A Real-Time Technique for Spatio-Temporal Video Noise Estimation

Mohammed Ghazal, *Student Member, IEEE,* Aishy Amer, *Member, IEEE,* and Ali Ghrayeb, *Senior Member, IEEE*

***Abstract*— This paper proposes a spatio-temporal technique for estimating the noise variance in noisy video signals, where the noise is assumed to be additive white Gaussian noise. The proposed technique utilizes domain-wise (spatial, temporal and spatio-temporal) video information independently for improved reliability. It divides the video signal into cubes and measures their homogeneity using Laplacian of Gaussian based operators. Then, the variances of homogeneous cubes are selected to estimate the noise variance. A least median of squares robust estimator is used to reject outliers and produce domain-wise noise variance estimates which are adaptively integrated to obtain the final frame-wise estimate. The proposed technique estimates the noise variance reliably in video sequences with both low and high video activities (e.g., fast motion or high spatial structure) and it produces a maximum estimation error of 1.7 dB PSNR. The proposed method is fast when compared to referenced methods.**

***Index Terms*— Homogeneity measurement, robust estimator, spatio-temporal estimation, video noise, white noise.**

I. Introduction

The need for fast and accurate video noise estimation algorithms rises from the fact that many fundamental video processing algorithms such as compression, segmentation, motion estimation and format conversion adapt their parameters and improve performance when the noise is known. For example, in segmentation [1], the binarization process can be adapted to the noise level for more stable object masks. Also, accurate information about the noise level is vital to robust motion estimation [2] to tune the outliers rejection phase for more reliable motion vectors or parameters.

Noise refers to unwanted stochastic variations as opposed to deterministic distortions such as shading or lack of focus. It can be added to the video signal or multiplied with the video signal. It can also be signal dependent or signal independent. Based on its spectral properties, noise is further classified as white or color noise. Many types of noise affect CCD cameras such as photon shot noise and read out noise [3]. Photon shot noise is due to the random arrival of photons at the sensor which is governed by Poisson distribution. Other sources of noise include output amplifier noise, camera noise and clock noise which can be combined in a single equivalent Gaussian noise source called read out noise. Because of the high counting effect of Photon arrivals and according to the central limit theorem, the aggregate noise effect can be well approximated by Gaussian distribution [4]. Consequently, in this paper, an additive white Gaussian noise (AWGN) model is assumed. The choice is also motivated by AWGN being the most common noise model for terrestrial TV broadcasting [5].

Algorithms for estimating the AWGN variance are either temporal [6], [7], spatial [8]–[14] or spatio-temporal approaches [15], [16]. There exist few methods for purely temporal or spatio-temporal noise estimation such as [6], [7], [15], [16]. These methods are challenged by the presence of object or global motion. Motion detection or motion compensation are commonly used as countermeasures. Hence, methods in this area such as [6], [7] tend to be computationally expensive. The method in [15] attempts to utilize temporal adaptation to stabilize the spatially estimated noise variance. The method in [16] uses spatio-temporal gradients to perform noise estimation.

Many methods for purely spatial noise estimation have been presented [8]–[14]. Difficulties with these methods rise from frames with very high or very low noise levels as well as highly structured frames. The difficulty lies in determining whether intensity variations are due to noise or to frame details. Spatial methods can be further categorized into smoothing-based, wavelet-based and block-based methods. Smoothing-based algorithms such as [8], [9] estimate noise from the difference between the noisy and smoothed frames. The assumption is that this difference represents an approximation of the noise. These approaches are computationally expensive and tend to overestimate the noise variance. Moreover, the algorithm in [9] depends on many parameters such as the number of process iterations and the shape of the fade-out cosine function to evaluate the variance histogram.

Wavelet-based methods were presented in [10], [11], [14], [16]. They utilize the wavelet transform to isolate the noise component and estimate the noise variance. Methods that use the wavelet transform are similar to smoothing-based methods in overestimating the noise variance, however, recent wavelet-based methods (e.g., [11], [14], [16]) produce much better results than smoothing based methods (e.g., [8], [9]) mostly because the wavelet transform can better isolate various sub-bands of the signal and provide a much sparser representation than the two sub-bands (high and low) associated with smoothing based methods. Wavelet-based noise estimation techniques have the advantage of efficient and easy integration into wavelet denoising and compression frameworks. In non-

The authors are with the Electrical and Computer Engineering Department, Concordia University, Montréal, Quebéc, Canada. {E-mails:{moha_mo,amer,aghrayeb}@ece.concordia.ca}.

This work was supported, in part, by the *Fonds de la recherche sur la nature et les technologies du Quebéc* (NATEQ) under grant numbers F00365 and F00361.

This work was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, 2006.

wavelet based applications in which noise estimation is key to improved performance, especially ones which operate in real-time, the wavelet transform for noise estimation is considered an overhead.

Block-based methods in [12] and [13] are less computationally demanding. These methods attempt to locate regions with the least amount of signal variations where any intensity variations is assumed to be due to noise. The algorithm in [12] uses the variance to measure block homogeneity. The problem with this approach is that the variance is not always a reliable measure of homogeneity. The algorithm in [13] proposes a homogeneity test in which a number of high-pass operators are applied directionally. The variance of the noise is estimated from the local variances of blocks selected to be the most homogeneous. The algorithm in [13], however, does not exploit the temporal information present in the video signal.

The proposed method attempts to estimate the global (frame-wise) variance of the noise from the local (domain-wise) variances of selected cubes in the video signal. The selected cubes have the common characteristic of being intensity homogeneous in the 2D or 3D space. Cube inhomogeneity is due to fine details and structures in the spatial domain, motion in the temporal domain or noise. The proposed algorithm starts by dividing the 3D space defined by the video signal into cubic subspaces in an interpretation different from the one in [15] which treats the video signal as a sequence of 2D images.

The contributions of the proposed methods are 1) considering domain-wise (spatial, temporal, and spatio-temporal) estimations independently for improved estimation reliability, 2) proposing Laplacian of Gaussian (LoG) based operators for local homogeneity measurement, 3) utilizing a least median of squares (LMS) robust estimator to reject outliers, and 4) integrating domain-wise estimates adaptively to produce the frame-wise estimate. The proposed method is an extension of the method in [17], but in this paper, we propose new methods to improve the performance. First, we use LoG-based local homogeneity analysis instead of Laplacian-based to reduce the effect of noise. Second, we use an LMS robust estimator to reject outliers and consider multiple observations as opposed to the simple median which considers only one observation. Finally, we consider the possibility of an estimation failure in a whole domain by measuring the domain reliability and using it to adaptively integrate the domain-wise estimates to produce the frame-wise estimate.

The remainder of the paper is as follows. In section II, we present the proposed approach theoretically and give an interpretation of its good performance. In section III, we propose an LMS robust estimator to reject outliers that pass the homogeneity test. We discuss objective simulation results in section IV. Finally, we conclude with section V.

