A No-reference Perceptual Quality Metric for Videos Distorted by Spatially Correlated Noise

Chao Chen
Google Inc.
1600 Amphitheatre Parkway
Mountain View, California
chaochen@google.com

Mohammad Izadi
Google Inc.
1600 Amphitheatre Parkway
Mountain View, California
izadi@google.com

Anil Kokaram
Google Inc.
1600 Amphitheatre Parkway
Mountain View, California
anilkokaram@google.com

ABSTRACT

Assessing the perceptual quality of videos is critical for monitoring and optimizing video processing pipelines. In this paper, we focus on predicting the perceptual quality of videos distorted by noise. Existing video quality metrics are tuned for "white", i.e., spatially uncorrelated noise. However, white noise is very rare in real videos. Based on our analysis of the noise correlation patterns in a broad and comprehensive video set, we build a video database that simulates the commonly encountered noise characteristics. Using the database, we develop a perceptual quality assessment algorithm that explicitly incorporates the noise correlations. Experimental results show that, for videos with spatially correlated noises, the proposed algorithm presents high accuracy in predicting perceptual qualities.

CCS Concepts

•Human-centered computing → User models; Laboratory experiments; Empirical studies in HCI;

Keywords

Subjective Video Quality Assessment, Noise, Power Spectrum Density

1. INTRODUCTION

During video acquisition, compression and communication, various types of distortion may be introduced. Among them, noise is typically introduced in the acquisition process, and it may propagate throughout the video processing and communication pipeline. Accurately predicting the impact of noise on perceptual quality helps in optimizing the overall efficiency of video services. For example, using the estimated noise level, we may be able to predict the likelihood of severe blocking in a subsequent transcoding operation[1]. We can then select a suitable pre-processor to reduce the knock-on effects both on bitrate and quality.

The best way of assessing video quality is to collect human scores by conducting subjective tests. However, human studies are labor intensive and not applicable to real-time applications. It is, therefore, appealing to devise an algorithm that can accurately predict perceptual quality. Perceptual image and video quality assessment have been extensively studied, and several high-performance metrics have been invented. For image quality assessment, SSIM (Structural SIMilarity index [29]) achieves very high correlation with perception and has been widely used due to its low complexity. For video quality assessment, MOVIE (MOtion-based Video Integrity Evaluation index [24]) and STMAD (Spatial-Temporal Most Apparent Distortion [28]) report the best accuracy.

A common drawback of these high-performance metrics is that they need a pristine reference of the video clip under test. In practice, for instance in quality assessment of YouTube uploads, such a reference is unavailable. This limitation motivated research on no-reference image/video quality metrics. One approach to this problem is to measure the "unnaturalness" of a picture or a video. It has been found that certain statistics of a pristine image/video follow certain models[27] [34]. Quality can then be evaluated by measuring the deviation of these statistics from the models. The image quality metrics proposed by [16], [18], [26] and video quality metrics proposed by [23] and [33] follow this methodology. Another method of no-reference quality prediction is based on the free-energy principle in brain science [9]. In this theory, when a visual signal is perceived, the human brain tries to infer the signal using an internal generative model. The free energy of this inference process, i.e., the discrepancy between the visual signal and the inference of the internal generative model thus can be used to quantify the level of distortion. The algorithms proposed in [15] and [32] follow this method to predict image quality.

Besides the aforementioned general-purpose quality metrics, several algorithms are specifically designed for noise [14, 30, 31]. These algorithms, along with the general-purpose algorithms are designed or tuned for predicting the videos distorted by white noise. However, the noise on neighboring pixels is typically correlated. As demonstrated in [25], the demosaicing operation in video acquisition is performed over neighboring pixels and thus introduces correlation. We verified this in the videos uploaded to YouTube and found that the noise all presents some degree of correlations (see section 2.1 for more details).

In this paper, we propose a no-reference video quality metric for realistic noise. To this end, we first analyzed the noise

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2964302

correlation patterns on a vast and comprehensive video set. Using the typical noise correlation patterns we found in the video set, we construct a video database that simulates the commonly encountered noise. A series of subjective tests are then performed to evaluate the perceptual quality of the noisy videos in the database. Then, we develop a perceptual quality metric that explicitly incorporates noise correlation patterns in quality assessment. Experimental results show that the proposed algorithm can accurately predict the perceptual quality of noisy videos.

The rest of the paper is organized as follows: section 2 describes the construction of our quality database for practical noisy videos. Section 3 details the proposed quality metric. Section 4 shows the performance of the proposed metric. The paper is concluded in section 5

2. CONSTRUCTING THE DATABASE

Since the number of sample videos in a subjective test session is limited, they should be carefully selected to represent noise patterns observed in practice. We analyzed the noise structure of a large set of consumer generated videos. Based on the distribution of the noise structures, we created a series of test videos that are representative. The perceptual quality of the videos is then assessed in a series of subjective tests for the development of our quality prediction algorithm.

2.1 Realistic Noise Analysis

Every type of video capture pipeline may present different noise statistics. Hence, there exist many different noise structures in practical videos. YouTube is a video-sharing platform with an enormous amount of consumer-generated video contents. We collect 46,000 videos uploaded to YouTube to analyze their noise patterns. These videos are selected such that

- they represent varying contents e.g. sports, movies, nature, music video, etc.
- they are captured with over 200 different types of cameras or device models.
- they represent varying resolutions including 360p, 720p, 1080p, 2K and 4K.
- their length is longer than 30 seconds to focus on contents that have a chance of being watched.

We analyze the noise patterns of the video collection with a 2D auto-regressive (AR) model. This model has been widely used to analyze noise structure and synthesize film grain [8, 6]. Denoting by $\mathsf{N}(x, y)$ the synthesised noise at position (x, y), the AR model is:

$$\mathsf{N}(x,y) = \sum_{(i,j)\in\mathcal{S}_\Delta} w_{i,j}.\mathsf{N}(x+i, y+j) + \mathsf{G}^\sigma(x,y), \qquad (1)$$

where $\mathsf{G}^\sigma(x, y) \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable with zero mean and variance σ^2. The set $\mathcal{S}_\Delta$ is the square neighborhood of $(0, 0)$ with size $(2\Delta + 1) \times (2\Delta + 1)$. The noise at a pixel (x, y) is modeled as the sum of a white noise sample and a weighted sum of the noise in its neighborhood. The weight assigned to the pixel at $(x+i, y+j)$ is the model parameter $w_{i,j}$. Note that this is a prediction equation and so $w_{i,j} = 0$ for the pixels succeeding to the center pixel in raster-scan order. In the following we denote by $\boldsymbol{w}$ the vector of weight parameters, i.e., $\boldsymbol{w} = (w_{i,j} : \forall(i, j) \in \mathcal{S}_\Delta)^\mathsf{T}$. The parameter σ^2 characterizes the power of noise. For a fixed $\boldsymbol{w}$, larger σ^2 leads to larger noise power.

