Blind Prediction of Natural Video Quality

Michele A. Saad, Alan C. Bovik, *Fellow, IEEE*, and Christophe Charrier, *Member, IEEE*

***Abstract*— We propose a blind (no reference or NR) video quality evaluation model that is nondistortion specific. The approach relies on a spatio-temporal model of video scenes in the discrete cosine transform domain, and on a model that characterizes the type of motion occurring in the scenes, to predict video quality. We use the models to define video statistics and perceptual features that are the basis of a video quality assessment (VQA) algorithm that does not require the presence of a pristine video to compare against in order to predict a perceptual quality score. The contributions of this paper are threefold. 1) We propose a spatio-temporal natural scene statistics (NSS) model for videos. 2) We propose a motion model that quantifies motion coherency in video scenes. 3) We show that the proposed NSS and motion coherency models are appropriate for quality assessment of videos, and we utilize them to design a blind VQA algorithm that correlates highly with human judgments of quality. The proposed algorithm, called video BLIINDS, is tested on the LIVE VQA database and on the EPFL-PoliMi video database and shown to perform close to the level of top performing reduced and full reference VQA algorithms.**

***Index Terms*— Video quality assessment, discrete cosine transform, egomotion, generalized Gaussian.**

I. Introduction

TODAY'S technology permits video content to be ubiquitously created, stored, transmitted, and shared between users on a multitude of devices ranging from hand-held PDAs and tablets, to very large high definition screens. Video content is being transmitted in exponentially increasing volumes via wireless and wired networks. The limited availability of bandwidth, and the physical properties of the transmission media and capture and display devices means that some information from the original source is likely to be lost. It is, however, important that the perceived visual quality at the end-user be maintained at an acceptable level, given rising consumer expectations of the quality of multimedia content delivered to them.

Image and video quality assessment (I/VQA) researchers have been working to understand how distortions introduced throughout the lossy path between the source and destination affect the statistics of multimedia signals and how these distortions affect perceived signal quality. The most accurate way to assess the quality of an image or a video is to collect the opinions of a large number of viewers of the image/video in the form of *opinion scores* that rate the visual quality of the image or video. These opinion scores are then averaged (usually after normalization with respect to each individual's score average). This average is known as the *mean-opinion-score* (MOS), and the overall process is referred to as subjective I/VQA. While subjective I/VQA is cumbersome, expensive, impractical and for many important applications infeasible (e.g. for real-time monitoring of video quality in a network), it is valuable for providing ground truth data for the evaluation of objective I/VQA algorithms.

Objective I/VQA refers to models that seek to predict the visual quality of a signal automatically, in the absence of human raters. Objective quality assessment methods fall into three categories: 1) full-reference (FR), 2) reduced-reference (RR), and 3) blind or no-reference (NR) approaches.

FR-I/VQA refers to I/VQA models that require the presence of a reference signal to predict the quality of a test signal. FR-IQA models now exist that achieve excellent levels of performance, as demonstrated by high correlations with human subjective judgments of visual quality. SSIM [1], MS-SSIM [2], VSNR [3], MAD [4], and the VIF index [5] are examples of successful FR-IQA algorithms. Prominent FR-VQA algorithms include MOVIE [6], VSSIM [7], VQM [8], DVQ [9], Tetra VQM [10], ST-MAD [11], and the work in [12] and [13]. These methods require the availability of a reference video against which to compare the test signal. In many applications, however, the reference is not available to perform a comparison against, which severely limits the application domain of FR-IQA algorithms.

RR-I/VQA refers to I/VQA models that require partial information about the reference signal in order to predict the quality of a test signal. Successful RR-I/VQA algorithms include the wavelet-based RR-IQA algorithm in [14], the divisive normalization transform-based RR-IQA algorithm in [15], the information theoretic RRED index in [16], and the wavelet-based RR-VQA method in [17].

NR-I/VQA models have potentially much broader applicability that FR and RR models since they can predict a quality score in the absence of a reference image/video or any specific information about it. The problem of "blindly" assessing the visual quality of images and videos requires dispensing with older ideas of quality such as fidelity, similarity, and metric comparison. Only recently have NR-IQA algorithms been devised that correlate highly with human judgments of quality.

Manuscript received October 10, 2012; revised June 25, 2013; accepted December 16, 2013. Date of publication January 9, 2014; date of current version February 6, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary Comer.

M. A Saad was with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA. She is now with Intel Corporation, Austin, TX 78746 USA (e-mail: michele.saad@utexas.edu).

A. C. Bovik is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

C. Charrier is with the Department of Electrical and Computer Engineering, The University of Caen, Saint-Lô 50000, France (e-mail: christophe.charrier@unicaen.fr).

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

Digital Object Identifier 10.1109/TIP.2014.2299154

Some are distortion-specific, i.e. they quantify one or more specific distortions such as blockiness [18], blur [19], [20], or ringing [21] and score the image accordingly. There are considerably fewer algorithms that work well across multiple classes of distortions. Examples of such NR-IQA approaches can be found in [22]–[25].

There are even fewer blind VQA algorithms than blind IQA algorithms. The problem is much more challenging owing to a lack of relevant statistical and perceptual models. Certainly, accurate modeling of motion and temporal change statistics in natural videos would be valuable, since these attributes play an important role in the perception of videos [26]–[28]. Indeed, considerable resources in the human visual system (HVS) are devoted to motion perception [26]–[28].

In [29] an H.264-specific algorithm was proposed that extracts transform coefficients from encoded bitstreams. A PSNR value is estimated between the quantized transform coefficients and the predicted non-quantized coefficients prior to encoding. The estimated PSNR is weighted using the perceptual models in [30] and [31]. The algorithm, however, requires knowledge of the quantization step used by the encoder for each macroblock in the video, and is hence not applicable when this information is not available. The authors of [32] propose a distortion-specific approach based on a saliency map of detected faces. However, this approach is both semantic dependent and distortion dependent.

There do not yet exist NR-VQA algorithms that have been shown to consistently correlate well with human judgments of temporal visual quality. Towards designing such a model, we have developed a framework that utilizes a spatio-temporal model of DCT coefficient statistics to predict quality scores. The attributes of this new blind VQA model are that it 1) characterizes the type of motion in the video, 2) models temporal as well as spatial video attributes, 3) is based on a model of natural video statistics, 4) is computationally fast, and 5) extracts a small number of interpretable features relevant to perceptual quality. Finally, we provide a Matlab implementation of the developed algorithm, which we have dubbed *Video BLIINDS* owing to its genesis from ideas on spatial IQA [25], which can be downloaded from the Laboratory of Image and Video Engineering (LIVE) website at `http://live.ece.utexas.edu/`.

The remainder of the paper is organized as follows. In Section 2 we describe the overall framework of the model. In Section 3 we discuss relevant attributes of motion and motion perception. In Section 4 we explain the temporal statistics model that underlies many of the features that are extracted for quality prediction. We also show how to assemble the overall quality prediction model there. In Section 5 we report and analyze experiment results, and we conclude in Section 6.

II. Algorithm Framework

We shall refer to pristine/undistorted videos that have not been subjected to distortions as *natural video scenes*, and statistical models built for natural video scenes as NVS (natural video statistics) models. Deviations from NVS models, caused by the introduction of distortions, can be used to predict the perceptual quality of videos. The study of the statistics of natural visual signals is a discipline within the field of perception. It has been shown that static natural scenes exhibit highly reliable statistical regularities. The general philosophy follows the premise that the human vision system has evolved in response to the physical properties of the natural environment [26], [28], and hence, the study of natural image statistics is highly relevant to understanding visual perception.

The field of NVS has not developed nearly as far as the study of still image statistics. Most authors have focused on trying to find models of optical flow statistics but with limited success [33], [34]. For example, the authors of [33] developed a limited model exhibiting regularities, but only under the assumption that the camera is in motion, yet no objects in the imaged scene move independently. Our own experiments on optical flow modeling have encountered similar difficulties, with some limited success on the perceptual side [35]. Yet, confident that the moving world does indeed exhibit statistical regularities, we have relied upon *Occam's Razer* and directed our modeling efforts to the simpler case of frame-differences only, where we have indeed found that regularities appear to exist, and more importantly, that these regularities are predictably disturbed by the presence of distortions. Thus, our approach to blind VQA design leverages the fact that natural, undistorted videos exhibit statistical regularities that distinguishes them from distorted videos where these regularities are destroyed. Specifically, we propose an NVS model of DCT coefficients of frame-differences.

The statistics of frame-differences have previously been explored. The authors of [36] found that frame-differenced natural videos reliably obey a (global) space-time spectral model. We have also found that a simple and regular local natural video statistic (NVS) model nicely describes filtered or transformed time-differential (or frame differenced) videos in the wavelet and DCT domains [25], [37].