II. Proposed Algorithm

The main steps of the proposed algorithm are: 1) divide the signal into cubes and profile the local homogeneity in each domain independently using LoG-based operators (section II-A), 2) select the most homogeneous cubes in each domain separately and calculate their local statistics (mean and variance) along homogeneous plains only to get the set of variance estimates in each domain (section II-B), 3) apply an LMS robust estimator to reject outliers and produce the domain-wise estimates, and 4) adaptively integrate domain-wise estimates to get the final frame-wise noise variance estimate while rejecting domains that are unsuitable candidates for estimation (due to high motion or complex spatial structures) (section III).

A. Local Homogeneity Measurement

Let V_η denote a noisy digital video signal defined by

$$V_\eta = V + \eta, \tag{1}$$

where V is the noise-free uncorrelated video signal and η is the added noise component. A pixel in V_η is denoted by $V_\eta(i,j,n)$ where i and j are the spatial coordinates and n is the temporal coordinate. $\eta(i,j,n)$ is the amount of noise added to $V(i,j,n)$. Since the proposed algorithm is designed to be context-free, there are no restrictions on the original signal V. The division of V_η into cubes C_{ijn} with spatial indexes i and j and temporal index n is done using

$$\begin{aligned} C_{ijn} &= \{V_\eta(i,j,n) | (i,j,n) \in \Psi_{ijn}\}; \\ \Psi_{ijn} &= \{(i,j,n) | i - \tfrac{W-1}{2} \le i \le i + \tfrac{W-1}{2}, \\ &\qquad j - \tfrac{W-1}{2} \le j \le j + \tfrac{W-1}{2}, \\ &\qquad n - \tfrac{W-1}{2} \le n \le n + \tfrac{W-1}{2}\} \end{aligned}, \tag{2}$$

where Ψ_{ijn} is a set of indexes (locations) making a cube of size W^3 centered around the 3D point $(i,j,n) \in V_\eta$ and C_{ijn} is the set of pixel values at those locations. A single spatial plain in Ψ_{ijn} is denoted $\Psi_{ij\rho} \subset \Psi_{ijn}$ which can be isolated using

$$\begin{aligned} \Psi_{ij\rho} = \{(i,j,n) | & i - \tfrac{W-1}{2} \le i \le i + \tfrac{W-1}{2}, \\ & j - \tfrac{W-1}{2} \le j \le j + \tfrac{W-1}{2}, \\ & n = \rho\} \end{aligned}. \tag{3}$$

For example, if $\rho = n$, $\Psi_{ij\rho}$ will represent coordinates to pixels in the middle spatial plain of the cube (non-zero plain in Fig. 2 (c)).

To locate the homogeneous cubes in the video signal, we define a set $\{\zeta_D\}$ of low-complexity homogeneity measures with (4), where D is the domain index. We consider five domains, the spatio-temporal ST, the purely temporal T, the purely spatial S, the spatially vertical and temporal VT and the spatially horizontal and temporal HT. These measures represent the quantities in (5)-(9).

$$\{\zeta_D\}, D \in \{ST, T, S, VT, HT\} \tag{4}$$

$$\zeta_{ST} = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial j^2} + \frac{\partial^2 V_\eta}{\partial n^2}|; \tag{5}$$

$$\zeta_T = |\frac{\partial^2 V_\eta}{\partial n^2}|; \tag{6}$$

$$\zeta_S = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial j^2}|; \tag{7}$$

$$\zeta_{VT} = |\frac{\partial^2 V_\eta}{\partial j^2} + \frac{\partial^2 V_\eta}{\partial n^2}|; \quad (8)$$

$$\zeta_{HT} = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial n^2}|. \quad (9)$$

The proposed homogeneity measures are the magnitudes of dimensional Laplacian (second-order) operators. We use second-order and not first-order operators because of their 1) rotational invariance which accounts for unpredictable object shapes or movements, and 2) sensitivity to fine details and structures. To calculate the measures, masks can be applied to pixels of cubes C_{ijn} in various domains. For example, the second order (Laplacian) masks in Fig. 1(a) respond to change in different orientations in the spatial domain. More precisely they are invariant to 45^o rotations.

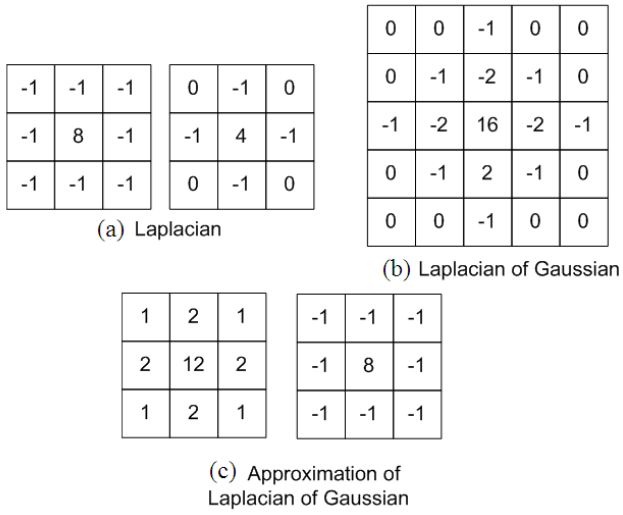


Fig. 1. Different discrete structure-detecting masks. The masks in (a) approximate the Laplacian with $W = 3$. The masks in (b) approximate the LoG with $W = 5$. The masks in (c) approximate the LoG with $W = 3$ broken into Gaussian and Laplacian operations.

On the other hand, the disadvantages of using the Laplacian are threefold: 1) sensitivity to noise, 2) double edges, and 3) inability to determine edge directions. Since our objective is to locate homogeneous areas, edge direction is of no interest. Also, producing double edges is not significant as no segmentation is needed. The most important problem facing the proposed homogeneity profile is the Laplacian's sensitivity to noise. To overcome such sensitivity, we use approximations of the LoG filter which is the result of convolving a Gaussian smoothing filter with the Laplacian filter and has the continuous-time impulse response

$$h(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right]\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right). \quad (10)$$

The impulse response $h(x,y)$ is sampled to produce the discrete-time mask in Fig. 1(b), which combines both Gaussian smoothing and Laplacian structure detection and has less sensitivity to noise than the Laplacian alone. A window size of $W \geq 5$ is needed to have enough samples of $h(x,y)$ to observe the effect of both Gaussian smoothing and Laplacian structure detection. If the mask is smaller than $W = 5$, there is not enough samples to observe the effect of Gaussian smoothing and the Laplacian and LoG filters will have the same effect. The drawback of using the mask in Fig. 1(b) is that extending it to the third temporal dimension requires five frame delays which is not practical in real-time and online video processing systems.

