STRUCTURE-ORIENTED SPATIO-TEMPORAL VIDEO NOISE ESTIMATION

Mohammed Ghazal, Aishy Amer and Ali Ghrayeb

Electrical and Computer Engineering, Concordia University
Montréal, Québec, Canada
{moha_mo,amer,aghrayeb}@ece.concordia.ca

ABSTRACT

Noise can highly impact the performance of video processing algorithms. This paper proposes a new real-time spatio-temporal method for estimating the noise variance in video signals. The proposed algorithm selects 3D regions or cubes in the video signal with high intensity uniformity. The noise variance is estimated from the selected set of spatially, temporally and spatio-temporally intensity-uniform cubes using local variances calculated along homogeneous plains. The proposed algorithm works well for sequences with high structure and motion activity and outperforms other methods with a worst-case estimation error of 2 dB. It works well for highly noisy and non-noisy sequences.

1. INTRODUCTION

Many video processing algorithms such as video quality enhancement, compression, deinterlacing, motion estimation and file format conversion require a priori knowledge of the noise present in the signal in order to adapt their parameters and improve performance. Hence, the need for a fast, accurate and robust video noise estimation algorithms.

Proposed algorithms for estimating the variance (σ_η^2) of the additive white Gaussian noise (AWGN) are either inter-frame or intra-frame based. There exist few methods for inter-frame noise estimation [1,2]. These methods are challenged by the presence of object or global motion. Motion detection or motion compensation are commonly used countermeasures. Hence, methods in this area such as the one in [1] tend to be computationally expensive. The method in [2] attempts to incoperate temporal adaptation to stabilize the spatially estimated noise variance.

Many methods for intra-frame noise estimation has been presented. Difficulties with these methods rise from frames with very high or very low noise levels as well as highly structured frames. The problem lies in determining if the intensity variations are due to noise or frame details. Intra-frame methods are categorized into smoothing-based, wavelet-based and block-based methods. Smoothing based algorithms such as the one in [3] estimate noise from the difference of the noisy frame and its smoothed version. The assumption is that the difference frame represents an approximation of the noise signal. These approaches are computationally expensive and tend to overestimate the noise variance.

The authors in [4] use the wavelet domain to decompose the frame into sub-bands. The coefficients of the diagonal details or the *HH* (*H*igh-*H*igh) band are used to estimate the noise variance. Methods that use the wavelet domain are similar to the smoothing-based methods in overestimating the noise variance because the *HH* band has also high frequency frame information.

Block-based methods in [5] and [6] are less computationally demanding. These methods attempt to locate regions with the least amount of signal information. The intensity variations in these regions is assumed to be due to noise. The algorithm in [5] 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 [6] proposes a novel 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 the blocks selected to be the most homogeneous.

This paper proposes a low-complexity algorithm that uses both intra-frame and inter-frame information to yield a stable and accurate estimate of the noise variance. The proposed method divides the video signal into cubes and measures their homogeneity. The noise variance is then estimated from a set of selected cubes along the homogeneous plains only.

The remainder of the paper is as follows. Section 2 presents the proposed approach theoretically and gives an interpretation of its good performance. Objective simulation results are presented and discussed in Section 3. Finally, Section 4 concludes the paper.

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

2. PROPOSED APPROACH

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

Define a noisy digital video signal V_η

$$V_\eta(i,j,n) = V(i,j,n) + \eta(i,j,n), \tag{1}$$

where i and j are the spatial coordinates, n is the temporal coordinate and $\eta(i,j,n)$ is the amount of noise. Since the algorithm is designed to be context-free, there are no restrictions on the original signal V. We assume that pixels in $\{V_\eta(i,j,n)\}$ are independent and identically distributed (iid) but not necessarily zero mean. The division of V_η into cubes C_{klm} with spatial indecies k and l and temporal index m is done using

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

where Ψ_{klm} is a cubic window of size W^3 ($W \in$ odd $\mathbb{Z}^+$) centered around the 3D point $(k,l,m) \in V_\eta$. To locate the homogeneous cubes in the video signal, we define a set of low-complexity homogeneity measures with (3). Theoretically, these measures represent the quantities in (4)-(8).