Fig. 2 plots an example of the statistics of DCT coefficient frame differences. Specifically, the empirical probability distributions of frame difference coefficients (from 5×5 spatial blocks) in a pristine video and in a video distorted by a simulated wireless channel are shown. Fig. 1 shows a sample frame from the pristine and distorted videos corresponding to the distributions in Fig. 2. Notice how the distribution of the pristine video DCT coefficients is more heavy-tailed than that of the distorted video DCT coefficients. Examples similar to this one are consistently observed over a wide range of pristine and distorted videos [25], [37]. In Fig. 3 we show plots of the frame difference DCT coefficient histograms obtained from multiple frames of pristine and distorted videos. Similar histogram deviations are observed on the four distortions on which the algorithm was tested (MPEG-2 distortions, H.264 distortions, IP packet-loss, and wireless distortions).

The new blind VQA model is summarized in Fig. 5. A local 2-dimensional spatial DCT is applied to frame-difference-patches, where the term *patch* is used to refer to an $n \times n$ block of frame differences. This captures spatially and temporally local frequencies. The frequencies are spatially local since the

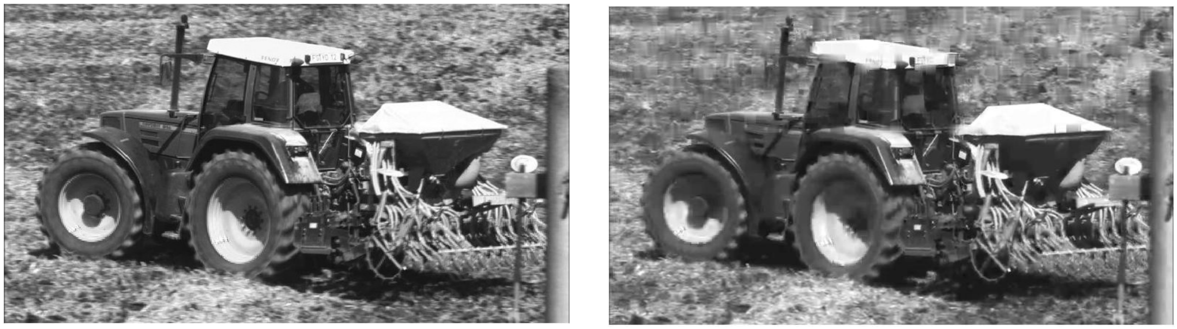


Fig. 1. Left: frame from pristine video. Right: frame from distorted video.

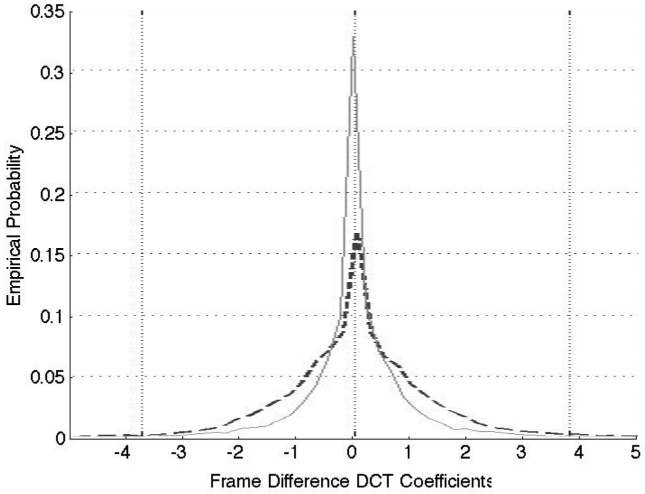


Fig. 2. Empirical probability distribution of frame-difference DCT coefficients of pristine and distorted videos. Dashed line: pristine video. Solid line: distorted video.

DCT is computed from $n \times n$ blocks, and they are temporally local since the blocks are extracted from consecutive frame differences. The frequencies are then modeled as generated from a specific family of probability density functions. It is observed that the parameters of this family differ for pristine and distorted videos. Fig. 4 is an example of how the parameters of the spatio-temporal NVS model vary according to the level of perceived distortion. It is a plot of one parameter (γ) of the NVS model (to be described in the following sections) for each frame-difference in three 250 fps, 10 second videos having three broadly different quality levels. It may be observed that γ decreases as the amount of perceived distortion in the video increases.

The interaction between motion and spatio-temporal change is of particular interest, especially with regards to whether motion is implicated in the masking of distortions. The type of motion which occurs in a video is a function of object and camera movement. In our model, image motion is characterized by a coherency measure which we define and use in conjunction with the parameters derived from the spatio-temporal NVS model of DCT coefficients. These features extracted under the spatio-temporal NVS model are then used to drive a linear kernel support vector regressor (SVR), which is trained to predict the visual quality of videos.

In this new model, the spatial and temporal dimensions of video signals are jointly analyzed and assessed. The behavior of a video is analyzed along the temporal dimension in two distinct ways: 1) By frame differencing: the statistics of frame differences are analyzed under the NVS model, and 2) By analyzing the types of motion occurring in the video and quantifying the motion in terms of a coherency measure.

III. Relevant Properties of Motion and Motion Perception

Both spatial and temporal distortions afflict videos. Examples of commonly occurring spatial distortions include blocking, ringing, false contouring, and blur. Blocking effects result from block-based compression techniques such as MPEG-1, MPEG-2, MPEG-4, and H.264. Ringing distortions are often visible around edges or contours of processed videos, manifesting as a rippling effect in the neighborhood of edges. Ringing occurs, for example, in wavelet based compression systems such as Motion JPEG-2000. False contouring arises from inadequate quantization. Blur is the loss of high frequency information and can occur as a result of compression-induced loss of high frequencies or as a by-product of the video acquisition system.

Many temporal distortions are highly annoying. Examples of commonly occurring temporal artifacts include *ghosting*, motion-compensation mismatch, jitter, mosquito noise, and stationary area fluctuations [38]. *Ghosting* appears as a blurred remnant trailing behind fast moving objects. Motion-compensation mismatch occurs as a result of the assumption that all constituents of a macroblock undergo identical motion shifts from one frame to another. Jitter may occur due to transmission delays in a network. Mosquito noise is a temporal artifact seen as fluctuations in smooth regions surrounding high contrast edges or moving objects, while stationary area fluctuations resemble the mosquito effect but occur in textured regions of scenes.

Temporal content and the type of motion occurring in videos plays a major role in the visibility of distortions and in the perception of the quality of dynamic image sequences. A major unresolved question affecting VQA model design

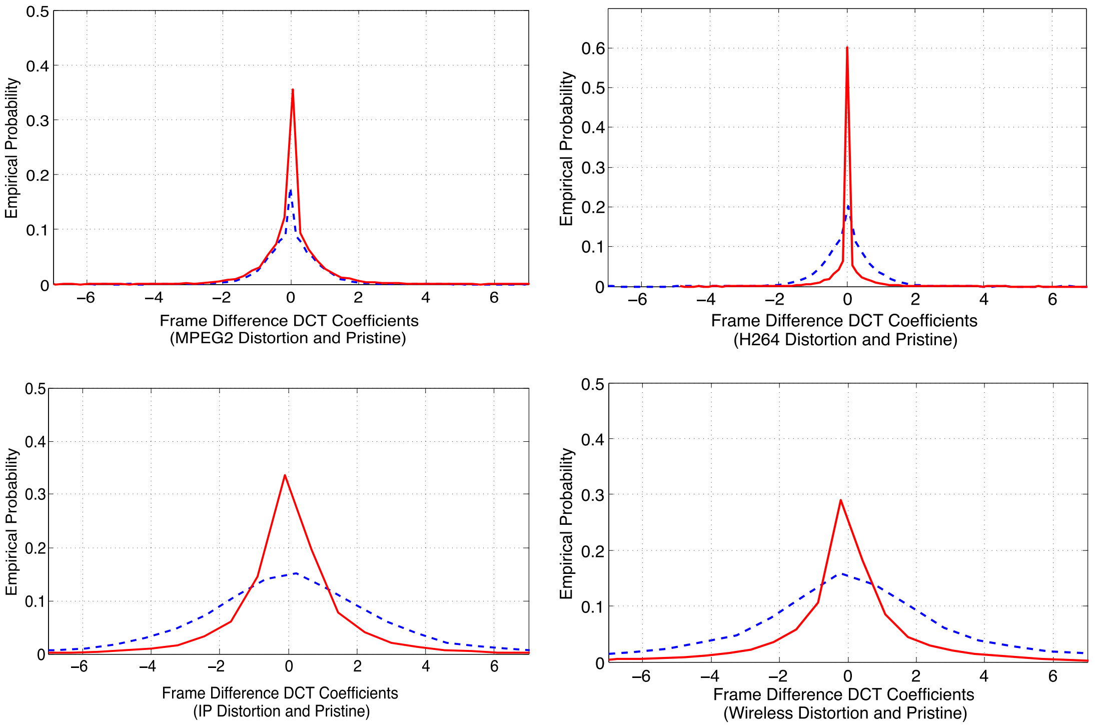


Fig. 3. Empirical probability distribution of frame-difference DCT coefficients of pristine and distorted videos for 4 distortions (MPEG-2, H.264, IP, and wireless distortions). Dashed line: pristine video. Solid line: distorted video.