As an alternative approach, we propose to use the separability of the LoG and apply first the Gaussian mask with $W = 3$ (to the left of Fig. 1(c)) followed by the Laplacian mask with $W = 3$ (to the right of Fig. 1(c)) to approximate the Laplacian of Gaussian with $W = 3$ in the spatial domain. The same procedure applies for other domains too. After applying the Gaussian mask to reduce the noise, we then proceed to apply each of the masks in Fig. 2 to obtain the quantities in (5)-(9). Fig. 2(a) is a 3D Laplacian operator used to measure spatio-temporal homogeneity or ζ_{ST} in (5). The central coefficient of the mask accumulates to $W^3 - 1$ as a result of combining the second derivatives in all directions. The mask in Fig. 2(b) evaluates homogeneity along the temporal direction or ζ_T in (6). It acts as a local low-complexity motion detector. The mask in Fig. 2(c) is the spatial domain Laplacian operator. It measures purely spatial homogeneity or ζ_S defined in (7). This mask's response is an approximation of the sum of directional responses of the masks defined in [13]. The mask in Fig. 2(d) measures both the homogeneity along the spatial vertical direction and the temporal direction or ζ_{VT} in (8). Similarly, the mask in Fig. 2(e) measures the homogeneity along the spatial-horizontal and the temporal directions or ζ_{HT} in (9).

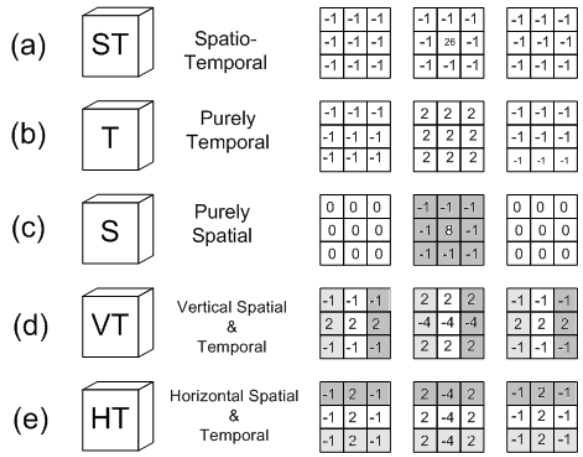


Fig. 2. Homogeneity analyzer cubical masks where pixels in the same gray-level belong to one plain.

To study the effect of using the proposed combination of the Gaussian mask followed by the Laplacian masks as opposed to using the Laplacians in Fig. 2 alone, an expression of the Laplacian's sensitivity to noise is developed using Fig. 3 which shows a spatial 3×3 window of the original signal. Let $\{X(z)\}, z \in \{1, 2, ..., 8\}$ be a set of random variables denoting pixels in the eight-neighborhood of pixel $X(0) \in V$ (see (1)), $\{N(z)\}$ are the random variables corresponding to the noise signal η and $\{G(z)\}$ are the random variables in the noisy video signal V_η. The response Y of applying the spatial

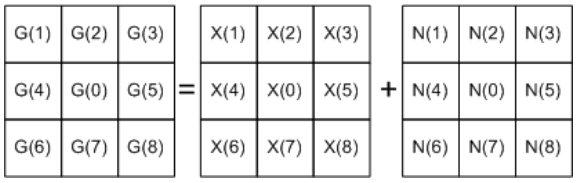


Fig. 3. Evaluation of sensitivity to noise in Laplacian operators.

Laplacian mask in Fig. 2(c) can be calculated with

$$Y = 8G(0) - \sum_{z=1}^{8} G(z) \tag{11}$$

$$= 8(X(0) + N(0)) - \sum_{z=1}^{8}(X(z) + N(z)) \tag{12}$$

$$= \left[8X(0) - \sum_{z=1}^{8} X(z)\right] + \sum_{z=1}^{8} N(z) + 8N(0). \tag{13}$$

Since the noise model is AWGN, $\left[8X(0) - \sum_{z=1}^{8} X(z)\right]$ measures the homogeneity of the window as $\sum_{z=1}^{8} N(z) \approx 0$. We can see that the term $8N(0)$ can impair the Laplacian's response. By applying the Gaussian mask before the Laplacian mask, the effect of noise is reduced. At a later stage, robust estimation of the noise variance is used to exclude cubes that falsely pass the homogeneity test due to noise or other sources of outliers.

B. Homogeneous Cubes Selection

The proposed homogeneity profile for every cube C_{ijn} is the set $\{\zeta_D\}$ populated by applying the Gaussian mask in Fig. 1(c) followed by masks in Fig. 2 to the video signal cubes C_{ijn} (see (2)). We select the $L \in \mathbb{Z}^+$ most homogeneous cubes only based on each ζ_D for noise variance estimation. Formally, let U_D be the set of all selected homogeneous cubes based on ζ_D or

$$U_D = \left\{ C_{ijn} | \min_{ijn}^{L}(\zeta_D) \right\}. \tag{14}$$

Recall from (4) that $D \in \{ST, T, S, VT, HT\}$. Eq. (14) indicates that we are considering the set of the L most homogeneous cubes selected independently based on each ζ_D (i.e., ζ_S, ζ_T, ζ_{ST}, ζ_{HT} and ζ_{VT}). L was fixed to 10% of the total number of blocks in [8] and [13]. In the proposed algorithm, L is a monochromically decreasing positive function of the initial estimate $\text{PSNR}^{\eta}_{init}$ of the Peak Signal to Noise Ratio of V_η, and is computed as

$$L = L_{max} - \frac{\text{PSNR}^{\eta}_{init}}{\beta}, \tag{15}$$

where L_{max} is the maximum number of cubes to be used and β is a scaling factor. We control with L_{max} and β the mapping between the expected range of noise and the suitable number of cubes to consider (as observed by our simulations). $\text{PSNR}^{\eta}_{init}$ is calculated from the median of the variances of the three most homogeneous cubes over all ζ_D. We use L to include more cubes in case of noisy video sequences and less cubes in case of less noisy ones. Using more cubes in case of noisier video sequences is to account for the effect of noise on homogeneity analysis. Other positive decreasing functions can also be used as L. Note that homogeneity measures of a cube are not combined because a cube that is highly homogeneous temporally (low ζ_T) can be spatially non-homogeneous (high ζ_S).