To analyze the structure of the noise in a video sequence we employ aspects of the noise estimation method proposed in [13]. That method estimates the Power Spectral Density [20] of the noise in texture-less areas of video frames. This idea has been used in well-known post-production toolkits e.g. NUKE and Neat Video. The rationale of this method is that the noise is easier to separate in flat regions of a frame. The method first searches for flat patches from all the frames in a video and then applies hyper-planes to approximate the true values of the patches. The approximation error on the patch is considered as the noise. We applied this method on our video collections and fit the AR model to the separated noise to obtain minimum mean square error estimations for the model parameters $\boldsymbol{w}$ and σ. From this analysis, we have the following observations:

1. An AR model with $\Delta = 4$ is sufficient to model the noise in all the 46000 clips ($\Delta > 4$ did not improve the fitting error significantly compared to $\Delta = 4$). This implies that the noise at two pixels separated by a distance of > 4 is uncorrelated in a typical video. In the following analysis, we use $\Delta = 4$.

2. The histogram of $||\boldsymbol{w}||_2$ estimated on the video collection is shown in Figure 1. According to (1), $\boldsymbol{w}$ reflects the correlation among neighboring pixels. As $||\boldsymbol{w}||_2^2$ decreases to zero, the noise $\mathsf{N}(x, y)$ converges to white noise $\mathsf{G}(x, y)$. It is shown in the figure that a large fraction of videos in our collection has non-zero $||\boldsymbol{w}||_2^2$. This observation suggests that the noise in real videos is not white in general.

3. We note that the estimation of the model parameter $\hat{\boldsymbol{w}}$ and $\hat{\sigma}$ may be inaccurate. For videos with textures, the noise separation method may confuse the texture with noise and overestimate the correlation. Therefore, we cannot directly use the noise pattern obtained from the video collection to synthesize the noise for our subjective test.

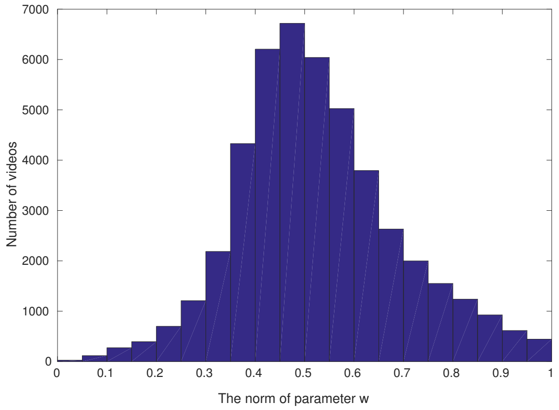


Figure 1: The histogram of $||\hat{\boldsymbol{w}}||_2^2$ estimated on a video collection of 46000 videos.

For $\Delta = 4$, the number of non-zero weights $w_{i,j}$ of our AR model is $((2\Delta+1)^2-1)/2 = 40$. To find out the most typical noise pattern, we may perform a Principal Component Analysis (PCA) on all the $\hat{\boldsymbol{w}}$ in the 40-dimensional space and use the eigenvectors as typical noise patterns for noise synthesis. However, as discussed above, the estimated $\hat{\boldsymbol{w}}$ may contain outliers, and PCA is not robust for the analysis of data with outliers [3].

To rule out poorly estimated vectors $\boldsymbol{w}$, we applied K-means clustering [2] to cluster the estimated vectors into K groups. We vary the number of groups K from 3 to 15 and, for a given K, calculate the mean vector of each group as $\left\{\bar{\boldsymbol{w}}^{\mathrm{K}}_{\ell} : 1 \leq \ell \leq \mathrm{K}\right\}$. Thus we obtain the least correlated and most correlated noise parameter as

$$\bar{\boldsymbol{w}}^{\mathrm{K}}_{\min} = \arg\min_{\bar{\boldsymbol{w}}^{\mathrm{K}}_{\ell}} ||\bar{\boldsymbol{w}}^{\mathrm{K}}_{\ell}||_2, \qquad (2)$$

and

$$\bar{\boldsymbol{w}}^{\mathrm{K}}_{\max} = \arg\max_{\bar{\boldsymbol{w}}^{\mathrm{K}}_{\ell}} ||\bar{\boldsymbol{w}}^{\mathrm{K}}_{\ell}||_2, \qquad (3)$$

We observed that both $\bar{\boldsymbol{w}}^{\mathrm{K}}_{\min}$ and $\bar{\boldsymbol{w}}^{\mathrm{K}}_{\min}$ is not changing further as we increase K over 10. Therefore, we fix K = 10 and define a linear interpolation function between $\bar{\boldsymbol{w}}^{10}_{\min}$ and $\bar{\boldsymbol{w}}^{10}_{\max}$ as:

$$\mathbf{w}(\alpha) = \bar{\boldsymbol{w}}^{10}_{\min} + \alpha\left(\bar{\boldsymbol{w}}^{10}_{\max} - \bar{\boldsymbol{w}}^{10}_{\min}\right). \qquad (4)$$

For any $\alpha \in (0, 1)$, $\mathbf{w}(\alpha)$ gives a noise pattern with intermediate correlation level between $\bar{\boldsymbol{w}}^{10}_{\min}$ and $\bar{\boldsymbol{w}}^{10}_{\max}$. To verify that this interpolation yields points close to the 10 clusters, we computed the mean distance of $\bar{\boldsymbol{w}}^{10}_{\ell}$s (excluding $\bar{\boldsymbol{w}}^{10}_{\min}$ and $\bar{\boldsymbol{w}}^{10}_{\max}$) from the interpolating line, normalized by the length between $\bar{\boldsymbol{w}}^{10}_{\min}$ and $\bar{\boldsymbol{w}}^{10}_{\max}$. The normalized mean distance of 0.11 indicates that the line is fairly close to the cluster centers. We also project all $\bar{\boldsymbol{w}}^{10}_{\ell}$s on the interpolating line and compute the distance between every two neighboring projected points on that line. The standard deviation of the gaps is around 0.08 of the length, which implies $\bar{\boldsymbol{w}}^{10}_{\ell}$s are reasonably scattered along the line. In sum, the noise patterns given by (4) provides a reasonable approximation for realistic noise patterns.