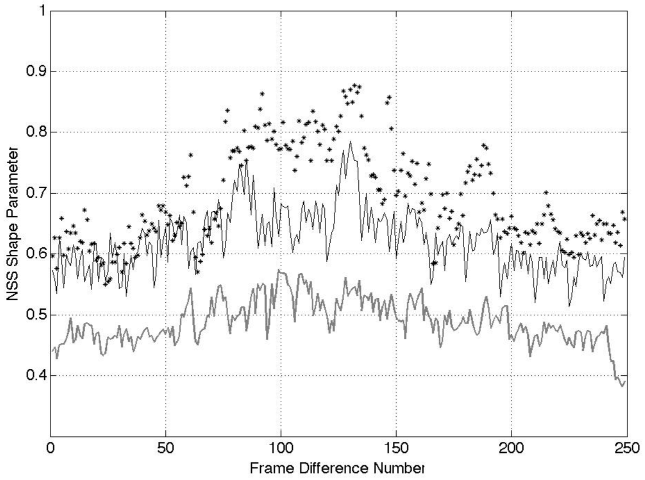


Fig. 4. Plot of video quality parameter γ over time for three videos. Stars (top): pristine video, DMOS = 0. Thin black (middle): medium quality video, DMOS = 56.1328. Gray (bottom): low quality video, DMOS = 72.1356.

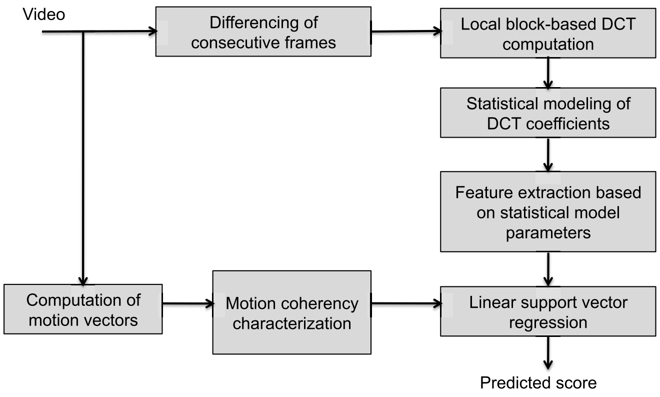

Video
Differencing of consecutive frames
Local block-based DCT computation
Statistical modeling of DCT coefficients
Feature extraction based on statistical model parameters
Computation of motion vectors
Motion coherency characterization
Linear support vector regression
Predicted score

Fig. 5. Blind VQA framework.

is whether a phenomenon of temporal masking of distortions exists, and if it does, whether it can be modeled and measured. While there is a 'standard model' of spatial contrast masking [39], [40], no such model exists that has been observed to accurately predict temporal masking of local temporal video perturbations. However, very recently Suchow *et al.* demonstrated a powerful temporal change *silencing* phenomenon that is triggered by the presence of large temporal image flows [41]. In a series of 'illusions' they devised, objects changing in hue, luminance, size, or shape appear to stop changing when they move in concert with other objects. Although this phenomenon is not yet well-modeled, our theory seeks to predict temporal change visibility as a function of cohesive, collective motion [42]. Highly localized space-time changes in video appearance (brightness, hue, size, shape) are rendered much less conspicuous or even invisible by large coherent motions in the scene. This suggests that localized space-time distortions in videos may be masked/silenced by large motions.

The presence of highly visible, predominantly temporal artifacts in videos and the complexity of perceptual motion processing are major reasons why still image quality assessment algorithms applied on a frame-by-frame basis fail to accurately predict human visual judgments of video quality. The type of motions in a scene may serve to either mask or enhance the visibility of distortions. It is hence important to take the type of motion into account in the design of VQA algorithms. Our model characterizes motion by utilizing a *coherencey measure*, which we describe next.

A. Motion Coherency

The experiments in [41] strongly suggest that large, coherent motion silences transient temporal change or "flicker", which is a reasonable description of many temporal video distortions. Following this observation, we characterize *motion coherence* using a 2D *structure tensor* model applied to a video's computed motion vectors. If motion vectors are not readily available, then a simple motion vector estimation algorithm is applied on $n \times n$ blocks to determine the corresponding spatial location of the blocks in one frame in the consecutive frame in time. The motion estimation is performed via a simple three-step search algorithm [43].

The motion coherence tensor summarizes the predominant motion directions over local neighborhoods, as well as the degree to which the local directions of motion flow are coherent. The 2D motion coherence tensor at a given pixel is given by:

$$S = \begin{bmatrix} f(M_x) & f(M_x.M_y) \\ f(M_x.M_y) & f(M_y) \end{bmatrix} \tag{1}$$

where

$$f(V) = \sum_{l,k} w[i,j]V(i-l, j-k)^2, \tag{2}$$

and $M_x(i,j)$ and $M_y(i,j)$ are horizontal and vertical motion vectors at pixel (i,j) respectively, and w is a window of dimension $m \times m$ over which the localized computation of the tensor is performed. The eigenvalues of the motion coherence tensor convey information about the spatial alignment of the motion vectors within the window of computation. The relative discrepancy between 2 eigenvalues is an indicator of the degree of anisotropy of the local motion (in the window), or how strongly the motion is biased towards a particular direction. This is effectively quantified by the coherence measure

$$C = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2, \tag{3}$$

where λ_1 and λ_2 are the eigenvalues of the motion coherence tensor. We use this measure in Video BLIINDS to characterize motion coherence over spatial patches of frame differences. The average value of these parameters, over all the frame differences in the video sequence, is computed and used as a feature for quality prediction.

B. Egomotion

In addition to object motion, global motion or egomotion may be present due to the motion of the camera or other large dominating motion. The velocity of global motion can affect the perception of scene content. Accordingly, our model accounts for the magnitude of global motion. This is computed simply as the mode of the motion vector magnitudes between every two consecutive frames. Motion vectors are computed according to the *three-step-search* algorithm in [43]. The absolute value of the difference between the mode and average motion vector magnitude per frame is computed and divided by the average motion magnitude per frame. In other words, let $M_{X(i)}$ and $M_{Y(i)}$ be the horizontal and vertical motion vector components of motion vector i respectively (corresponding to one frame difference). Also, let M and E be the mode and mean of the motion vector magnitudes (corresponding to two consecutive frames) respectively.

$$M = mode_{\{i=1...m\}}\left(\sqrt{(M_{X(i)})^2 + (M_{Y(i)})^2}\right), \tag{4}$$

$$E = \frac{1}{m}\sum_{i=1}^{m}\left(\sqrt{(M_{X(i)})^2 + (M_{Y(i)})^2}\right), \tag{5}$$

where m is the number of motion vectors per frame.

The quantities M and $|E - M|$ are then averaged over the frames of a video sequence resulting in M_{ave} and $|E - M|_{ave}$, respectively. Then the global motion characterization measure is given by

$$G = \frac{|E - M|_{ave}}{1 + M_{ave}} \tag{6}$$

This quantity represents the fraction of motion attributed to non-global motion ($|E - M|_{ave}$) over global motion (M_{ave}). By subtracting M (global motion) from the average motion E, we get a residual, and determine what fraction of the average motion is contributed to by that residual. G is used as a feature during the score prediction phase.

IV. NVS Model-Based Features

A good NVS (natural video statistics) model should capture regular and predictable statistical behavior of natural videos. Such models could be used to measure the severity of distortions in video signals since distortions may predictably modify these statistics. NVS models may be regarded as duals of low-level perceptual models since the HVS is hypothesized to have evolved with respect to the statistics of the surrounding visual environment over the millennia [26], [40], [44].

In the following we propose an NVS model of frame-differences that is expressed in the DCT domain and define a number of perceptually relevant features that are extracted from the model parameters. We begin by describing an NVS model of the DCT coefficients of patch frame differences. We then discuss the motion analysis process and how it is used to weight the parameters of the spatio-temporal DCT model.

A. Spatio-Temporal Statistical DCT Model

Consider a video sequence containing M frames. Each frame indexed $i + 1$ is subtracted from frame i, for $i \in \{1, ..., M - 1\}$, resulting in $M - 1$ difference-frames.

Each difference frame is then partitioned into $n \times n$ patches or blocks. The 2-D DCT is then applied to each $n \times n$ patch. The DCT coefficients from every block in each difference frame are modeled as following a generalized Gaussian probability distribution. Given an $m \times l$ video frame, there are $\frac{m \times l}{n \times n}$ DCT blocks per frame, each containing $n \times n$ frequency

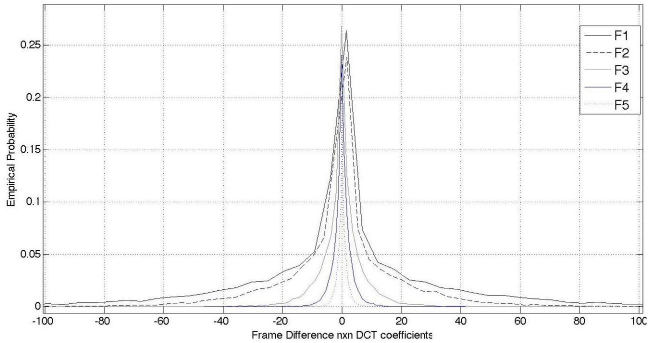


Fig. 6. Empirical distribution of DCT coefficients at 5 different frequencies from an $n \times n$ DCT decomposition of a frame-difference.