$$\{\zeta_X\}, X \in \{ST, T, S, VT, HT\}; \tag{3}$$

$$\zeta_{ST} = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial j^2} + \frac{\partial^2 V_\eta}{\partial n^2}|; \tag{4}$$

$$\zeta_T = |\frac{\partial^2 V_\eta}{\partial n^2}|; \tag{5}$$

$$\zeta_S = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial j^2}|; \tag{6}$$

$$\zeta_{VT} = |\frac{\partial^2 V_\eta}{\partial j^2} + \frac{\partial^2 V_\eta}{\partial n^2}|; \tag{7}$$

$$\zeta_{ZT} = |\frac{\partial^2 V_\eta}{\partial i^2} + \frac{\partial^2 V_\eta}{\partial n^2}|. \tag{8}$$

Our proposed homogeneity measures are the magnitudes of directional Laplacian operators. For (4)-(8) to be useful, they must be expressed in discrete form. For this purpose, we define the masks in Fig. 1.

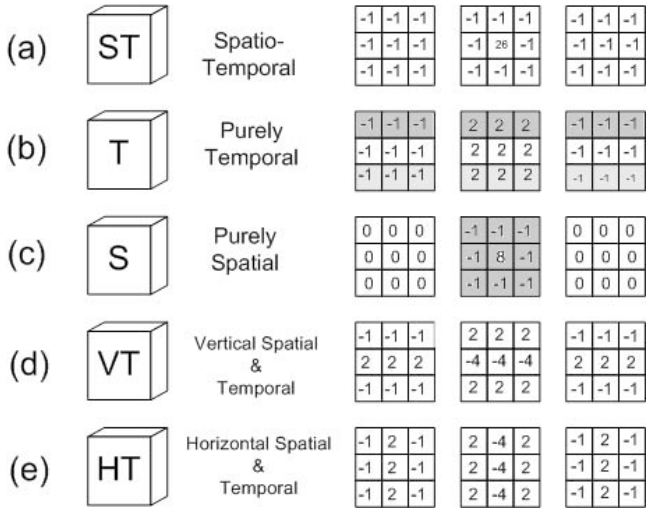

Fig. 1. Homogeneity analyzer cubical masks where pixels in the same gray level belong to one plain.

Fig. 1(a) is a 3D Laplacian operator used to measure spatio-temporal homogeneity or ζ_{ST} in (4). The central coefficient of the mask (mask's 3D midpoint) can be calculated using W^3-1. The central coefficient accumulates to this value as a result of combining the 2^{nd} derivatives in all directions. The mask in Fig. 1(b) evaluates homogeneity along the temporal direction or ζ_T in (5). It acts as a local low-complexity motion detector. The mask in Fig. 1(c) is the spatial domain Laplacian operator. It measures purely spatial homogeneity or ζ_S defined in (6). This mask's response is an approximation of the sum of directional responses of the masks defined in [6]. The mask in Fig. 1(d) measures both the homogeneity along the spatial vertical direction and the temporal direction or ζ_{VT} in (7). Similarly, the mask in Fig. 1(e) measures the homogeneity along the spatial horizontal direction and the temporal direction or ζ_{HT} in (8). All masks were designed to have a degree of rotational invariance to account for unpredictable object shapes or movements. For example, the mask in Fig. 1(c) is 45^o isotropic in the spatial domain. This means that the masks response is invariant to 45^o rotations.

Other masks can be designed to measure homogeneity along the diagonal or other directions. The only restriction on the mask is that it fulfills the following 2^{nd} derivative definition requirements: The mask response must be (1) zero in flat areas; (2) nonzero at the onset and end of an intensity step or ramp; and (3) zero along ramps. The reason the second and not the first derivative

was chosen is because it has a stronger response to fine details. To overcome the second derivative sensitivity to noise, we will use the variance to exclude overestimated or underestimated noise power using the median function.

The quantities $\{\zeta_X\}$ in (4)-(8) are calculated for every cube C_{klm} by applying the masks in Fig. 1 to the video signal. We now define a set, U_C, to be the set of all selected homogeneous cubes as depicted by

$$U_C = \left\{ C_{klm} | \min_{klm}^{L}(\zeta_X) \right\}, L \in \mathbb{Z}, \tag{9}$$

which indicates that we are considering the set of the L most homogeneous cubes selected by ζ_X. The five different masks in Fig. 1 are used, therefore the cardinality of set U_C is equal to $5L$. L was fixed to 10% of the total number of blocks in [3] and [6]. In our proposed algorithm, L is variable and is computed as