2.2 Test Clip Generation

The visual quality of a clip depends both on the level of noise and the spatio-temporal characteristics of the video. Therefore, we need to include videos that represent a broad spectrum of spatial and temporal complexities in our test. A simple way to evaluate the spatial/temporal complexity is to measure the average size of their I frames and P frames. However, videos with large I frames tend to have large P frames because high spatial complexity also gives rise to more prediction residuals in motion compensation. To decouple the correlation between I frame size and P frame size, we normalize the size of P frame by the I frame size and use $\frac{\text{P frame size}}{\text{I frame size}}$ as the indicator for the temporal complexity. We selected 3226 video clips with 4K resolutions from YouTube such that their encoding bit rates are all higher than 100 Mbps, hence ensuring that they contain minimal noise at source. We then encoded them using the H.264 encoder with FFmpeg. To make the frame size a better indicator for spatial-temporal complexity, we apply a constant quantization parameter of 28 to all videos. The distribution of spatial-temporal features is shown in Fig. 2. It is seen that the values of the I frame sizes and the normalized P frame sizes are scattered widely and loosely coupled.

We partition the space of I frame size and normalized P frame size using the percentile of their marginal distributions, respectively. In particular, we calculated the 33% and 66% percentile of the marginal distribution of I frame size and the 50% percentile of the normalized P frame size, respectively. Then we partition the space of spatial-temporal complexity using these percentiles into six regions as shown in Figure 2. In each region, we selected 20 videos that are closest to the centroid of the region (shown by the ▽ markers in Figure 2). We manually reviewed each of these 20 clips and selected one video that was free from noise and other artifacts such as out-of-focus and over-exposure (shown by △ markers in Figure 2). The chosen videos are shown to be close to the centroid of the respective regions and thus are representative of a wide range of spatial-temporal complexity. From each video, we extract a 10-second clip and resize the video from 4K to 1080p resolution using Ffmpeg's scaling filter. The scaled videos are saved as Y'UV raw videos without further compression. The scaling is necessary because we would like to show two video clips side by side on our 4K display. A brief description of the selected clips are given in Table. **??**.

Table 1: A brief description of the video clips in our database.

Name	Description
beach	shot from a airplane, beach landscape.
goose	still camera, gooses in front of a lake.
singer	handhold camera, a singer in a concert.
snow	still camera, snow storm over a forest.
sony	still camera, a SONY smartphone.
talkingman	handhold camera, a man is talking.

We then employ the AR model (1) to synthesize noise. We considered 3 models, $\mathbf{w}(0)$, $\mathbf{w}(0.5)$ and $\mathbf{w}(1)$, where $\mathbf{w}(\cdot)$ is defined in (4). The noise synthesized using the three parameters thus present low, medium, and high noise correlations, respectively. For each noise pattern, we select 4 different σ to create 4 test clips with different noise strength. We manually checked the 4 clips to make sure their perceptual quality is visually distinguishable and covers a wide range. In total, we created $3 \times 4 = 12$ distorted versions for each

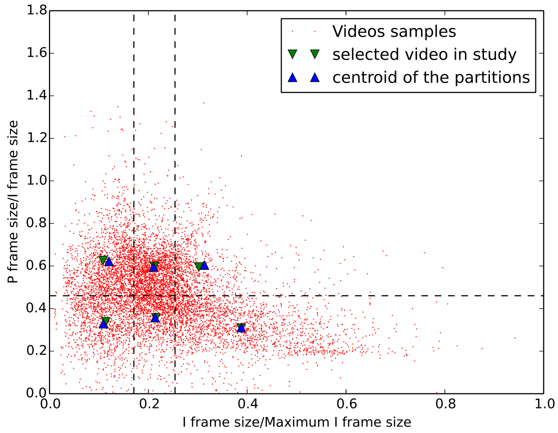


Figure 2: The joint distribution of I frame size and $\frac{\text{P frame size}}{\text{I frame size}}$ of 3226 high quality 4K videos uploaded to YouTube.

video. Because the duration of a test session is limited to be less than 30 minutes, 9 randomly sampled examples from each set of 12 clips were used for our subjective test [12]. Thus, we used $6 \times 9 = 54$ clips in total.

2.3 Subjective Experiment

We conducted a Double Stimulus Continuous Quality Scale (DSCQS) subjective test following the ITU guidelines [12]. We use a 55 inch Samsung TV for our test, and the participants sit 2.0 m away from the screen. The screen of the 3840 x 2160 display was equally split into two parts. The original video and the noisy video were synchronously played in the two parts such that the subjects could view and compare the two clips easily. There were 28 subjects involved in the test. After each pair of clips had been played, the subject was asked to grade the test video using the original video as a reference.

Fig. 3 shows the Mean Opinion Scores (MOS) obtained from the subjective test versus the power of the synthesized noise in each test video. It can be seen that the MOS obtained in our tests covers a wide range of values. The plot also demonstrates that, as the power of noise increases, the perceptual quality tend to decrease. However, the noise power itself cannot provide a good prediction for MOS. The rank-order correlation between noise power and MOS is only 0.755. That is because the visibility of noise depends on the video contents. The MOS of the videos with dark scenes such as ‘talkingman’ is lower than videos with brighter contents such as ‘snow’. The videos with complex textures or motions such as ‘beach’ and ’goose’ shows better MOS than other videos.

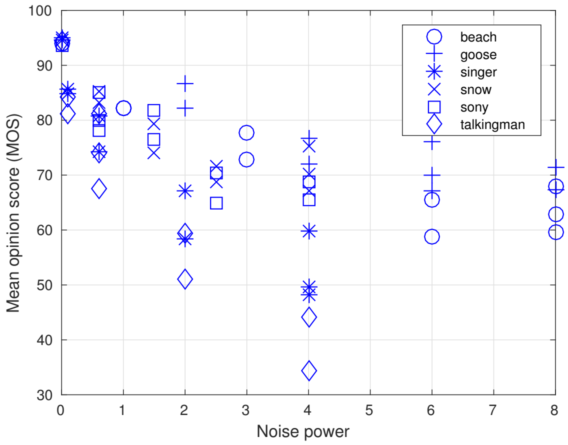


Figure 3: The Mean Opinion Scores of the test videos.