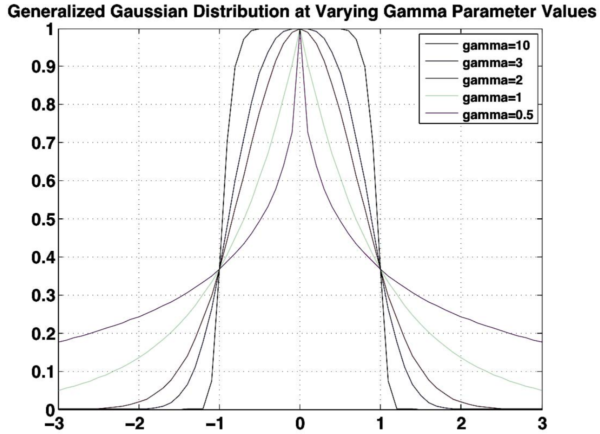


Fig. 7. Generalized Gaussian density plots for different values of the shape parameter γ.

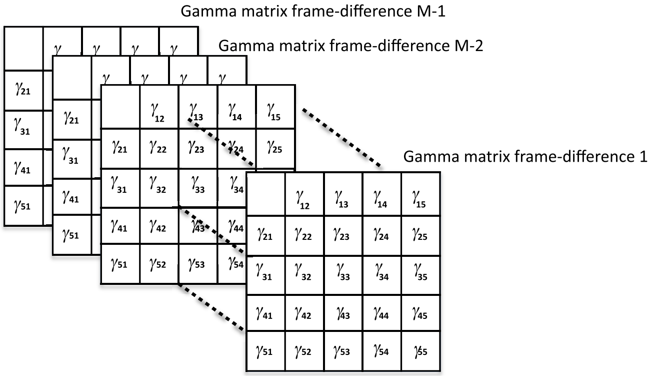

Gamma matrix frame-difference M-1
Gamma matrix frame-difference M-2
Gamma matrix frame-difference 1
γ12 γ13 γ14 γ15
γ21 γ22 γ23 γ24 γ25
γ31 γ32 γ33 γ34 γ35
γ41 γ42 γ43 γ44 γ45
γ51 γ52 γ53 γ54 γ55

Fig. 8. $n \times n$ matrix of the shape parameter γ values is obtained for each frame-difference, by fitting a GGD model to the histogram of each frequency in the $n \times n$ DCT block over all blocks in a frame-difference.

coefficients. Thus each of the $n \times n$ frequency coefficients in a DCT block occurs $\frac{m \times l}{n \times n}$ times per difference-frame. We fit the histogram of each frequency coefficient from all $n \times n$ patches in each difference frame with a parametric density function. Fig. 6 shows a histogram of the DCT coefficients at five different spatial frequencies F_1, F_2, ... F_5 in an $n \times n$ DCT decomposition of difference frames from a video that was not distorted. It may be observed that the coefficients are symmetrically distributed around zero and that the coefficient distributions at different frequencies exhibit varying levels of peakedness and spread about their support. This motivates the use of a family of distributions that encompasses a range of tail behaviors. The 1D generalized Gaussian density is a good fit to these coefficient histograms:

$$f(x|\alpha, \beta, \gamma) = \alpha e^{-(\beta|x-\mu|)^{\gamma}}, \tag{7}$$

where μ is the mean, γ is the shape parameter, and α and β are normalizing and scale parameters given by

$$\alpha = \frac{\beta\gamma}{2\Gamma(1/\gamma)}, \tag{8}$$

$$\beta = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}, \tag{9}$$

where σ is the standard deviation, and Γ denotes the ordinary gamma function

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt. \tag{10}$$

This family of distributions includes the Gaussian distribution ($\gamma = 2$) and the Laplacian distribution ($\gamma = 1$) [45]. As $\gamma \to \infty$ the distribution converges to a uniform distribution. Fig. 7 shows the generalized Gaussian distribution for a variety of values of the shape parameter (γ).

A variety of methods have been proposed to extract the parameters of this model. We deploy the reliable method given in [46].

After fitting a generalized Gaussian density to the histogram of each of the frequency coefficients from frame-difference patches across the image, we form an $n \times n$ matrix of shape parameters[1] per difference-frame. The motivation behind this approach is to characterize the statistical behavior of each of the frequencies in the local DCT blocks over time, as well as interactions among those frequencies. This is captured in the matrix of shape parameters obtained from each of the difference-frames. Fig. 8 depicts the matrix of shape parameter values obtained for each frame difference. This characterization is typically different for natural videos as opposed to distorted ones. The Video BLIINDS model aims to capture this statistical disparity and quantify it for perceptual video quality score prediction. We do not fit a GGD to the histograms of the DC values. These are however utilized for quality prediction as will be described shortly in Section IV-C.

B. Model-Based Sub-Band Features: Spectral Ratios

In order to capture the spectral signatures of videos (pristine and distorted), each $n \times n$ matrix of shape-parameters per difference frame is partitioned into three sub-bands as depicted in Fig. 9, where the top left band corresponds to shape-parameters modeling low-frequency coefficients, the middle partition corresponds to mid-band frequencies, and the lower right partition corresponds to high-frequency coefficients.

[1] The other parameters of the GGD did not contribute to higher quality prediction. We hence only retained the shape parameters of the model fits.

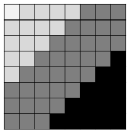
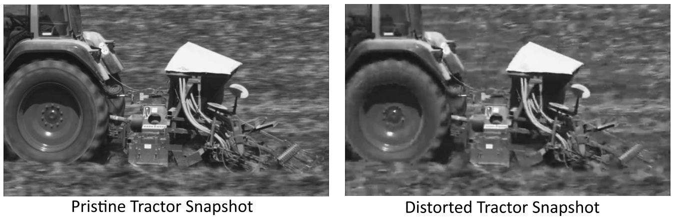


Fig. 9. Frequency band partition of frame differences. Top left: low frequency. Bottom right: high frequency.



DC	0.6110	0.5850	0.5290	0.5210
0.6330	0.6170	0.5860	0.5460	0.5730
0.6100	0.6380	0.5720	0.5930	0.5910
0.5890	0.6400	0.6370	0.5840	0.5820
0.5570	0.6110	0.6340	0.6380	0.6050

Pristine Tractor Gamma Matrix

DC	0.5420	0.4890	0.4430	0.4610
0.5710	0.5330	0.4870	0.4470	0.4810
0.5270	0.5440	0.4730	0.4640	0.4980
0.4990	0.5340	0.5100	0.4610	0.4750
0.4540	0.5250	0.5140	0.5230	0.4900

Distorted Tractor Gamma Matrix

Fig. 10. Snapshots of the pristine and distorted *Tractor* sequence, along with the computed γ matrices from corresponding frame-difference DCT coefficients. Notice how the γ values differ between the pristine and distorted sequences.

DC	0.2120	0.2140	0.2210	0.2330
0.2030	0.2150	0.2140	0.2190	0.2260
0.2040	0.2080	0.2140	0.2250	0.2360
0.2040	0.2150	0.2170	0.2330	0.2500
0.2060	0.2080	0.2170	0.2420	0.2440

Pristine Ice-Skating Gamma Matrix

DC	0.2130	0.2150	0.2200	0.2210
0.2040	0.2140	0.2140	0.2150	0.2190
0.2060	0.2090	0.2170	0.2250	0.2300
0.2030	0.2150	0.2170	0.2320	0.2480
0.2030	0.2100	0.2210	0.2410	0.2410

Distorted Ice-Skating Gamma Matrix

Fig. 11. Snapshots of the pristine and distorted *Ice-Skating* sequence, along with the computed γ matrices from corresponding frame-difference DCT coefficients. Notice how the γ values differ between the pristine and distorted sequences.

Before we describe the sub-band NSS features, we pause to show some examples of γ-matrix values obtained from a couple 'pristine' and distorted videos.

Fig. 10 shows the γ-matrices from one frame-difference of the 'pristine' *Tractor* video and a distorted counterpart of the same video. The shape-parameters changed significantly, indicating a distortion-induced modification of the shape of the distribution of the coefficients.

Fig. 11 depicts the same thing as Fig. 10 on a different video sequence.

It is instructive to observe the way the parameters became modified by distortion. First, it is more noticeable in the higher frequency band. Also, the ranges of the γ values are highly dependent on the nature of the content of the video. For the *Tractor* sequence, which is rich in spatial activity, the γ values (for both reference and distorted videos) ranged between 0.4 and 0.65, whereas the γ values corresponding to the spatially smooth *Ice-Skating* sequence fell in a completely different range. This is not surprising since frame difference DCT coefficients may be expected to have a more peaky distribution on highly smooth regions/ sequences.