$$L = L_{max} - \frac{PSNR_{init}}{\beta}, \tag{10}$$

where $PSNR_{init}$ is the initial estimate of the Peak Signal to Noise Ratio (PSNR) calculated from the median of the local variances of the 3 most homogeneous cubes over all ζ_X. L_{max} is the maximum number of cubes to be used and β is a scaling factor. The choice of L_{max} is arbitrary between 5 and 30. In our simulations, L_{max} was set to 15 and $\beta = 5$. The function $L(PSNR_{init})$ can be replaced by any monochromically decreasing positive function of $PSNR_{init}$ and is used to ensure the inclusion of more cubes in case of noisy video sequences and less cubes in case of non-noisy ones. Using less cubes in case of non-noisy video results in a more reliable estimate. 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 (see Fig. 1) found to be most homogeneous. For all cubes found to be spatially homogeneous (by substituting $X = S$ in (9), we use

$$\begin{aligned} \mu_S &= \frac{\sum_{(i,j)\in\Psi_{kl}} V_\eta(i,j)}{W^2}; \\ \sigma_S^2 &= \frac{\sum_{(i,j)\in\Psi_{kl}} (V_\eta(i,j)-\mu_S)^2}{W^2-1} \end{aligned}, \tag{11}$$

where Ψ_{kl} states that we use only pixels along the middle spatial plain of the cube (See Fig. 1(c)). For cubes found to be temporally homogeneous we use

$$\begin{aligned} \mu_{T_\rho} &= \frac{\sum_{(i,n)\in\Psi_{km}} V_\eta(i,n)}{W^2}; \\ \sigma_{T_\rho}^2 &= \frac{\sum_{(i,n)\in\Psi_{km}} (V_\eta(i,n)-\mu_{T_\rho})^2}{W^2-1} \end{aligned}, \tag{12}$$

where Ψ_{km} indicates that we use only pixels along temporal plains (See Fig. 1(b)). Using (12), we calculate the sample mean μ_{T_ρ} and variance $\sigma_{T_\rho}^2$ along each plain $\rho = \{1, .., W\}$ and then compute the average over the W plains. It is important that the noise variance is estimated using plains determined to be most homogeneous only as we have no information about the homogeneity along other plains. For cubes that are chosen to be spatio-temporally most homogeneous (i.e., $\min(\zeta_Y)$, $Y = \{ST, HT, VT\}$), the sample mean and variance are calculated over all pixels in the cube using

$$\begin{aligned} \mu_Y &= \frac{\sum_{(i,j,n)\in\Psi_{klm}} V_\eta(i,j,n)}{W^2}; \\ \sigma_Y^2 &= \frac{\sum_{(i,j,n)\in\Psi_{klm}} (V_\eta(i,j,n)-\mu_Y)^2}{W^3-1} \end{aligned}. \tag{13}$$

The dimension-based (i.e., spatial, temporal and spatio-temporal) noise variance is estimated using the set of L local noise variances of the selected homogeneous cubes. For the spatio-temporal dimension, this set is denoted $\{\sigma_{ST\alpha}^2\}, \alpha \in \{1, 2, ..., L\}$. The overall noise variance for that set, $\widehat{\sigma}_{ST}^2$, is calculated using the median variance over the set as

$$\widehat{\sigma}_{ST}^2 = median(\sigma_{ST\alpha}^2), \quad \alpha = \{1, 2, ..., L\}. \tag{14}$$

Similarly, the quantities $\widehat{\sigma}_S^2$, $\widehat{\sigma}_T^2$, $\widehat{\sigma}_{HT}^2$ and $\widehat{\sigma}_{VT}^2$ are calculated. An alternative approach to (14) is to use

$$\widehat{\sigma}_X^2 = \sum_{\alpha\in\{1,2,..,L\}} w(\zeta_X)\sigma_{X_\alpha}^2, \tag{15}$$

where $w(\zeta_X) = \gamma\zeta_X$ is the weight assigned to $\sigma_{X_\alpha}^2$. Notice that if $w(\alpha) = 1/L$, $\widehat{\sigma}_X^2$ will be the average variance.