We also analyzed the relation between noise correlation and MOS. The average MOS for the test clips distorted by noise of low, medium, and high correlation are 76.11, 70.88, and 68.7, respectively. It seems videos with highly correlated noise tend to have worse quality.

In sum, the perceptual quality of noisy video depends not only on noise power but also on video contents and the structure of noise. In the next section, we present a no-reference quality metric that can accurately predict perceptual qualities by taking into account the noise pattern and video characteristics.

3. PERCEPTUAL QUALITY METRIC

The flow chart of the proposed no-reference perceptual quality assessment algorithm is shown in Fig. 4. The noise is first estimated from an input video using the method proposed in [13]. Then, the Power Spectral Densities (PSD) of the noise and the input video are estimated in the frequency domain, respectively (module ① and ② in Fig. 4). For a given frequency channel, the ratio between the PSD of the noise and that of the input video (biased by a regularizing constant α) are considered as the noise features of that frequency channel (module ③). These features are further adjusted according to the brightness of the video to incorporate contrast sensitivity (module ④). A visual importance pooling method is applied to the adjusted features to estimate the visibility of noise in each frequency (module ⑤). The final quality metric is then calculated by summing the estimated visibility over the frequencies that show high correlation with subjective quality (module ⑥). Details follow.

3.1 Noise Feature Extraction

We denote by $\{\mathrm{v}(x,y,t) : 0 \le x < W, 0 \le y < H, t \in \mathbb{N}^+\}$ a noisy video signal with resolution $W \times H$. To analyze the local characteristics of the video, we partition the video frames into $B \times B$ non-overlapping blocks and denote by $\{\mathrm{v}^{p,q}(x,y,t) = \mathrm{v}(Bp+x, Bq+y, t) : 0 \le x, y < B\}$ be the (p,q)'th block of a frame.

We model $\mathrm{v}^{p,q}(x,y,t)$ as the sum of a noise-free video block $\mathrm{u}^{p,q}(x,y,t)$ and a noise signal $\mathrm{n}(x,y,t)$, i.e.,

$$\mathrm{v}^{p,q}(x,y,t) = \mathrm{u}^{p,q}(x,y,t) + \mathrm{n}^{p,q}(x,y,t), \tag{5}$$

where $\mathrm{n}^{p,q}(x,y,t)$ is a realization of a zero-mean random variable $\mathsf{N}(x,y)$.

As shown in [4, 7, 19], the visual cortex of human brain decomposes the received visual signal into different orientations and frequencies. In fact, as shown in [10, 24], an overcomplete wavelet transform can be used to imitate such behavior and extract spatial-temporal features for visual quality assessment. In this paper, we approximate the frequency decomposition in visual cortex using the 3D Discrete Fourier Transform (3D-DFT) and propose to use the ratio of power spectrum density (PSD) of $\mathrm{n}^{p,q}(\cdot)$ and $\mathrm{v}^{p,q}(\cdot)$ as the noise feature.

For a discrete video signal $\{\mathrm{f}(x,y,t) : 0 \le x, y < B, 0 \le t < T\}$, its 3D-DFT is given by

$$\hat{\mathrm{f}}(\ell_x, \ell_y, \ell_t) = \sum_{t=0}^{T-1} \omega_T^{\ell_t t} \left(\sum_{y=0}^{B-1} \omega_B^{\ell_y y} \left(\sum_{x=0}^{B-1} \mathrm{f}(x,y,t) \omega_B^{\ell_x x} \right) \right).$$

Here, $\omega_B = \exp(-2\pi\mathrm{i}/B)$ and $\omega_T = \exp(-2\pi\mathrm{i}/T)$ are spatial and temporal discrete Fourier basis, respectively. The PSD of $\mathrm{f}(x,y,t)$ [1] is defined by

$$\mathrm{S_{ff}}(\ell_x, \ell_y, \ell_t) = |\hat{\mathrm{f}}(\ell_x, \ell_y, \ell_t)|^2. \tag{6}$$

PSD captures the signal energy corresponding to spatial frequency $(2\pi\ell_x/B, 2\pi\ell_y/B)$ and temporal frequency $2\pi\ell_t/T$. We define the following distortion feature:

$$\mathrm{r}(\ell_x, \ell_y, \ell_t) = \frac{\mathrm{S_{nn}}(\ell_x, \ell_y, \ell_t)}{\mathrm{S_{vv}}(\ell_x, \ell_y, \ell_t) + \alpha}, \tag{7}$$

[1] Strictly speaking, $\mathrm{S_{ff}}$ should be called periodgram.

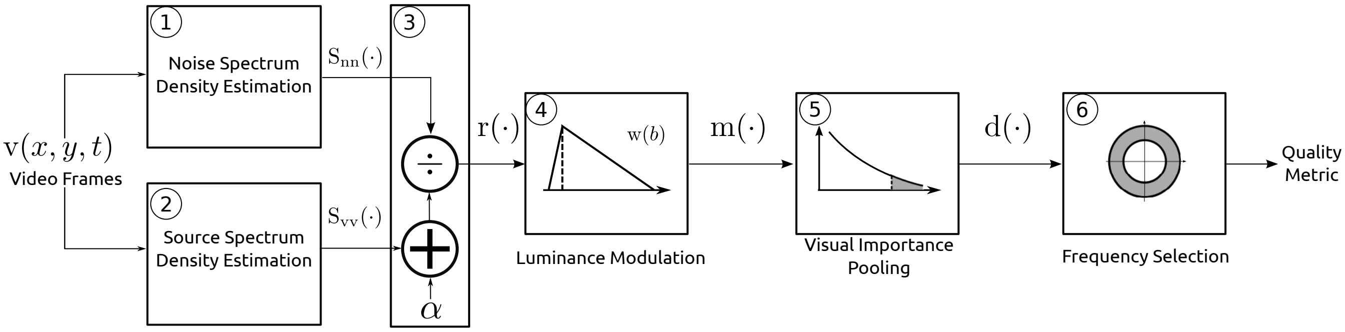


Figure 4: The flow chart of the proposed noisy video quality metric.

where $\mathrm{S_{nn}}$ and $\mathrm{S_{vv}}$ are the PSDs for noise $\mathrm{n}(x, y, t)$ and video signal $\mathrm{v}(x, y, t)$, respectively. The constant α is employed to stabilize potential numerical errors. The feature $\mathrm{r}(\ell_x, \ell_y, \ell_t)$ therefore gives the relative noise strength at a given spatial-temporal frequency.