This kind of content-dependency however, poses a challenge in blind quality assessment since the absolute parameter values are less important than relative values between bands. To capture the inter-relationships of features between the different bands (low, medium, and high frequency) in a less content-dependent manner, we compute ratios of parameters between the bands. Ratios tend to reduce content-dependency (since the γ parameters in different bands fall in comparable ranges within similar content while still maintaining sensitivity to distortion).

The geometric mean of the shape parameters in each of the low, mid, and high frequency bands is first computed as

$$G_f = \left(\prod_{i=1}^{m} \gamma_i\right)^{1/m}, \tag{11}$$

where $f \in \{low, mid, high\}$.

The low frequency band γ's in each 5×5 matrix depicted in Fig. 8 are denoted $\{\gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \gamma_{33}\}$. The γ-parameters corresponding to the mid-band are notated as $\{\gamma_{14}, \gamma_{15}, \gamma_{24}, \gamma_{33}, \gamma_{42}, \gamma_{43}, \gamma_{51}, \gamma_{52}\}$, while the γ-parameters corresponding to the high frequency band are $\{\gamma_{25}, \gamma_{34}, \gamma_{35}, \gamma_{44}, \gamma_{45}, \gamma_{53}, \gamma_{54}, \gamma_{55}\}$. Once G_{low}, G_{mid}, and G_{high} are computed the following spectral ratios are obtained per shape-parameter matrix (i.e per frame difference):

$$R_1 = \frac{G_{high}}{G_{low}}, \tag{12}$$

$$R_2 = \frac{G_{high}}{G_{mid}}, \tag{13}$$

$$R_3 = \frac{G_{mid}}{G_{low}}, \tag{14}$$

$$R_4 = \frac{(G_{high} + G_{mid})/2}{G_{low}}, \tag{15}$$

and

$$R_4 = \frac{G_{high}}{(G_{low} + G_{mid})/2}, \tag{16}$$

Finally, the geometric mean of each ratio is computed over all frame differences. The geometric mean makes it possible to account for changes in parameter values that fall in different ranges because of content differences without having to attempt alignment of these parameter ranges.

C. Temporal Variation of Mean DC Coefficients

To track temporal variations in the average intensity of differenced video frames (from all $n \times n$ DCT blocks), the discrete temporal derivative of the average intensity per video

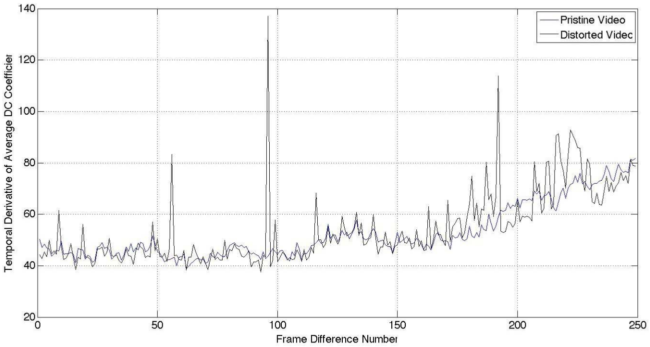


Fig. 12. Plot of the temporal derivative of mean DC coefficients for a pristine and a distorted video.

frame is also computed. An example is shown in Fig. 12. This is a simple measure of sudden local changes which may arise from various temporal distortions that result in local 'flicker. 'Let D_i be the average DC coefficient value per frame i. The absolute discrete temporal derivative of D_i is estimated then as

$$T_i = |D_{i+1} - D_i|, \tag{17}$$

where D_{i+1} and D_i are the average DC coefficients at frames indexed $i + 1$ and i respectively. The mean of the absolute discrete temporal derivatives is computed as a feature for prediction along with the other extracted features.

D. Spatial Naturalness

In addition to the above described spatio-temporal features (which are based on frame-differences), we also utilize the image naturalness index NIQE features described in [47], to predict crude frame-by-frame naturalness scores. These naturalness scores are predicted from the frame statistics in the pixel domain. While these features do not yield high video quality prediction performance when used in isolation, they do capture spatial aspects of distortion that are not contained in the other features and thereby boost prediction performance.

E. Prediction

Given a database of distorted videos and associated human judgments, the extracted features are used to train a linear kernel support vector regressor (SVR) to conduct video quality score prediction. The SVR based on the implementation in [48] was used to conduct quality score prediction.

The complete list of features used for video quality prediction is: the motion coherency measure and the global motion measure which are key characterizations of the temporal behavior exhibited by a video sequence, the five NVS γ (shape-parameter) spectral ratios, absolute temporal derivative of mean DC coefficients, and the purely spatial frame-naturalness measure described in [47].

Each feature is computed from each frame difference (except the spatial naturalness measure), then temporally pooled over a 10 second interval. Prior to feeding the features into the SVR, the spatio-temporal features (other than the naturalness index) are subjected to a logarithmic nonlinearity, as depicted in Fig. 13. Quality prediction is then performed on the entire video segment.

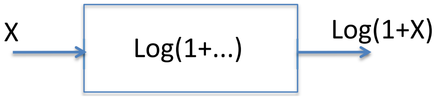


Fig. 13. The spatio-temporal features (DC feature, coherency measure, global motion measure, and shape-parameter spectral ratios are logarithmically transformed before being used as features for quality prediction by the SVR).

V. Experiments and Result

The algorithm was evaluated on the publicly available LIVE VQA database [38]. The LIVE VQA database has a total of 160 videos derived from 10 reference videos of highly diverse spatial and temporal content. The database contains videos distorted by four distortion types: 1) MPEG-2 compression, 2) H.264 compression, 3) wireless distortions, and 4) IP distortions. We first evaluated Video BLIINDS by applying it on each distortion type in isolation, then we mixed the distortions together and applied the method on the mixture. We split the database into content-independent train and test sets: 80% of the content was used for training and the remaining 20% was used for testing. We compute the Spearman rank order correlation coefficient (SROCC) between predicted scores and the subjective scores of the database for every possible combination of train/test split.

The patch size for the DCT computation that was used is 5×5. This is similar to the feature extraction block size chosen in BLIINDS-2 [25]. The motion vectors involved in the computation of the motion coherency tensor and the global motion characterization measure are derived from 10×10 pixel blocks.

A. Feature Contribution to Prediction Performance

In order to understand the contribution of each individual conceptual feature to the overall prediction performance of Video BLIINDS, each was used in isolation of the other features to predict quality, and the correlation between predicted and actual quality scores was computed. Table I shows the Spearman rank order correlation coefficients obtained when using each conceptual feature in isolation of the other features for prediction of video quality. The NVS parameter ratios result in the highest prediction performance among the all features. Note that the coherency and global motion measures are not quality features per se. In fact, these are features that help identify and characterize the type of the video content, which can affect the perception of video quality.

B. Algorithm Prediction Performance

There are no existing blind VQA approaches that are non-distortion specific, which makes it difficult to compare our algorithm against other methods. Full-reference and reduced reference approaches have the enormous advantage of access

TABLE I
SROCC CORRELATION ON EVERY POSSIBLE COMBINATION OF TRAIN/TEST SET SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS) USING EACH CONCEPTUAL FEATURE IN ISOLATION OF OTHER FEATURES, FOR QUALITY PREDICTION. 80% OF CONTENT USED FOR TRAINING

Spatial Naturalness	DC Feature	NVS Shape-Parameter Ratios	Coherency Measure	Global Motion Measure
0.324	0.486	0.587	0.204	0.220

TABLE II
FULL-REFERENCE AND REDUCED-REFERENCE MEDIAN SROCC CORRELATIONS ON EVERY POSSIBLE COMBINATION OF TRAIN/ TEST SET SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS). 80% OF CONTENT USED FOR TRAINING

Distortion	PSNR	SSIM	VQM	STMAD	MOVIE	RRED
MPEG-2	0.643	0.786	0.905	0.929	0.905	0.905
H.264	0.714	0.881	0.786	0.952	0.881	0.905
Wireless	0.691	0.691	0.762	0.810	0.786	0.762
IP	0.600	0.543	0.771	0.771	0.771	0.771
ALL	0.677	0.650	0.745	0.834	0.807	0.826

TABLE III
FULL-REFERENCE AND REDUCED-REFERENCE MEDIAN LCC CORRELATIONS ON EVERY POSSIBLE COMBINATION OF TRAIN/TEST SET SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS). 80% OF CONTENT USED FOR TRAINING

Distortion	PSNR	SSIM	VQM	STMAD	MOVIE	RRED
MPEG-2	0.696	0.805	0.943	0.942	0.955	0.904
H.264	0.698	0.851	0.850	0.947	0.919	0.892
Wireless	0.798	0.634	0.943	0.904	0.920	0.806
IP	0.733	0.726	0.896	0.901	0.895	0.816
ALL	0.722	0.625	0.780	0.861	0.852	0.725

TABLE IV
NO-REFERENCE MEDIAN SROCC AND LCC CORRELATIONS ON EVERY POSSIBLE COMBINATION OF TRAIN/TEST SET SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS). 80% OF CONTENT USED FOR TRAINING

	SROCC		LCC	
Distortion	NIQE	Video BLIINDS	NIQE	Video BLIINDS
MPEG-2	0.523	0.869	0.490	0.924
H.264	0.541	0.839	0.579	0.893
Wireless	0.280	0.815	0.387	0.951
IP	0.276	0.779	0.443	0.946
ALL	0.151	0.759	0.317	0.881

to the reference video or information about it. Blind algorithms generally require that the algorithm be trained on a portion of the database. We do however, compare against the naturalness index NIQE in [47], which is a blind IQA approach applied on a frame-by-frame basis to the video, and also against top performing full-reference and reduced reference algorithms.