The frame-wise noise variance is then estimated using the median of domain based noise variances as shown in

$$\widehat{\sigma}_\eta^2 = median(\widehat{\sigma}_X^2). \tag{16}$$

3. RESULTS AND EVALUATION

To evaluate the performance of the algorithm, 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 $E_k = |\sigma_\eta - \widehat{\sigma}_\eta|$, is used. The estimation error average μ_{E_k} and variance $\sigma_{E_k}^2$ are computed using (17)

$$\mu_{E_k} = \frac{\sum_{i=1}^{N} E_k(i)}{N}; \quad \sigma_{E_k}^2 = \frac{\sum_{i=1}^{N} (E_k(i) - \mu_{E_k})^2}{N}, \tag{17}$$

where N is the total number of test frames used. While μ_{E_k} measures the performance of a noise estimation algorithm, $\sigma_{E_k}^2$ measures the reliability of that performance.

The standard video sequences *Prlcar, Tennis, Train, Football, Car and Flowergarden* were corrupted with 20, 30 and 40 dB AWGN. Simulation was run on the first 50 frames of each sequence using $W = 3$ cubic windows. Average time needed for the proposed and reference algorithms was measured and the Time Ratio (TR) between them was calculated accordingly. Implementation was using C++ under an Intel(R) Xeon(TM) CPU 2.40GHz machine running Linux. The proposed method was found to be faster than all reference methods except [6].

Table 1 shows that the proposed algorithm has the most reliable performance for different noise levels. Fig. 3 shows the estimation error over time, μ_{E_k}, averaged over all test sequences. As can be seen from Fig. 3, the proposed method gives a lower average estimation error than reference methods and is temporally stable. Due to space constraints, a figure showing the estimation error standard deviation, σ_{E_k}, over time was excluded. The figure showed that the reliability of the proposed method is better than reference method for all noise levels. The proposed method requires more memory than spatial methods such as [6], however, already delayed frames for compression or other video processing can be used.

Table 1. The average and the standard deviation of the estimation error for 20, 30 and 40 dB noise.

	20 dB		30 dB		40 dB		
Alg.	μ_{E_k}	σ_{E_k}	μ_{E_k}	σ_{E_k}	μ_{E_k}	σ_{E_k}	TR
Inter-frame							
Ours	0.61	0.83	0.87	0.91	0.98	1.08	1.0
[2]	1.30	1.77	1.53	1.79	5.54	5.78	1.1
Intra-frame							
[3]	1.99	1.20	3.21	1.42	4.34	1.70	4.7
[4]	1.75	1.26	2.12	1.81	3.36	2.70	2.4
[5]	0.79	1.13	1.01	1.20	1.10	1.24	2.5
[6]	1.60	1.55	2.39	1.25	1.91	1.16	0.8

4. CONCLUSION

This paper proposed a video noise estimation 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 3D Laplacian operators. The noise variance is estimated from the local variances of selected homogeneous cubes calculated along intensity uniform plains. The proposed algorithm works well for video sequences with high structure and motion activity. It performs reliably with different noise levels with a maximum estimation error of 2 dB.

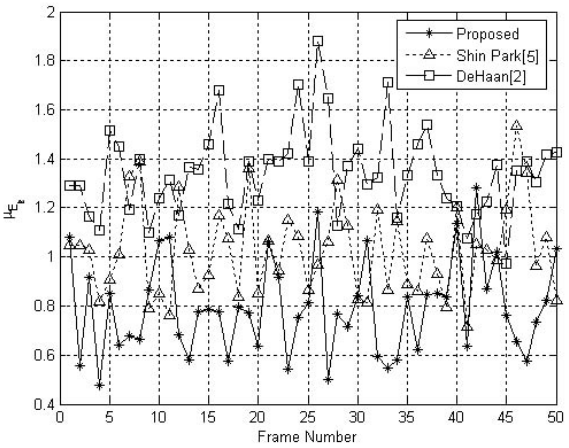

Fig. 2. Average error μ_{E_k} over time for proposed and reference methods over all test sequences.

5. REFERENCES

[1] 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.

[2] T. Kwaaitaal-Spassova G. deHaan and O.A. Ojo, "Automatic 2-d and 3-d noise filtering for high quality television receivers," *International Workshop on Signal Processing and HDTV*, vol. VI, pp. 221–230, 1996.

[3] S. I. Olsen, "Estimation of noise in images: An evaluation," *Graphical Models and Image Process.*, vol. 55, pp. 319–323, July 1993.

[4] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, Apr 1994.

[5] D. H. Shin and R. H. Park, "Block-based noise estimation using adaptive gaussian filtering," *IEEE Transactions on Consumer Electronics*, vol. 51, pp. 218–226, Feb 2005.

[6] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 113–118, January 2005.