising. In particular, we compare the proposed algorithm against GSM denoising applied to each individual video frame independently (abbreviated as 2DGSM [18]). The results are shown in Table I, where ST-GSM performs consistently better for all test sequences at all noise levels.

Second, the effectiveness of motion estimation/motion compensation in the proposed algorithm is tested. Fig. 4 shows the results of a straightforward comparison, where the same ST-GSM algorithm was applied to denoise the same video sequence, but with and without the involvement of the motion estimation/motion compensation stages. The PSNR and SSIM results computed on a frame-by-frame basis show that motion estimation/compensation leads to markable improvement. Similar results were observed when denoising other video sequences, especially those video segments that involve significant amount of global motion.

In the third experiment, we compare ST-GSM against other video denoising algorithms, including both baseline and state-of-the-art schemes. These include: 1) Wiener2D—MATLAB's *Wiener2* function (a spatially adaptive Wiener filter) applied on a frame-by-frame basis; 2) Wiener3D—our implementation of a simple extension of *Wiener2* to three dimensions, with a window size of 3 × 3 × 3; 3) WRSTF [5]—wavelet-domain reliability-based spatio-temporal filtering; 4) SEQWT [10]—sequential wavelet domain temporal filtering; 5) 3DWTF [6]—3-D dual tree wavelet transform denoising; 6) IFSM [9]—inter-frame statistical modeling; 7) 3DSWDCT [7]—3-D sliding window discrete cosine transform; 8) VBM3D [14]—block matching and 3-D filtering; 9) 2DGSM [18]—as in the first experiment described above; 10) STA [15]—space-time adaptive patch-based filtering; 11) K-SVD [16]—video denoising based on sparse and redundant representation; and 12) 3-D-patch [17] —Bayesian variational 3-D patch-based method. The results for WRSTF, SEQWT and 3DWTF were obtained from the processed video sequences available at [50]. The PSNR results of STA, K-SVD and 3-D-patch denoising are directed cited from the corresponding publications [15]–[17]. We performed the denoising experiments for the rest of the methods.

Table I compares PSNR and SSIM results of 10 denoising algorithms for six video sequences at five noise levels. Table II shows PSNR and SSIM performance averaged over sequences. Table III presents PSNR comparisons with STA, K-SVD, and 3-D-patch methods. It can be observed that the proposed ST-GSM algorithm demonstrates competitive performance when compared with the state-of-the-art. Finally, Fig. 5 demonstrates the visual effects of different denoising algorithms. Specifically, we show a frame extracted from the *Miss America* sequence, together with a noisy version of the same frame, and the denoised frames obtained by four video denoising algorithms. It can be seen that ST-GSM is quite effective at suppressing background noise while maintaining the edge and texture details and thus, the structural information of the objects in the scene. This is further verified by examining the SSIM quality maps of the corresponding frames.

VI. Conclusion

We proposed a wavelet-domain ST-GSM model for natural video signals and applied it to the restoration of video signals corrupted by additive white Gaussian noise. We found that applying motion estimation/motion compensation is effective in enhancing the correlations between temporal neighboring wavelet coefficients and thus, improving the performance of ST-GSM denoising. We proposed a Fourier domain NRCC scheme to provide reliable motion estimation in the presence of noise. Our experimental comparisons with state-of-the-art algorithms showed that ST-GSM is competitive in terms of both subjective and objective (PSNR and SSIM) evaluations.

There are a number of potential improvements and extensions that may be done in the future. First, the current implementation of ST-GSM denoising is rather slow. A rough estimate for the computational complexity can be inferred from the observation that our un-optimized MATLAB code took 120 s per CIF frame on an Intel 2.4 GHz workstation. Both algorithmic and software optimizations are needed to accelerate the algorithm. Second, the current algorithm is applied to each color channel independently. Denoising all color channels jointly by including color wavelet coefficient neighbors in GSM modeling may further improve the algorithm. Finally, there are advanced models in describing the statistical motion properties of natural video signals (e.g., [24]) that may be employed to impose stronger statistical prior in a Bayesian framework and in turn improve the denoising performance.

ACKNOWLEDGMENT

The authors would like to thank S. M. M. Rahman for providing the code for IFSM algorithm, and the anonymous reviewers for their valuable comments that greatly helped us improve this paper.

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, no. 9, pp. 1272–1292, Sep. 1995.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, May 2005.

[3] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1. Sep. 1996, pp. 379–382.

[4] F. Jin, P. Fieguth, and L. Winger, "Wavelet video denoising with regularized multiresolution motion estimation," *Eup. Assoc. Speech, Signal, Image Process. J. Appl. Singal Process.*, vol. 2006, no. 72705, pp. 1–11, 2006.

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

[6] W. I. Selesnick and K. Y. Li, "Video denoising using 2-D and 3-D dualtree complex wavelet transforms," in *Proc. SPIE, Wavelets: Appl. Signal Image Process. X*, vol. 5207. Nov. 2003, pp. 607–618.

[7] D. Rusanovskyy and K. Egiazarian, "Video denoising algorithm in sliding 3-D DCT domain," in *Proc. ACIVS*, Sep. 2005, pp. 618–625.

[8] N. Lian, V. Zagorodnov, and Y. Tan, "Video denoising using vector estimation of wavelet coefficients," in *Proc. IEEE Int. Sym. Circuits Syst.*, May 2006, pp. 2673–2676.

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

[10] A. Pizurica, V. Zlokolica, and W. Philips, "Combined wavelet domain and temporal video denoising," in *Proc. IEEE Conf. Adv. Video Signal-Based Surveillance*, Jul. 2003, pp. 334–341.