The algorithms were separately tested on those portions of the LIVE VQA database that contain specific distortions (MPEG2, H264, wireless distortions, and IP distortions), as well as on the entire database containing all the distortions mixed together in "the same bucket." Consequently, Video BLIINDS was trained and tested on each distortion of the database separately, and on all of the distortions mixed together. The median SROCCs (Spearman rank order correlation coefficient) and PLCCs[2] (Pearson's linear correlation coefficient) between subjective and predicted scores for the top-performing full-reference and reduced reference VQA algorithms are given in Tables II and III respectively, (including full-reference PSNR and SSIM image quality indices). VQM [8] and Video RRED [37] are top-performing reduced reference VQA approaches, with VQM being a standardized approach. On the other hand, MOVIE [6] and ST-MAD [11] are highly competitive (in terms of prediction performance) full-reference VQA algorithms. The median SROCCs and PLCCs for the blind IQA approach NIQE and Video BLIINDS are shown in Table IV. We chose to report the results for the full and reduced reference methods in separate tables than those of the no-reference methods. The reason for this is to allow a fairer comparison of algorithms. Full and reduced reference approaches utilize a reference video for quality prediction. Hence the reference videos cannot be included in the test sets as including them would lead to misleadingly higher correlations. On the other hand, it is informative to include the pristine/reference videos in the test sets of no-reference algorithms since one needs to know how well the algorithm is able to predict the quality of a relatively "pristine" video.

Video BLIINDS clearly outperforms the blind NIQE index and the full-reference PSNR and SSIM measures. Video BLIINDS does not quite attain the performance level of state-of-the-art full-reference VQA measures, (MOVIE and ST-MAD), but its performance is nearly as good and with much less computational cost. Of course, Video BLIINDS does not rely on any information from the pristine version of the video to make quality predictions. It does, however, rely on being trained *a priori* on a set of videos with associated human quality judgments.

A statistical analysis of the SROCCs obtained for each of the QA approaches (PSNR, SSIM, VQM, NIQE, and Video BLIINDS) was performed using a multi-comparison analysis of variance (ANOVA) test. Fig. 14 shows the spreads of distributions of the SROCCs for each algorithm. The plot shows that the reduced-reference VQM and Video BLIINDS perform very similarly on the LIVE VQA database, and outperform PSNR, SSIM, and NIQE. Table V shows the results of the ANOVA test indicating whether each algorithm is superior than another by a statistically significant SROCC margin.

In addition to testing on the LIVE VQA database, we also tested the performance of Video BLIINDS on the 4-CIF

[2] Since the relationship between predicted and actual scores is not necessarily a linear one, a nonlinear function between the predicted and actual variables is fit prior to computing the PLCC.

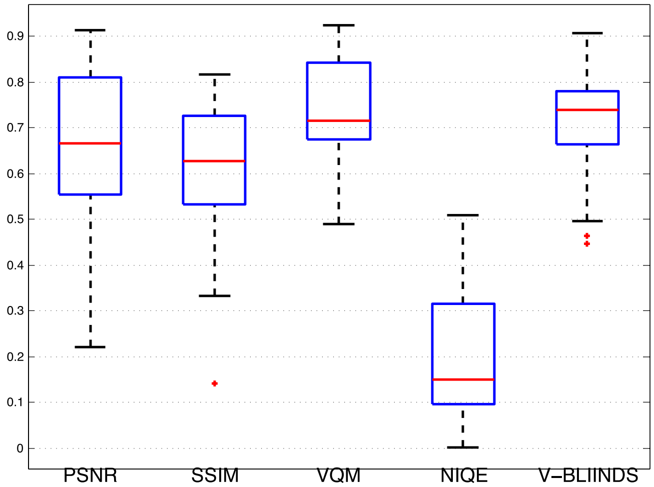


Fig. 14. Plot of median SROCC distribution for PSNR, SSIM, VQM, NIQE, and Video BLIINDS.

TABLE V

MULTI-COMPARISON ANOVA TEST RESULTS. −1 MEANS COLUMN OUTPERFORMS ROW BY A STATISTICALLY SIGNIFICANT DIFFERENCE. 0 MEANS ROW AND COLUMN ARE NOT STATISTICALLY DIFFERENT. +1 MEANS ROW OUTPERFORMS COLUMN BY A STATISTICALLY SIGNIFICANT DIFFERENCE

Algorithm	PSNR	SSIM	VQM	NIQE	V-BLIINDS
PSNR	0	0	-1	+1	-1
SSIM	0	0	-1	+1	-1
VQM	+1	+1	0	+1	0
NIQE	-1	-1	-1	0	-1
V-BLIINDS	+1	+1	0	+1	0

TABLE VI

NO-REFERENCE MEDIAN SROCC AND LCC CORRELATIONS ON EVERY POSSIBLE COMBINATION OF TRAIN/TEST SET SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS). 80% OF CONTENT USED FOR TRAINING ON THE EPFL-PoliMi DATABASE

SROCC		LCC	
NIQE	Video BLIINDS	NIQE	Video BLIINDS
0.555	0.807	0.124	0.752

EPFL-PoliMi database.[3] The median LCC and SROCC scores for NIQE and Video BLIINDS are shown in Table VI.

VI. ALGORITHM COMPLEXITY

Let $m \times k$ be the frame dimension, $n \times n$ the dimension of the blocks from which the model-based features are extracted (in our model $n = 5$), and let $w \times w$ be the dimension of the motion vector structure tensor. The computational complexity of Video BLIINDS is largely determined by the complexity of the DCT transform, the generalized Gaussian density parameter estimation, and by the motion coherency computation.

The computational complexity of the DCT computation and of the generalized Gaussian density parameter estimation is of the order of $\frac{m \times k}{n^2} \times n^2 \mathrm{log} n = m \times k \times \mathrm{log} n$. Fast algorithms exist for DCT computation that are of the order $O(n^2 \mathrm{log} n)$ [49], where n is the dimension of the frame patches. Parameter estimation of the generalized Gaussian is of the order of computing moments of the data within each block ($O(n^2)$), and of numerically estimating the shape parameter γ . From empirical data of natural scenes, it is observed that $0 < \gamma < K$. We set $K = 10$, since it was observed that $\gamma << 10$. The interval $[0, K]$ was partitioned in steps of size ϵ, and the parameter γ was determined by solving an inverse function by numerically sweeping the interval $[0, K]$ in increments of size ϵ [46]. The complexity of such an operation is of the order $O(log(1/\epsilon))$. ϵ was chosen to be 0.001 Hence $log(1/\epsilon) << min(m, k)$.

The complexity of computing motion coherency is determined by the complexity of computing motion vectors using the three-step search algorithm in [43], which is an $O(n^2)$ operation, and from computing the eigenvalues of the $w \times w$ structure tensor. In the most general case, eigenvalue computation is an $O(w^3)$ operation.

The algorithm is highly parallelizable as one can perform computations on the image blocks in parallel. A further computational advantage can be attained by bypassing DCT computation when DCT coefficients have already been computed, e.g. by an encoder. We envision that the Video BLIINDS approach is easily extensible to scenarios involving DCT-like transforms such as the H.264 integer transforms.

VII. SOME PRACTICAL APPLICATIONS OF VIDEO BLIINDS

The results in the previous section demonstrate that the Video BLIINDS features are well suited for predicting the visual quality of videos compressed using the H.264 standard. We now show that the Video BLIINDS features can be used in two useful applications involving H.264 compression.

The first application addresses the following question: Given an uncompressed video, how much can it be compressed (i.e. what minimum bit rate is required) to achieve a desired level of quality (as expressed by DMOS or MOS)? Note that different videos generally require different compression bit rates to be represented at a specific visual quality, depending on their spatial and temporal content. In the second application we ask: Given a video compressed by H.264, can the bit-rate at which it has been compressed be predicted? We show that the Video BLIINDS features can be used to address both of these timely questions.

In the first application, which we call the *Video BLIINDS Bit Rate Selector*, we design an algorithm that selects the bit rate at which to compress a video at a given level of perceptual quality. It takes as input an uncompressed video

[3] Regarding training and testing on EPFL, since there is so little content to train on (6 4-CIF reference videos) 80% of the content is only 5 references. Consequently a leave-one-out (in this case leave one ”reference and corresponding distorted videos”' out) train/test analysis is performed to predict the video scores. Thus the scores for each reference video and corresponding distorted versions are predicted by an SVM trained on all the other reference videos and their corresponding distorted counterparts. Each video thus has a predicted MOS coming from an SVM that was trained on all the content except its own (the content is completely separate between the training and test sets). However, when computing the SROCC, every combination of 2 different contents/reference videos was taken and the MOS predicted in the test phase was used to compute a median SROCC. This is to ensure that more than one type of content/reference is present in the test set. Otherwise, homogeneous content could result in deceptively high SROCC values.