After homogeneous cubes are selected, we calculate their sample mean and variance along the plains (pixels with the same gray level in Fig. 2) found to be most homogeneous. For all cubes in U_S, we use

$$\begin{aligned} \mu_S &= \frac{\sum_{(i,j,n)\in\Psi_{ij\rho|_{\rho=n}}} V_\eta(i,j,n)}{W^2}; \\ \sigma_S^2 &= \frac{\sum_{(i,j,n)\in\Psi_{ij\rho|_{\rho=n}}} (V_\eta(i,j,n)-\mu_S)^2}{W^2-1}, \end{aligned} \tag{16}$$

where $\Psi_{ij\rho|_{\rho=n}}$ indicates that we use only pixels along the middle spatial plain of the cube (pixels in the same gray level in Fig. 2(c)) (see (3)). The set of all local variances calculated spatially from cubes in U_S is denoted by $U_S^{\sigma^2}$. For cubes in U_{HT}, we use

$$\begin{aligned} \mu_{HT}(\rho) &= \frac{\sum_{(i,j,n)\in\Psi_{i\rho n}} V_\eta(i,j,n)}{W^2}; \\ \sigma^2_{HT}(\rho) &= \frac{\sum_{(i,j,n)\in\Psi_{i\rho n}} (V_\eta(i,j,n)-\mu_{HT}(\rho))^2}{W^2-1}, \end{aligned} \tag{17}$$

where $\Psi_{i\rho n}$ indicates that we use only pixels along horizontal and temporal plains only. By varying ρ between $j - \frac{W-1}{2}$ and $j + \frac{W-1}{2}$, we consider each time pixels in one plain only (pixels in the same gray level in Fig. 2(e)). This makes $\mu_{HT}(\rho)$ and $\sigma^2_{HT}(\rho)$ functions of ρ. We calculate their value for every plain ρ and consider the mean μ_{HT} and variance σ^2_{HT} of the cube as the average of the means and variances for all ρ. It is important that the noise variance is estimated using only plains found to be homogeneous as we have no information about the homogeneity along other plains. In the same manner, for cubes that are chosen to be spatio-temporally most homogeneous (i.e., U_{ST}), the sample mean and variance are calculated over all pixels in the cube using

$$\begin{aligned} \mu_{ST} &= \frac{\sum_{(i,j,n)\in\Psi_{ijn}} V_\eta(i,j,n)}{W^3}; \\ \sigma^2_{ST} &= \frac{\sum_{(i,j,n)\in\Psi_{ijn}} (V_\eta(i,j,n)-\mu_{ST})^2}{W^3-1}. \end{aligned} \tag{18}$$

The corresponding set of variances calculated spatio-temporally is denoted by $U^{\sigma 2}_{ST}$. The sets $U^{\sigma 2}_{T}$ and $U^{\sigma 2}_{VT}$ are calculated using the same procedure (considering only pixels in the homogeneous plains). In the next section, we propose a robust estimator to estimate the domain-based (i.e., spatial, temporal and spatio-temporal) noise variance from these sets of variances. This estimator aims at rejecting outliers that pass the homogeneity test.

III. Robust Estimation using the LMS

This section presents how the dimension-based (i.e., spatial, temporal and spatio-temporal) noise variances are robustly estimated from $U_D^{\sigma^2}$. Robustness is defined in [18] as the ability

to deal with the possible consequences of deviations from the assumed statistical model. In computer vision literature, the M-*Estimators* and the LMS are the most commonly used robust estimators [19]. In this paper, we utilize the LMS by defining the search range R around the initial variance estimate σ^2_{init} to be

$$R = \left[\sigma^2_{init} - \frac{\sigma^2_{th}}{2} : \frac{\sigma^2_{th}}{Q} : \sigma^2_{init} + \frac{\sigma^2_{th}}{2}\right]. \tag{19}$$

Let σ^2_p be a variance in R which assumes values between $\sigma^2_{init} - \sigma^2_{th}/2$ and $\sigma^2_{init} + \sigma^2_{th}/2$ in steps of $\frac{\sigma^2_{th}}{Q}$ as illustrated in Fig. 4. Q is the number of search steps which controls the accuracy versus the complexity. Larger Q means more computations but more accurate estimation and vice versa. σ^2_{th} is a threshold related to the maximum allowable estimation error in the application. The overall noise variance estimate $\widehat{\sigma}^2_D$ in domain D, is calculated using the LMS robust estimator from $U_D^{\sigma^2}$ as

$$\widehat{\sigma}^2_D = \underset{\sigma^2_p \in R}{\operatorname{argmin}} \underset{\sigma^2_{D\alpha} \in U_D^{\sigma^2}}{\operatorname{median}} |\sigma^2_p - \sigma^2_{D\alpha}|, \tag{20}$$

where $\sigma^2_{D_\alpha}$ is a variance in $U_D^{\sigma^2}$.

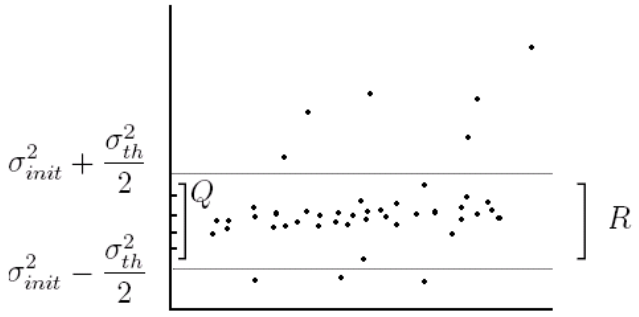


Fig. 4. Utilization of a robust LMS estimator in the proposed algorithm.

The breakdown point of a robust estimator is defined as the maximum percentage of outliers that can be injected into the assumed model before it fails (deviates largely from the expected behavior). It measures the robustness to outliers of an estimator [18]. The LMS is used because it has the highest breakdown point of 0.5. The mean and eventually any least squares based estimator has a breakdown point of 0, which means that a single outlier can impair the estimation result as opposed to 50% outliers in case of the LMS. When efficiency is more needed than accuracy, the simple median can be used instead of the LMS, which can be expressed by

$$\widehat{\sigma}^2_D = median(\sigma^2_{D\alpha}), \quad \sigma^2_{D\alpha} \in U_D^{\sigma^2}. \tag{21}$$

Note that the proposed LMS as opposed to the simple median [17] considers multiple estimates. The median estimate is impaired if the number of outliers exceeds the breakdown point of 0.5 due to large motion, complex structure, or noise. The proposed LMS allows us to reduce the number of outliers below the breakdown point, hence achieve better performance as the following example illustrates. We can decompose the set $U_T^{\sigma^2}$ of temporal noise estimates, when there is significant motion between frames, into three subsets: underestimates, accurate estimates, and overestimates of noise. Assume the overestimates are more than the underestimates. If the underestimates and overestimates are more than half of $U_T^{\sigma^2}$, then the median breakdown point of 0.5 is exceeded. To account for this, we find an initial estimate and limit the search for the best estimate to around it using an LMS estimator. The initial estimate is most likely drawn from the set of underestimates or accurate estimates. If we use the median instead as in [17] without initial estimation, the noise variance will be overestimated. If we use the median and limit the search to around the initial estimate, then the median can give an underestimate of noise if the initial estimate is drawn from the set of underestimates. On the other hand, an LMS estimator will be influenced by the remaining estimates and pulled up or down hence reducing the estimation error. Therefore, the LMS depends less on the accuracy of the initial estimate than the simple median.