Another motivation for using $\mathrm{r}(\cdot)$ as the distortion feature is that it captures the correlation pattern of noise. In fact, it can be shown that the PSD of a signal equals to the Fourier transform of its auto-correlation function. The correlation pattern of noise is captured by $\mathrm{S_{nn}}$ and thus by $\mathrm{r}(\cdot)$.

In our implementation, following the parameters used in [13], we fix the blocksize to be $B = 32$ and the temporal length to be $T = 3$. For each block in a video frame, we concatenate the block with the two co-located blocks in the adjacent frames to form a 3-D array, i.e., $\{\mathrm{v}^{p,q}(x, y, t-1), \mathrm{v}^{p,q}(x, y, t), \mathrm{v}^{p,q}(x, y, t+1) : 0 \leq x, y < 32\}$. Then the 3D-FFT is applied to the array to obtain $\{\mathrm{S}^{p,q,t}_{\mathrm{vv}}(\ell_x, \ell_y, \ell_t) : 0 \leq \ell_x, \ell_y < 32, 0 \leq \ell_t < 3\}$. The noise PSD $\mathrm{S_{nn}}(\ell_x, \ell_y, \ell_t)$ are estimated with the method proposed in [13]. Then, using (7), we get the distortion features $\{\mathrm{r}^{p,q,t}(\ell_x, \ell_y, \ell_t) : 0 \leq \ell_x, \ell_y < 32, 0 \leq \ell_t < 3\}$, which is a $32 \times 32 \times 3$ array. The stabilizing constant α is fixed to be 0.3.

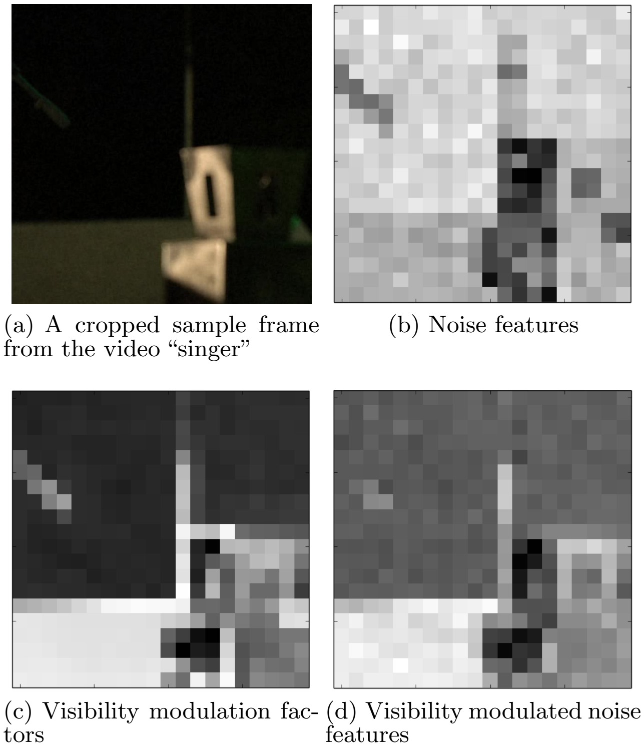
3.2 Visibility Modulation

The visibility of noise depends on the background luminance. As shown in [5, 10], in the dark regions of a video, the noise is much less visible than in the regions of mid-grey brightness. On the other hand, for regions of higher than mid-grey luminance, the noise visibility decreases with luminance, which follows Weber's law. To incorporate the impact of luminance, we modulate the distortion feature $\mathrm{r}^{p,q,t}(\cdot)$ by a function of local luminance. Assuming the luminance of pixels are normalized to $[0, 1]$, we define the following function that approximately maps brightness to noise visibility:

$$\mathrm{w}(b) = \begin{cases} b/\beta & 0 \leq b \leq \beta, \\ 1 - (b-\beta)/(1-\beta) & \beta < b \leq 1. \end{cases} \tag{8}$$

It is a piecewise linear function which connects points $(0, 0)$, $(\beta, 1)$, and $(1, 0)$. The function $\mathrm{w}(b)$ is defined so as to capture the relationship between error visibility threshold and background luminance given by [5]. The parameter β is set to 0.15 to approximate the luminance under which the noise is most visible. Letting $b^{p,q,t} = \sum_{x,y} \mathrm{v}(x, y, t)/B^2$ be the average luminance of the (p, q)'th block, the visibility modulated noise feature of the block is given by

$$\mathrm{m}^{p,q,t}(\ell_x, \ell_y, \ell_t) = \mathrm{r}^{p,q,t}(\ell_x, \ell_y, \ell_t)\, \mathrm{w}\left(b^{p,q,t}\right). \tag{9}$$

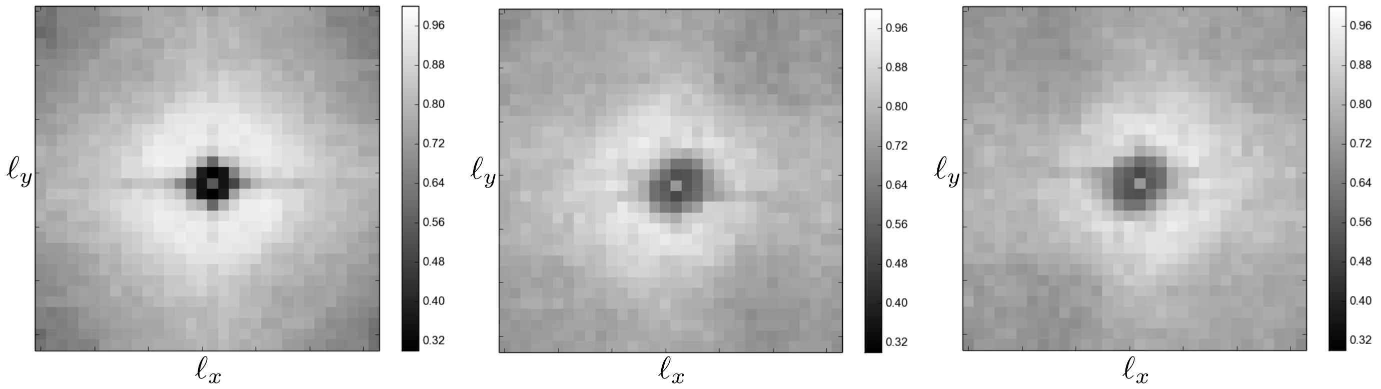

(a) A cropped sample frame from the video "singer"