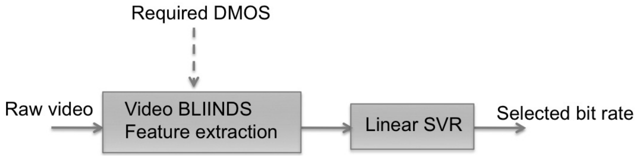


Fig. 15. Application 1: Perceptual bit rate selector.

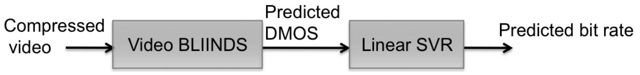


Fig. 16. Application 2: Bit rate prediction.

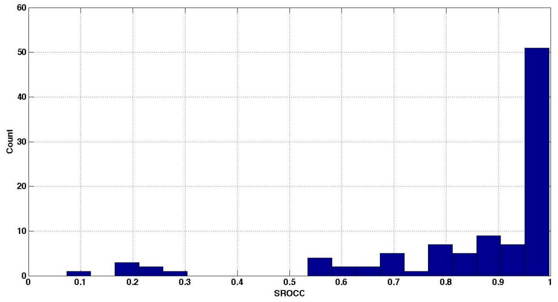


Fig. 17. Application 1: Histogram of SROCC between predicted and actual bit rates over 100 iterations of train/test splits.

and the desired quality level to be achieved by compression. It then extracts global Video BLIINDS features (pooled over 10 second intervals), and uses a linear SVR to predict the bit rate at which the video needs to be compressed. The overall framework of the perceptual bit rate selection algorithm is depicted in Fig. 15.

The second application which we call the *Video BLIINDS Bit Rate Predictor*, seeks to predict the rate at which a video has already been compressed, using Video BLIINDS quality features. This process is summarized in Fig. 16.

At this point it is important to mention that the above two applications assume a particular choice of the H.264 encoder parameters. These are specified in [50]. In other words, given a particular configuration of the H.264 encoder parameters, it is possible to derive a mapping from desired visual quality to an appropriate bit rate. This is inherent to the H.264 encoder parameters used on the videos comprising the training set from which the mapping was derived. The same assumption applies for the second application.

Both applications were tested on the H.264 compressed portion of the LIVE VQA database which contains a total of 50 videos derived from 10 reference videos. The details of the H.264 encoding parameters can be found in [38]. The compressed videos spanned bit rates between 0.2MB to 6MB. 80% of the content was used for training and the remaining 20% was used for testing. The process was repeated over 100 iterations of randomly selecting the train and test sets. In Application 1 (Bit Rate Selector), a median SROCC of 0.954 was achieved between the predicted and actual bit rates. The histogram of the obtained SROCC values is shown in Fig. 17.

Notice how there is a concentration of SROCC values between 0.8 and 1, with a few outliers below 0.5.

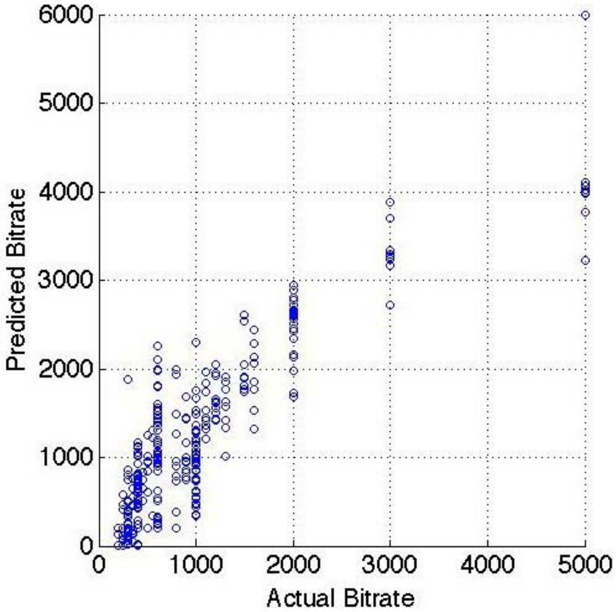


Fig. 18. Application 1: Scatter plot of predicted versus actual bit rates.

The performance of Application 1 depends on the cumulative error of first predicting the visual quality of the video (DMOS), and then using the predicted DMOS to predict the bit rate at which the video is compressed. The median mean square error between predicted and actual bit rates over the 100 iterations was also computed, and it was found to be 0.374 MB. A scatter plot of predicted versus actual bit rates is shown in Fig. 18, depicting the linear correlation between the two. Although we might expect subjective video quality to vary monotonically with compression level, this relationship need not be strict. For example, the perceived quality of a video might remain level over a fairly wide range of compression levels. For this reason, Video BLIINDS features may not necessarily be expected to yield precision bit rate selection. However, they can be expected to deliver reliable subjective quality in the resulting compressed video.

In Application 2 (Bit Rate Predictor), a median SROCC of 0.860 was achieved between the selected bit rate and the bit rate of the actual compressed videos in the database. The challenge in the second application is that the SVM that learns a mapping from the tuple of features plus desired DMOS to bit rate only sees the features extracted from the pristine videos of the database and not from the compressed videos. The histogram of the obtained SROCC values is shown in Fig. 20. The median mean square error between predicted and actual bit rates over the 100 iterations was also computed, and it was found to be 0.471 MB. A scatter plot of selected versus actual bit rates is further shown in Fig. 19. In the first application, the Video BLIINDS features deliver excellent quality predictor and generally correct, if imprecise, selected bit rates. Again, this may be attributed to a non-strict monotonic relationship between video quality and bit rate.

Similar to the results for Application 1, while the SROCC scores are concentrated above 0.8, there are a number of

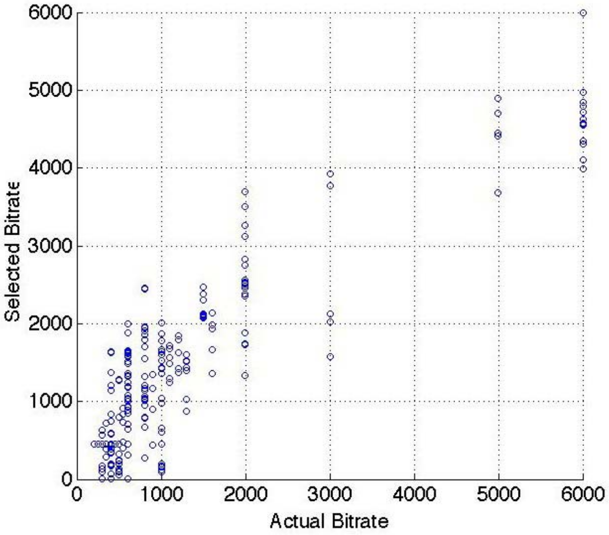


Fig. 19. Application 1: Scatter plot of selected versus actual bit rates.

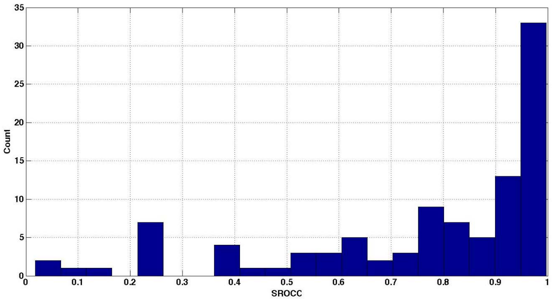


Fig. 20. Application 1: Histogram of SROCC between selected and actual bit rates over 100 iterations of train/test splits.

outliers below 0.5, showing the challenge in learning the mapping from desired quality to bit rate given only a few features from the original non-compressed video.

These two applications are good examples of how Video BLIINDS features can be used in practical ways. It remains for future work to explore how NVS features such as those used in Video BLIINDS can be exploited for other perceptual optimization problems, such as tracking, denoising, deblocking, and so on.

VIII. Challenges and Future Work

Several challenges remain to be tackled on the blind VQA problem. Our aim is to achieve correlations as high as those obtained via full-reference algorithms.

There is still much room for improvement on developing motion models that can be effectively incorporated into blind VQA models. Research avenues in this direction include more complete modeling of temporal filtering in the lateral geniculate nucleus (LGN) and motion processing in Areas MT/V5 and MST of extrasriate cortex [27], [51], [52].

As we continue our quest to better understand the mechanisms of motion processing in the HVS, we also are faced by the challenge of finding more complete models of natural video statistics. Models that are uniform across content, while still being predictably disturbed by distortion levels should contribute to better predicted quality.