[11] E. J. Balster, Y. F. Zheng, and R. L. Ewing, "Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 16, no. 2, pp. 220–230, Feb. 2006.

[12] A. Buades, B. Coll, and J. Morel, "Denoising image sequences does not require motion estimation," in *Proc. IEEE Conf. Adv. Video Signal-Based Surveillance*, Sep. 2005, pp. 70–74.

[13] A. Buades, B. Coll, and J. Morel, "Nonlocal image and movie denoising," *Int. J. Comput. Vision*, vol. 76, no. 2, pp. 123–139, 2008.

[14] K. Dabov, A. Foi, and K. Egiazarian, "Video denoising by sparse 3-D transform-domain collaborative filtering," in *Proc. Eur. Signal Process. Conf.*, Sep. 2007, pp. 145–149.

[15] J. Boulanger, C. Kervrann, and P. Bouthemy, "Space-time adaptation for patch-based image sequence restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1096–1102, Jun. 2007.

[16] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–36, Jan. 2009.

[17] X. Li and Y. Zheng, "Patch-based video processing: A variational Bayesian approach," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 19, no. 1, pp. 27–40, Jan. 2009.

[18] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[19] D. L. Ruderman, "The statistics of natural images," *Netw.: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1996.

[20] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annua. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.

[21] D. W. Dong and J. J. Atick, "Statistics of natural time-varying images," *Network: Comput. Neural Syst.*, vol. 6, no. 3, pp. 345–358, 1995.

[22] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," in *Proc. R. Soc. Lond. B*, vol. 265. 1998, pp. 2315–2320.

[23] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1. Sep. 2003, pp. 41–44.

[24] Z. Wang and Q. Li, "Statistics of natural image sequences: Temporal motion smoothness by local phase correlations," in *Proc. Human Vision Electron. Imag. IX, SPIE*, vol. 7240. Jan. 2009, pp. 72400W-1–72400W-12.

[25] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[26] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1056–1068, Jul. 2001.

[27] D. Andrews and C. Mallows, "Scale mixtures normal distributions," *J. R. Statist. Soc.*, vol. 36, p. 99, 1974.

[28] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," *Adv. Neural Inform. Process. Syst.*, vol. 12, pp. 855–861, 2000.

[29] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am. A*, vol. 24, pp. B61–B69, Dec. 2007.

[30] S. Zhu and K.-K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.

[31] B. K. P. Horn and B. G. Rhunck, "Determining optical flow," *Artificial Intell.*, vol. 17, pp. 185–203, Apr. 1981.

[32] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vision*, vol. 12, pp. 1573–1405, Feb. 1994.

[33] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vision*, vol. 5, no. 1, pp. 77–104, 1990.

[34] T. S. Huang and R. Y. Tsai, "Motion estimation," in *Image Sequence Analysis*, T. S. Huang, Ed. Berlin, Germany: Springer-Verlag, 1981, pp. 1–18.

[35] R. Manduchi and G. A. Mian, "Accuracy analysis for correlation-based image registration algorithms," in *Proc. IEEE Int. Sym. Circuits Syst.*, 1993, pp. 834–837.

[36] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ: Prentice Hall, 1995.

[37] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 188–200, Mar. 2002.

[38] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Reading, MA: Wiley-Interscience, 1992.

[39] M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.

[40] S. Mallat, *A Wavelet Tour of Signal Processing*. Cambridge, U.K.: Academic Press, 1999.

[41] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[42] D. Robinson and P. Milanfar, "Fundamental performance limits in image registration," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1185–1199, Sep. 2004.

[43] S. Alliney and C. Morandi, "Digital image registration using projections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 2, pp. 222–233, Mar. 1986.

[44] W. K. Pratt, "Correlation techniques for image registration," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 10, no. 3, pp. 353–358, May 1974.

[45] *Video Sequence Database* [Online]. Available: http://www.cipr.rpi.edu/resource/sequences

[46] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[48] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.* (special issue on objective video quality metrics), vol. 19, pp. 121–132, Feb. 2004.

[49] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[50] V. Zlokolica. Results of "Wavelet-domain video denoising based on reliability measures." *Universiteit Gent: Department of Telecommunications and Information Processing* [Online]. Available: http://telin.ugent.be/∼vzlokoli/Results_J

Gijesh Varghese (S'05–M'07) received the B. Tech. degree in electrical engineering from the National Institute of Technology, Calicut, India, in 2003, and the M.S. degree in electrical engineering from the University of Texas, Arlington, in 2007.

He is currently a Video Algorithm Engineer with Maxim Integrated Products, Sunnyvale, CA, where he works on video compression, and post- and preprocessing algorithms. His current research interests include video signal processing and compression.

Zhou Wang (S'97–M'02) received the Ph.D. degree in electrical and computer engineering from the University of Texas, Austin, in 2001.

He was an Assistant Professor with the University of Texas, Arlington, a Research Associate with the Howard Hughes Medical Institute, Chevy Chase, MD, and New York University, New York, and a Research Engineer with AutoQuant Imaging, Inc., Watervliet, New York. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada. He has more than 80 publications and one U.S. patent in these fields, and is an author of *Modern Image Quality Assessment* (Morgan & Claypool, 2006). His current research interests include image processing, coding, quality assessment, computational vision and pattern analysis, multimedia communications, and biomedical signal processing.

Dr. Wang is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and *Pattern Recognition*, and a Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and *EURASIP Journal on Image and Video Processing*. He was a recipient of the 2009 IEEE Signal Processing Society Best Paper Award and 2009 Ontario Early Researcher Award.