Using (20), the quantities $\widehat{\sigma}^2_S$, $\widehat{\sigma}^2_T$, $\widehat{\sigma}^2_{ST}$, $\widehat{\sigma}^2_{HT}$ and $\widehat{\sigma}^2_{VT}$ are calculated. The variance v_D of the estimates in each $U_D^{\sigma^2}$ measures the reliability of the domain-wise estimate $\widehat{\sigma}^2_D$. The domain D that has the least v_D is considered the most reliable domain and will be the reference domain because its estimates are the most consistent (least varying). The estimate of that domain will be the reference estimate ($\widehat{\sigma}^2_{ref} = \underset{v_D}{\operatorname{argmin}}\, \widehat{\sigma}^2_D$). This step is important to account for the case of a domain completely failing when calculating the frame-wise estimate (e.g., when there is large global motion between frames and the temporal domain estimation is failing). The frame-wise noise variance is estimated using the domain-wise noise variances using

$$\widehat{\sigma}^2_\eta = \frac{1}{N_D} \sum_D \hat{\sigma}^2_D, \tag{22}$$

where N_D is the number of domain-wise estimates used. We only include in the averaging process the domain-wise estimates that do not exceed σ^2_{ref} by more than $\frac{\sigma^2_{th}}{2}$ to account for the case of complete estimation failure in a given domain. This way we adapt the integration of domain-wise estimates to the reliability of those estimates.

The effectiveness of the proposed method is due to the use of a three stage estimation process where each stage does not heavily rely on the success of the previous one. First, we collect the suitable candidate parts of the signals for estimation. We perform the selection independently in different domains to account for different scenarios which commonly arise in real video sequences. For example, it could be the case that the temporal domain is not suitable for noise estimation due to high global motion, but the spatial domain is suitable for noise estimation due to low structure and so on. If the selection fails due to the noise effect on the homogeneity analysis (which we try to reduce with the LoG operators) or other reasons, then the robust estimation with the LMS will help reject outliers. If the LMS estimation itself fails due to outliers over-exceeding the breakdown point, we reject the entire domain in the final step as we adaptively integrate the estimates from multiple domains.

IV. Simulation Results

A. Algorithm Parameters

In our simulations, L_{max} was set empirically to 15 and $\beta = 5$. Q can be varied between 5 and 15. We use $Q = 5$ which achieves enough accuracy for the target application (segmentation and tracking) (see (19)). σ^2_{th} is set to correspond to a PSNR value of 2.75 dB to create a wide enough search range around σ^2_{init}. For example, if the initial noise estimate is $\text{PSNR}^{\eta}_{init} = 19.3$ dB of a 20 dB noisy video sequence, then the search range is $\left[19.3 - \frac{2.75}{2}, 19.3 + \frac{2.75}{2}\right]$ in steps of $\frac{2.75}{5}$.

B. Computational Complexity

Several measures are taken to increase the computational speed of the proposed method. For example, we select the mask coefficients to represent simple change of sign and multiples of two operations which lend themselves to possible hardware implementation. Moreover, we utilize the fact that mask coefficients repeat themselves in different structure detectors (Fig. 2(a-e)) and need not be recalculated for each of the masks applied. Also, to speed up the median calculation process, the proposed method uses an algorithm that calculates the median of a set of size M without resorting to sorting [20]. When calculating the median, the fact that it is smaller than half of the data and larger than the other half is utilized. The procedure starts by taking the first value of the data and counting the number of elements Ms in the rest of the set that are smaller than the first value and the number of elements Mb that are larger. If $Ms \neq Mb$, the first value is not the median. If $Ms < Mb$, then we do not need to consider any value less than the first value because we already know that the median is larger than the first value. On the other hand, if $Ms > Mb$, then we do not need to consider any value larger than the first value because we are sure the median is smaller than the first value. The search proceeds until the median is found when $Ms = Mb$ for a specific value. This procedure is similar in nature and complexity to linear search. This means we reduce the complexity (number of search elements) of median calculation in the best case scenario to $O(M)$.

The average time needed for the proposed and referenced algorithms to process a 512×512 frame is measured and the ratio of the time needed by referenced methods to the time needed by the proposed method calculated (see Time Ratio in Table I). We implemented all methods using C++ and ran simulations under an Intel(R) Xeon(TM) CPU 2.40GHz machine operated by Linux. The proposed method is faster than all referenced methods except [13] (Table I) due to the added modifications for temporal processing.

TABLE I
TIME RATIO COMPARISON BETWEEN THE PROPOSED AND REFERENCED METHODS.

Method	Ours	[15]	[8]	[12]	[13]	[14]
Time Ratio	1.0	1.5	4.5	2.3	0.6	3.4

C. Objective Performance

To evaluate the performance of the proposed algorithm, the estimation error defined to be the absolute difference between the true value of the standard deviation of noise σ_η and the estimated value $\widehat{\sigma}_\eta$, or $E = |\sigma_\eta - \widehat{\sigma}_\eta|$, is used. The estimation error mean μ_E and variance σ^2_E are

$$\mu_E = \frac{\sum_{n=1}^{N_F} E(n)}{N_F}; \qquad \sigma^2_E = \frac{\sum_{n=1}^{N_F} (E(n) - \mu_E)^2}{N_F - 1}, \quad (23)$$

where N_F is the total number of test frames used. While μ_E measures the performance of a noise estimation algorithm, σ_E measures the reliability of that performance. The standard video sequences *Prlcar, Tennis, Train, Football, Car and Flowergarden* are corrupted with 20, 30 and 40 dB AWGN. Noise is estimated for the first 50 frames of each sequence using $W = 3$ cubic windows.

Table II shows that the proposed algorithm has the most reliable performance for different noise levels. It enhances the performance of the approach in [17]. Figs. 5-9 show the individual estimation error for the test sequences used for the proposed and referenced methods at different noise levels. The proposed method produces less error for all sequences and noise levels. Fig. 10 shows the average estimation error over time, μ_E, and estimation error standard deviation, σ_E, averaged over all test sequences for every noise level. As can be seen from Fig. 10, the proposed method gives a lower average estimation error than referenced methods and is temporally stable. It also shows that the reliability of the proposed method is better than referenced methods for all noise levels.

TABLE II
THE AVERAGE (OVER ALL TEST VIDEO FRAMES) AND THE STANDARD DEVIATION OF THE ESTIMATION ERROR FOR 20, 30 AND 40 dB NOISE.