(b) Noise features

(c) Visibility modulation factors

(d) Visibility modulated noise features

Figure 5: (a) A cropped sample frame from the test clip "singer". (b) The average noise feature $\frac{1}{B^2T}\sum_{\ell_x,\ell_y,\ell_t} \mathrm{r}(\ell_x, \ell_y, \ell_t)$ for all the 32×32 blocks; (c) The average visibility modulation factor $\frac{1}{B^2T}\sum_{\ell_x,\ell_y,\ell_t} \mathrm{w}(\ell_x, \ell_y, \ell_t)$ for all blocks; (d) The average modulated noise feature $\frac{1}{B^2T}\sum_{\ell_x,\ell_y,\ell_t} \mathrm{m}(\ell_x, \ell_y, \ell_t)$ for all blocks. Brighter color indicate larger value.



(a) $\ell_t = 0$ (b) $\ell_t = 1$ (c) $\ell_t = 2$

Figure 6: The absolute value of the rank-order correlation coefficient between $\mathrm{d}(\ell_x, \ell_y, \ell_t)$ with the MOS obtained from the subjective test. The coefficients are rearranged such that $\ell_x = 0, \ell_y = 0$ is located at the center of image.

We show an example of the effect of this visibility modulation effect in Fig. 5. Although equal amount of noise is added everywhere in the frame shown in Fig. 5(a), the noise in the dark region is not quite visible. Fig. 5(b) shows that the perceptual noise level is overestimated by $\mathrm{r}(\ell_x, \ell_y, \ell_t)$ in the dark regions. By modulating the noise feature by $\mathrm{w}(\cdot)$ (see Fig. 5(c)), we can obtain a better estimation for the perceptual noise level (see Fig. 5(d)).

3.3 Visual Importance Pooling

In [17, 21], it has been recognized that the worst part of a video tends to attract more attention of the viewers and thus dominates the overall quality of the video. This fact has been exploited by a variety of visual importance pooling algorithms such as [17] and [22]. In our work, we applied a similar pooling method as [17]. For each frequency channel (ℓ_x, ℓ_y, ℓ_z), we first group the luminance modulated noise feature across all blocks, i.e., $\{\mathrm{m}^{p,q,t}(\ell_x, \ell_y, \ell_t) : \forall p, \forall q\}$. Then we calculate the p% percentile of the features, denoted by $\mathrm{k}^t(\ell_x, \ell_y, \ell_t)$. Finally, we obtain the average distortion score for frequency (ℓ_x, ℓ_y, ℓ_z) as

$$\mathrm{d}(\ell_x, \ell_y, \ell_t) = \frac{1}{\mathrm{F}} \sum_{t=1}^{\mathrm{F}} \mathrm{k}^t(\ell_x, \ell_y, \ell_t). \tag{10}$$

where F is the number of frames of the video. In our implementation, the parameter p for percentile pooling is set to 80%.

3.4 Frequency Selection

The final step of our algorithm is to combine the distortion scores $\mathrm{d}(\cdot)$ for different spatial-temporal frequencies to predict MOS. For all the test clips in our subjective test, we calculated their distortion scores $\mathrm{d}(\ell_x, \ell_y, \ell_t)$ for each frequency and then computed its rank-order correlation coefficients with the MOS obtained in our subjective test. The results are shown in Fig 6. It shows that for each temporal frequency ℓ_t, the rank-order correlation coefficients are roughly concentric around the frequency $\ell_x = 0, \ell_y = 0$. In other words, the level of correlation mainly depends on the radial frequency $\ell_r = \sqrt{\ell_x^2 + \ell_y^2}$. At medium radial frequency, $\mathrm{d}(\ell_x, \ell_y, \ell_t)$ achieves very strong correlation with MOS (more than 0.9).

Based on the above observations, we use the sum of distortion scores on selected frequencies to predict the perceptual quality. The overall distortion metric is given by:

$$\mathrm{D} = \sum_{t=0}^{T} \sum_{\ell_x=0}^{B} \sum_{\ell_y=0}^{B} \mathrm{d}(\ell_x, \ell_y, \ell_t) \mathrm{I}^{\ell_t}(\ell_x, \ell_y), \tag{11}$$

where $\mathrm{I}^{\ell_t}(\ell_x, \ell_y)$ is a binary frequency selection function. Because $\mathrm{d}(\ell_x, \ell_y, \ell_t)$ achieves maximum correlation with perceptual quality at medium radial frequencies, we use a simple "band-pass" frequency selection function as follows:

$$\mathrm{I}^{\ell_t}(\ell_x, \ell_y) = \begin{cases} 1 & \Psi_{\ell_t} \leq \sqrt{\ell_x^2 + \ell_y^2} \leq \Phi_{\ell_t}, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where the parameters Ψ_{ℓ_t} and Φ_{ℓ_t} defines the pass-band for the ℓ_t'th temporal frequency plane. In our implementation, we choose $\Psi_0 = \Psi_1 = \Psi_2 = 4$, $\Phi_0 = 8$, and $\Phi_1 = \Phi_2 = 6$. These parameters are optimized using out database.

4. PERFORMANCE EVALUATION

We evaluated the proposed metric on the videos tested in our subjective test. We employ a simple linear model to map our quality metric D to the MOS. The fitted model is

$$\hat{\mathrm{MOS}}(D) = -429.7171D + 90.9003.$$

Our metric gives accurate predictions for perceptual quality. The linear correlation coefficient and the rank order correlation coefficient between the predicted MOS and the MOS is 0.9490 and 0.9518, respectively.