	20 dB		30 dB		40 dB	
Alg.	μ_E	σ_E	μ_E	σ_E	μ_E	σ_E
Spatio-temporal						
Ours	0.23	0.33	0.50	0.41	0.65	0.68
[15]	2.80	0.77	2.53	1.19	3.1	5.78
[17]	0.61	0.83	0.87	0.91	0.98	1.08
Spatial only						
[8]	1.99	1.20	3.21	1.42	4.34	1.70
[12]	0.79	1.13	1.01	1.20	1.10	1.24
[13]	1.60	1.55	2.39	1.25	1.91	1.16
[14]	1.75	1.26	2.12	1.81	3.36	2.70

V. Conclusion

This paper proposed a technique in which the variance of the AWGN noise is estimated from selected homogeneous cubes in the 3D video signal. Spatial, temporal and spatio-temporal homogeneity are measured using LoG-based operators. The noise variance is estimated from the local variances of selected homogeneous cubes calculated along intensity uniform plains.

An LMS Robust estimator is utilized to obtain the domain-wise noise variance estimate. The domain-wise noise variance estimates are adaptively integrated to obtain the frame-wise final noise variance estimate. The proposed algorithm was tested on video sequences with high structure and motion activity. It performed reliably for different noise and video activity levels and with a maximum estimation error of 1.7 dB PSNR. This study shows that it is possible to incorporate temporal information in AWGN variance estimation to achieve significant accuracy gain with reduced complexity.

REFERENCES

[1] M. Spann and R. Wilson, "A quad-tree approach to image segmentation which combines statistical and spatial information," *Pattern Recognition*, vol. 18, no. 3-4, pp. 257–269, 1985.

[2] G. Calvagno, F. Fantozzi, R. Rinaldo, and A. Viareggio, "Model-based global and local motion estimation for videoconference sequences," *IEEE Transactions Circuits and Systems for Video Technology*, vol. 14, no. 9, pp. 1156–1161, 2004.

[3] F. D. Murtagh, A. Bijaoui, and J.-L. Starck, *Image Processing and Data Analysis: The Multiscale Approach*, pp. 46–74, Cambridge University Press, 1998.

[4] A. Bovik, *Handbook of Image and Video Processing*, pp. 275–285, Elsevier Academic Press, second edition, 2005.

[5] G. de Haan, T. G. Kwaaitaal-Spassova, M. Larragy, and O. A. Ojo, "Memory integrated noise reduction IC for television," *IEEE Transactions on Consumer Electronics*, vol. 42, no. 2, pp. 175–181, 1996.

[6] B. C. Song and K. W. Chun, "Noise power estimation for effective de-noising in a video encoder," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 357–360, March 2005.

[7] B. C. Song and K. W. Chun, "Motion-compensated noise estimation for efficient pre-filtering in a video encoder," *IEEE International Conference on Image Processing*, vol. 2, pp. 211–214, September 2003.

[8] S. I. Olsen, "Estimation of noise in images: an evaluation," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 4, pp. 319–323, July 1993.

[9] K. Rank, M. Lendl, and R. Unbehauen, "Estimation of image noise variance," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 146, no. 2, pp. 80–84, April 1999.

[10] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, April 1994.

[11] A. De Stefano, P. R. White, and W. B. Collis, "Training methods for noise level estimation on wavelet components," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 16, 2004.

[12] D.-H. Shin, R.-H. Park, Y. Seungjoon, and J.-H. Jung, "Block-based noise estimation using adaptive Gaussian filtering," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 218–226, February 2005.

[13] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 113–118, January 2005.

[14] E. J. Balster, Y. Zheng, and R. Ewing, "Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising," *IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions*, vol. 16, no. 2, pp. 220–230, February 2006.

[15] G. de Haan, T. G. Kwaaitall-Spassova, M. M. Larragy, O. A. Ojo, and R. J. Schutten, "Television noise reduction IC," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 1, pp. 143–153, 1998.

[16] V. Zlokolica, A. Pizurica, and W. Philips, "Noise estimation for video processing based on spatio-temporal gradients," *IEEE Signal Processing Letters*, vol. 13, no. 6, pp. 337–340, June 2006.

[17] M. Ghazal, A. Amer, and A. Ghrayeb, "Structure-oriented spatio-temporal video noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 845–848, May 2006.

[18] G. L. Shevlyakov and N. O. Vilchevski, *Robustness in Data Analysis: Criteria and Methods*, pp. 6–10, VSP BV, 2002.

[19] C. V. Stewart, "Robust parameter estimation in computer vision," *SIAM Review*, vol. 41, no. 3, pp. 513–537, 1999.

[20] B. Parhami, *Introduction to Parallel Processing: Algorithms and Architectures*, pp. 111–112, Springer, first edition, 1998.

Tennis

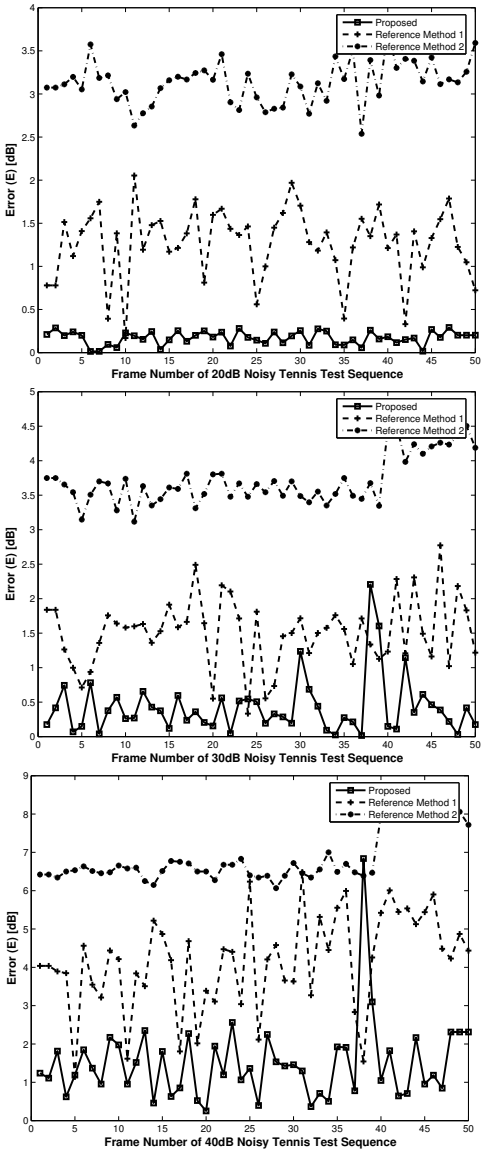
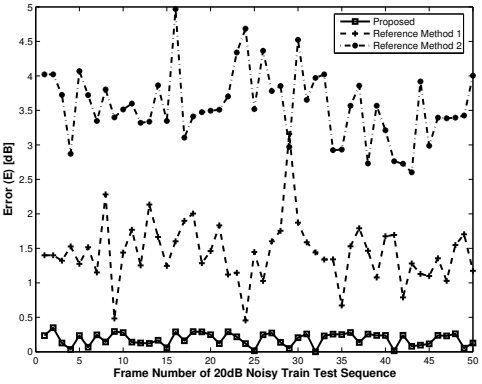
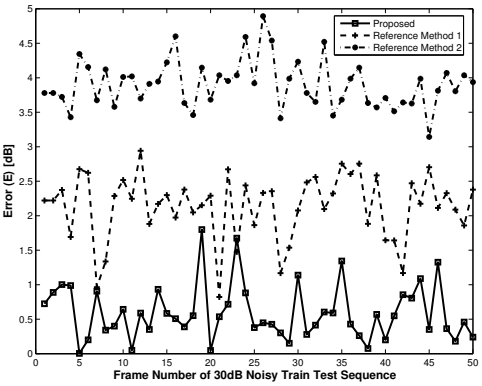


Fig. 5. Tennis: Estimation error over time for proposed, reference 1 [12], and reference 2 [15] methods.

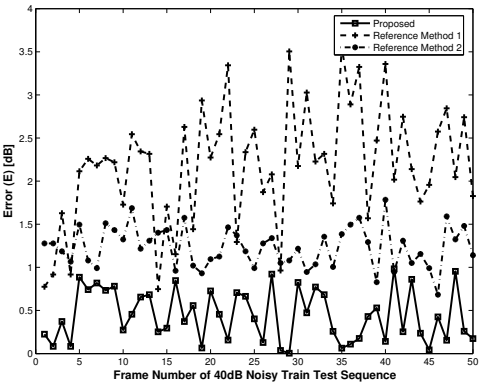
(a) Train



(b) 20 dB



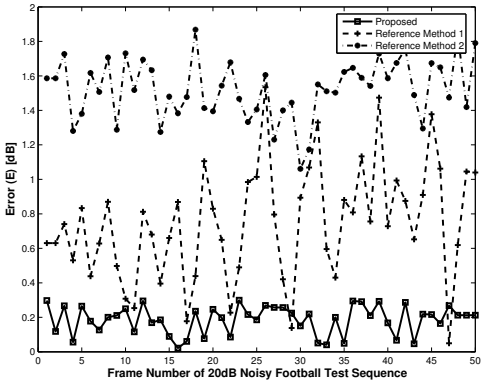
(c) 30 dB



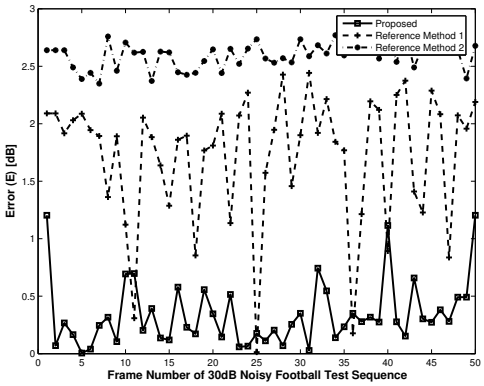
(d) 40 dB

Fig. 6. Train: Estimation error over time for proposed, reference 1 [12], and reference 2 [15] methods.

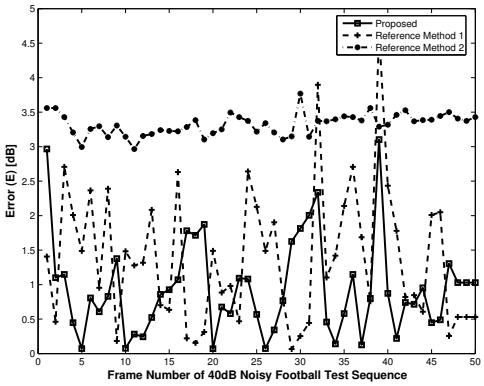
(a) Football



(b) 20 dB



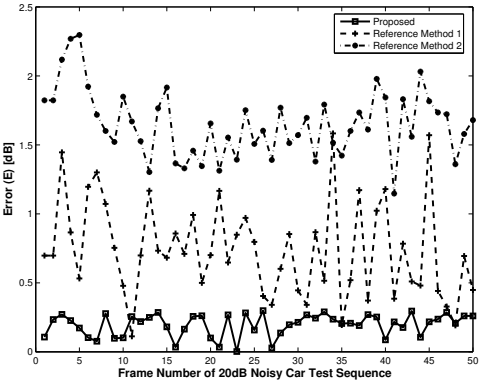
(c) 30 dB



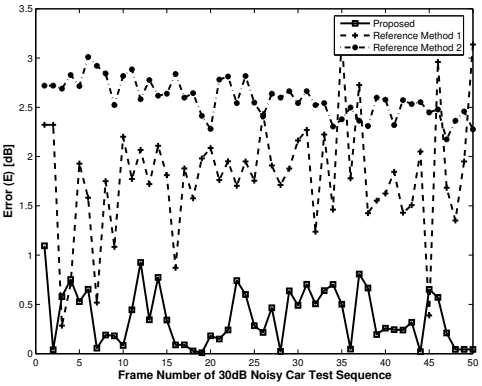
(d) 40 dB

Fig. 7. Football: Estimation error over time for proposed, reference 1 [12], and reference 2 [15] methods.

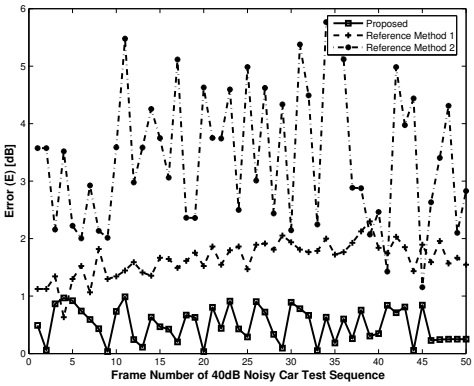
(a) BBCcar



(b) 20 dB



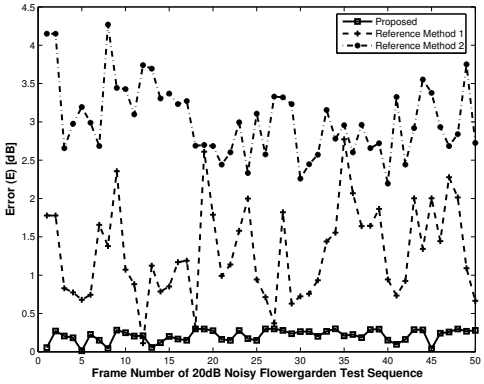
(c) 30 dB



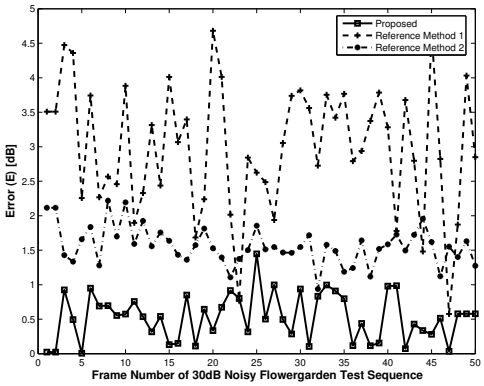
(d) 40 dB

Fig. 8. BBCcar: Estimation error over time for proposed, reference 1 [12], and reference 2 [15] methods.