We also compared the performance of our metric with several widely used video quality metrics, including PSNR, SSIM, VQM, and STMAD. The comparison is based on three measurements: 1) rank order correlation coefficients (SROCC); 2) linear correlation coefficients (LCC) and variance of residuals (VoR) between predicted MOS and the empirical MOS. For each metric, we employ the following monotonic logistic regression model recommended by ITU

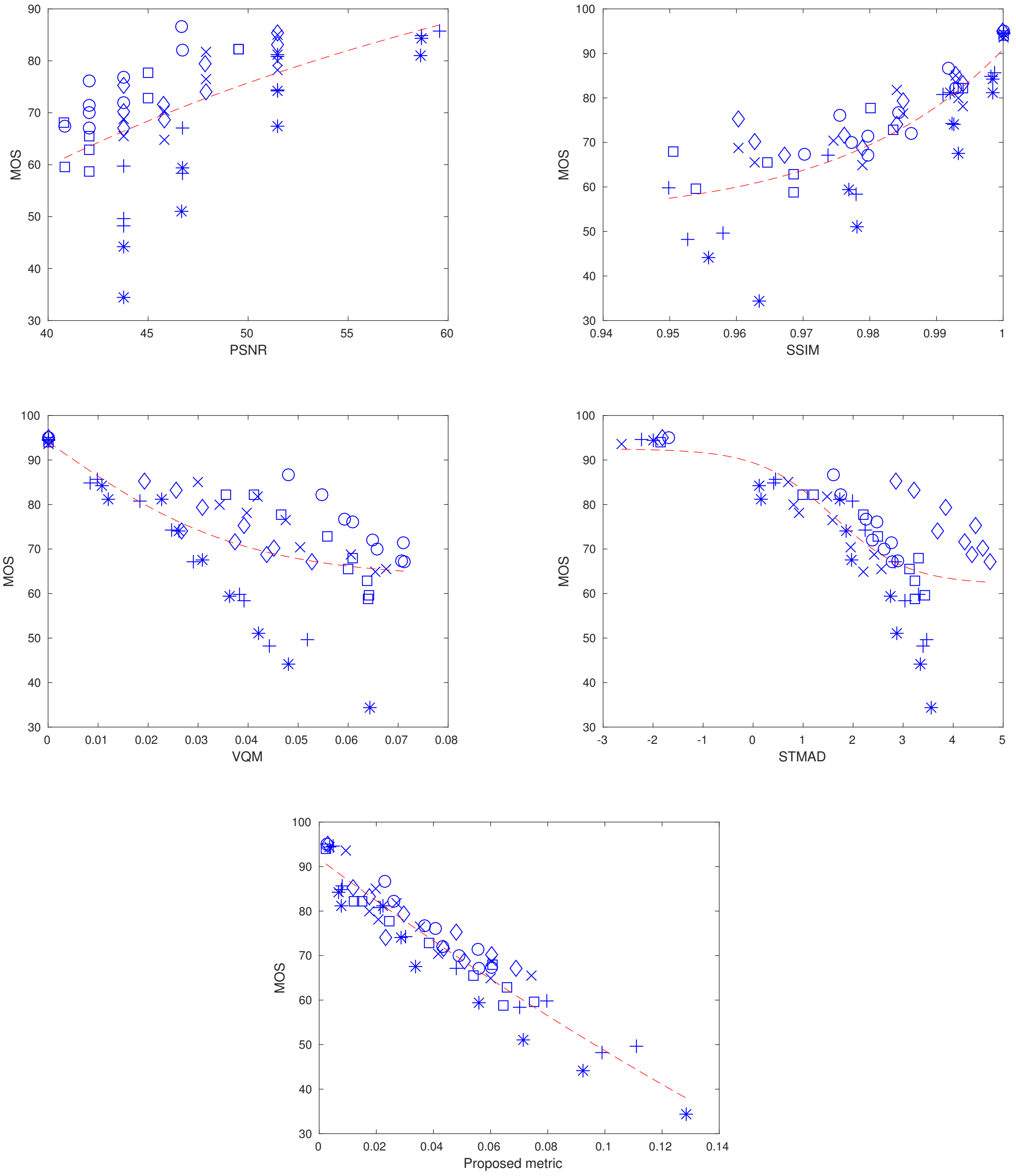


Figure 7: The regression results of PSNR, SSIM, VQM, STMAD, and the proposed noise metrics. The dashed lines show the fitted sigmoid model (13) for each metric. Different markers represents different video content. ○: 'goose', +: 'singer', ∗: 'talkingman', ×: 'sony', □: 'beach', ◇: 'snow'.

Table 2: Performance comparison with other full-reference quality metrics.

Quality Metrics	SROCC	LCC	VoR
Proposed	**0.9518**	**0.9497**	**16.4933**
PSNR	0.6382	0.5838	86.4300
SSIM	0.8560	0.8287	52.7269
VQM	0.6707	0.6917	87.7917
STMAD	0.7412	0.7533	72.8135

Table 3: Performance of the proposed metric on the testing sets when 60% of the video database is used for training.

Quality Metrics	SROCC	LCC	VoR
Proposed	**0.9417**	**0.9545**	**16.8952**
PSNR	0.7019	0.6549	86.6038
SSIM	0.8390	0.8103	65.0248
VQM	0.7450	0.7108	89.3259
STMAD	0.7100	0.7236	87.9041

in [11]:

$$\mathrm{MOS}^{\mathrm{pred}} = \frac{a_1 - a_0}{1 + \left(\frac{M + a_4}{a_2}\right)^{a_3}} + a_0, \qquad (13)$$

where M is the score given by a video quality metric and $\mathrm{MOS}^{\mathrm{pred}}$ is the MOS predicted by the quality metric. The function parameters $a_0 \cdots a_4$ are obtained by minimizing the mean square error of $\mathrm{MOS}^{\mathrm{pred}}$. The results are shown in Fig. 7. The SROCC, LCC, and VoR of all the metrics are summarized in Table 2. The proposed algorithm outperforms all the other quality metrics in all the performance measurements. Note that PSNR, SSIM, VQM and STMAD all need to have access to the original noise-free video, which is typically unavailable for video sharing services such as YouTube. The proposed metric does not need the original video and is thus more appealing for consumer generated video contents.