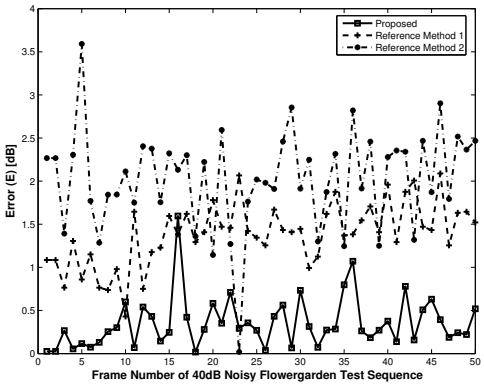
(a) Flowergarden



(b) 20 dB

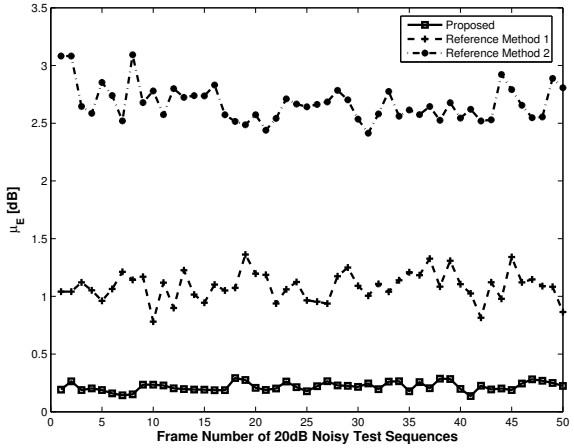


(c) 30 dB

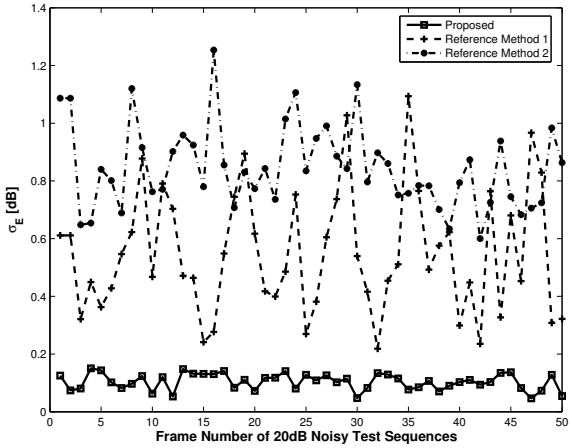


(d) 40 dB

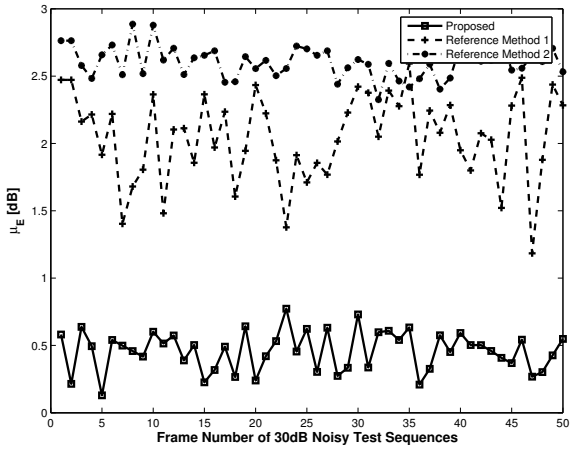
Fig. 9. Flowergarden: Estimation error over time for proposed, reference 1 [12], and reference 2 [15] methods.



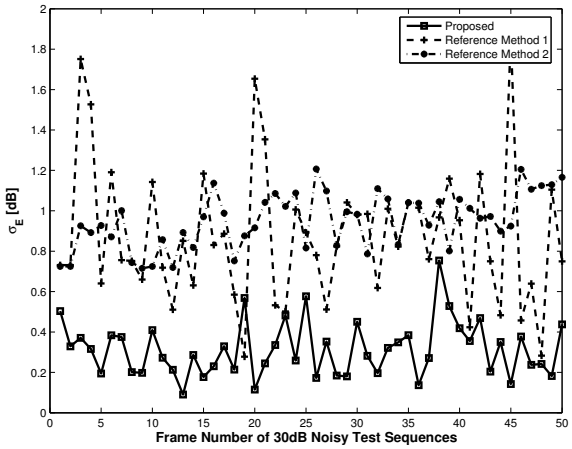
(a) Mean of Error for 20dB



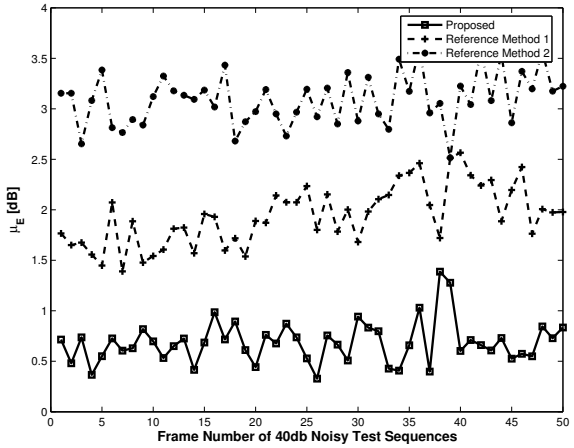
(b) Standard Deviation of Error for 20dB



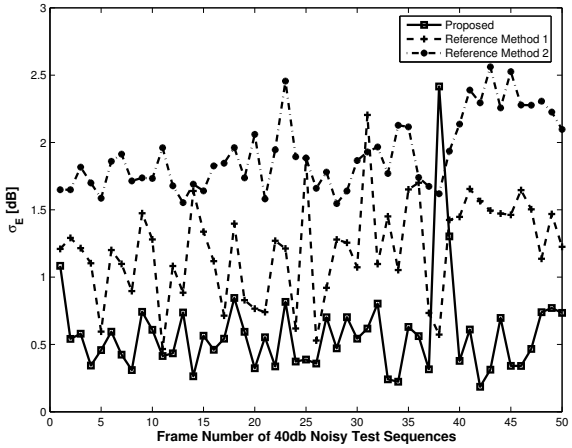
(c) Mean of Error for 30dB



(d) Standard Deviation of Error for 30dB



(e) Mean of Error for 40dB



(f) Standard Deviation of Error for 40dB

Fig. 10. All test sequences: Mean μ_E and standard deviation σ_E of error over time for proposed and referenced methods (reference 1 [12] and reference 2 [15]) for 20dB, 30dB and 40dB noisy test sequences.