The parameters of our algorithm such as the numerical stabilizing constant α, visual importance pooling parameter p and frequency selection parameters Φ, Ψ are tuned on our video database. To check if the database is large enough for training the algorithm parameters, we conducted cross-validations on our database. Each time, we partition our database into a training set and a test set. We tune the parameters of our algorithm on the training set and evaluate the VoR on the test set. We vary the size of the training sets from 20% to 90% of the whole database. For a given training set size, cross-validations are repeated on 1000 random training/test set partitions and the average VoR is then calculated and shown in Fig. 8. It shows that the VoR decreases quickly as the size of training set increases. When the size of training set covers more than 60% of the database, the average VoR does not decrease further. In other words, using 60% of our database is enough for tuning the parameters of our model and the model parameter trained on our database is reliable.

To further verify out algorithm, we applied the algorithm to 1% of the videos uploaded to YouTube. From these videos, we randomly select 7 videos such that their MOS predicted by the proposed algorithm is around 30, 40, . . . , 90, respectively. We found that their perceptual quality agree well with the predicted MOS.

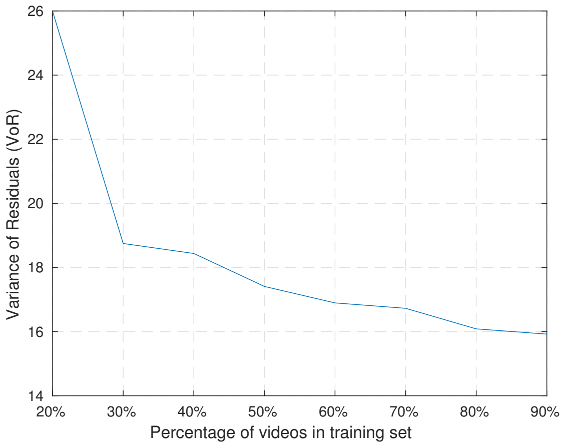


Figure 8: The variation of residuals (VoR) on the test set versus the size of the training set in cross validations.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a no-reference quality assessment algorithm for video distorted by noise. The algorithm captures the correlation structure of noise by using an estimated PSD. It also incorporated the impact of luminance and visual importance on the noise visibility in its design. The algorithm is trained and validated on a new database of noisy videos that simulates the typical noise patterns encountered in consumer-generated video content. The proposed algorithm can achieve up to 0.95 linear correlation with the MOS obtained from subjective studies, outperforming several high-performance video quality metrics.

The database we used for developing our metric only involves the videos distorted by noise. In practice, videos could be distorted by different types of artifacts in the same time. In the future, we plan to combine our noise metric with other artifacts detectors to predict the quality of videos suffering from multiple types of distortions.

6. REFERENCES

[1] O. K. Al-Shaykh and R. M. Mersereau. Restoration of lossy compressed noisy images. *IEEE Transactions on Image Processing*, 8(10):1348–1360, Oct 1999.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag Inc., Secaucus, NJ, USA, 2006.

[4] A. C. Bovik. Automatic prediction of perceptual image and video quality. *Proceedings of the IEEE*, 101(9):2008–2024, Sept 2013.

[5] C.-H. Chou and Y.-C. Li. A perceptually tuned subband image coder based on the measure of

just-noticeable-distortion profile. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6):467–476, Dec 1995.

[6] W. Collis, S. Robinson, and P. White. Synthesising film grain. In *European Conference on Visual Media Production*, pages 231–234, 2004.

[7] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2:1160–1169, July 1985.

[8] A. De Stefano, B. Collis, and P. White. Synthesising and reducing film grain. *Journal of Visual Communication and Image Representation*, 17(1):163–182, 2006.

[9] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, February 2010.

[10] B. Girod. The information theoretical significance of spatial and temporal masking in video signals. In *Proc. SPIE Human Vision, Visual Processing, and Digital Display*, pages 178–189, 1989.

[11] ITU. Objective perceptual assessment of video quality: Full reference television. https://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf, 2004.

[12] ITU-R Recommendation BT.500-13. Methodology for the subjective assessment of the quality of television pictures. http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf, Jan. 2012.

[13] A. Kokaram, D. Kelly, H. Denman, and A. Crawford. Measuring noise correlation for improved video denoising. In *IEEE International Conference on Image Processing (ICIP)*, Sept 2012.

[14] C. L. Lim and R. Paramesran. Blind image quality assessment for color images with additive gaussian white noise using standard deviation. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 039–041, Dec 2014.

[15] M. Liu, G. Zhai, Z. Zhang, Y. Sun, K. Gu, and X. Yang. Blind image quality assessment for noise. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–5, June 2014.

[16] A. Mittal, M. A. Saad, and A. C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, Jan 2016.

[17] A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):193–201, April 2009.

[18] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec 2011.

[19] B. A. Olshausen and D. J. Field. How close are we to understanding v1? *Neural Computation*, 17(8):1665–1699, 2005.

[20] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time Signal Processing (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[21] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik. Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, 22(2):610–620, Feb 2013.

[22] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322, Sept 2004.

[23] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, Aug 2012.

[24] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, Feb 2010.

[25] T. Seybold, C. Keimel, M. Knopp, and W. Stechele. Towards an evaluation of denoising algorithms with respect to realistic camera noise. In *IEEE International Symposium on Multimedia (ISM)*, pages 203–210, Dec 2013.

[26] J. Shen, Q. Li, and G. Erlebacher. Hybrid no-feference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images. *IEEE Transactions on Image Processing*, 20(8):2089–2098, Aug 2011.

[27] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.

[28] P. V. Vu, C. T. Vu, and D. M. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *IEEE International Conference on Image Processing (ICIP)*, pages 2505–2508, Sept 2011.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[30] G. Zhai, A. Kaup, J. Wang, and X. Yang. A dual-model approach to blind quality assessment of noisy images. In *Picture Coding Symposium (PCS)*, pages 29–32, Dec 2013.

[31] G. Zhai and X. Wu. Noise estimation using statistics of natural images. In *IEEE International Conference on Image Processing (ICIP)*, pages 1857–1860, Sept 2011.

[32] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1):41–52, Jan 2012.

[33] K. Zhu, C. Li, V. Asari, and D. Saupe. No-reference video quality assessment based on artifact measurement and statistical analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(4):533–546, April 2015.

[34] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. In *IEEE International Conference on Computer Vision*, pages 2209–2216, Sept 2009.