TABLE VII
No-Reference Median SROCC Correlation on Homogeneous Content of the EPFL-PoliMi Database

NIQE	Video BLIINDS
0.6978	0.989

We demonstrate how this challenge manifests by showing how our results on the EPFL database differ if tested on individual video sequences (instead of computing correlations on a mixture of video sequences). Table VII illustrates our point. In Table VII, we report median SROCC between predicted and subjective scores when the correlations are computed using each EPFL-PoliMi sequence (pristine and distorted counterparts) in isolation.

When content is kept homogeneous in this manner, Video BLIINDS almost perfectly ranks the videos according to. The problem becomes more challenging when there is significant content variation.

Capturing temporal distortions increases the computational complexity of VQA algorithms making real time processing more challenging. Current FR and RR VQA algorithms that correlate well with perception, such as those in [53], [11], and [37], can be very slow. Yet since many applications require real time monitoring of video quality in, there is considerable motivation to create VQA algorithms that are simple, perceptual, and fast.

To develop and validate accurate NVS and motion models, a large corpus of videos is needed. For many reasons, evaluating and benchmarking VQA algorithms is much more involved than IQA validation. A subject can requires significantly more time to view a video than a still image, which limits the sizes and availability of VQA databases.

Blind VQA algorithms that are trained on a database containing a specific set of distortions and associated human scores, are applicable to the set of distortions present in the training phase of the algorithm. It is also desirable that a learning-based blind VQA model be trained on a database containing a large number of videos of varying contents in order to learn as accurate a mapping as possible. If however, we were able to do away with training on human scores and only rely on models of perceptual and dual models of natural scenes (i.e., from a corpus of natural/pristine videos only), then it may be possible to avoid the limitations of regression (dependency on the distortion types in the database). This is a direction we have begun exploring as a promising avenue for future work.

IX. Conclusion

We have described a natural scene statistic model-based approach to the no-reference/blind video quality assessment problem. The new Video BLIINDS[4] model uses a small number of computationally convenient DCT-domain features.

[4] Regarding the resemblance between the IQA index BLIINDS [25] and the spatial IQA index in Video BLIINDS: Both model the distributions of local DCT coeffcients, but in different ways: Unlike [25], Video BLIINDS fits a histogram to each individual frequency in the 5 × 5 DCT block, over all blocks occurring in every frame-difference.

The method correlates highly with human visual judgments of quality. Additionally, we demonstrated two interesting applications of the Video BLIINDS features.

References

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

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst., Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.

[3] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[4] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, p. 011006-1–011006-21, Mar. 2010.

[5] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[6] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[7] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[8] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 10, no. 3, pp. 312–322, Sep. 2004.

[9] A. B. Watson, J. Hu, and J. F. McGowan, "DVQ: A digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[10] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bito, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 266–279, Apr. 2009.

[11] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most apparent distortion model for video quality assessment," in *Proc. IEEE ICIP*, Sep. 2011, pp. 2505–2508.

[12] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[13] C. Li and A. C. Bovik, "Content-weighted video quality assessment using a three-component image model," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011003-1–011003-9, Jan. 2010.

[14] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE*, vol. 5666, pp. 149–159, Jan. 2005.

[15] L. Qiang and Z. Wang, "Reduced-reference image quality assessment using divisive-normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[16] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.

[17] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 16, no. 2, pp. 260–273, Feb. 2006.

[18] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2000, pp. 981–984.

[19] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita, "No-reference image quality assessment for JPEG2000 based on spatial features," *Signal Process., Image Commun.*, vol. 23, no. 4, pp. 257–268, Apr. 2008.

[20] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *Proc. Int. Workshop Qual. Multimedia Exper.*, Jul. 2009, pp. 64–69.

[21] X. Feng and J. P. Allebach, "Measurement of ringing artifacts in JPEG images," *Proc. SPIE*, vol. 6076, pp. 74–83, Jan. 2006.

[22] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[23] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010.

[24] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[25] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[26] R. Blake and R. Sekuler, *Perception*, 5th ed. New York, NY, USA: McGraw-Hill, 2006.

[27] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neurosci.*, vol. 28, pp. 157–189, Mar. 2005.

[28] B. A. Wandell, *Foundations of Vision*. Sunderland, MA, USA: Sinauer Associates Inc., 1995.

[29] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[30] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision Models and Applications to Image and Video Processing*. New York, NY, USA: Springer-Verlag, 2001, pp. 179–200.

[31] D. H. Kelly, "Motion and vision. II: Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Amer.*, vol. 69, no. 10, pp. 1340–1349, Oct. 1979.

[32] H. Boujut, J. Benois-Pineau, T. A. O. Hadar, and P. Bonnet, "No-reference video quality assessment of H.264 video streams based on semantic saliency maps," *Proc. SPIE*, vol. 8293, pp. 82930T-1–82930T-9, Jan. 2012.

[33] S. Roth and M. J. Black, "On the spatial statistics of optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1. Oct. 2005, pp. 42–49.

[34] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *Proc. IEEE Conf. CVPR*, Jun. 1991, pp. 310–315.

[35] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE ICASSP*, Apr. 2007, pp. 869–872.

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

[37] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[38] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[39] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 9, pp. 181–197, Aug. 1992.

[40] Z. Wang and A. C. Bovik, "Reduced and no-reference visual quality assessment: The natural scene statistics model approach," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 29–40, Nov. 2011.

[41] J. W. Suchow and G. A. Alvarez, "Motion silences awareness of visual change," *Current Biol.*, vol. 21, pp. 140–143, Jan. 2011.

[42] L. K. Choi, A. C. Bovik, and L. K. Cormack, "A flicker detector model of the motion silencing illusion," *J. Vis.*, vol. 12, no. 9, p. 777, May 2012.

[43] R. Li, B. Zeng, and M. L. Liou, "A new three-step search algorithm for block motion estimation," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 2, no. 2, pp. 438–442, Aug. 1994.

[44] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neurosci.*, vol. 5, no. 6, pp. 598–604, Jun. 2002.

[45] A. C. Bovik, T. S. Huang, and D. C. Munson, "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1342–1350, Dec. 1983.

[46] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.

[47] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'Completely Blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[48] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "Kernlab—An S4 package for kernel methods in R," *J. Statist. Softw.*, vol. 11, no. 9, pp. 1–20, Oct. 2004.

[49] W. H. Chen, C. H. Smith, and S. Fralick, "A fast computational algorithm for discrete cosine transform," *IEEE Trans. Commun.*, vol. 25, no. 9, pp. 1004–1009, Sep. 1977.

[50] (2007). *H.264/MPEG-4 AVC Reference Software Manual* [Online]. Available: http://iphome.hhi.de/suehring/tml/

[51] E. P. Simoncelli and D. J. Heeger, “A model of neuronal responses in visual area MT,” *Vis. Res.*, vol. 38, no. 5, pp. 743–761, Mar. 1998.

[52] J. A. Perrone, “A visual motion sensor based in the properties of V1 and MT neurons,” *Vis. Res.*, vol. 44, no. 15, pp. 1733–1755, Jul. 2004.

[53] K. Seshadrinathan and A. C. Bovik, “Motion-based perceptual quality assessment of video,” *Proc. SPIE*, vol. 7240, pp. 72400X-1–72400X-12, Feb. 2009.

Michele A. Saad works for Intel Corporation. She received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2013, the B.E. degree in computer and communications engineering from the American University of Beirut, Lebanon, in 2007, and the M.S. degree in electrical and computer engineering from the University of Texas at Austin in 2009. Her research interests include statistical modeling of images and videos, motion perception, design of perceptual image and video quality assessment algorithms, and statistical data analysis and mining and machine learning. She is a recipient of the Intel Ph.D. Fellowship and the Microelectronics and Computer Development Fellowship from the University of Texas at Austin from 2007 to 2009. She is a Former Member of the Laboratory of Image and Video Engineering and the Wireless Networking and Communications Group, University of Texas at Austin.

Alan C. Bovik is the Curry/Cullen Trust Endowed Chair Professor with the Department of Electrical and Computer Engineering and the Institute for Neuroscience, University of Texas at Austin. His research interests include image and video processing, computational vision, and visual perception. He has published over 700 technical articles and holds several U.S. patents. His books include the recent companion volumes *The Essential Guides to Image and Video Processing* (Academic Press, 2009). He has received a number of major awards from the IEEE Signal Processing Society, including the Society Award in 2013, the Education Award in 2007, the Technical Achievement Award in 2005, and the Meritorious Service Award in 1998, as well as co-authoring papers that received the Best Paper Award in 2009, the Signal Processing Magazine Best Paper Award in 2013, and the Young Author Best Paper Award in 2013. He has been honored by other technical societies as well, including receiving the IST Honorary Member Award in 2013, the SPIE Technical Achievement Award in 2013, and the SPIE/IS&T Imaging Scientist of the Year Award in 2011. He received the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana in 2008, the IEEE Third Millennium Medal in 2000, and the two Journal Paper Awards from the International Pattern Recognition Society in 1988 and 1993. He is a fellow of the Optical Society of America, the Society of Photo-Optical and Instrumentation Engineers, and the American Institute of Medical and Biomedical Engineering. He has been involved in numerous professional society activities, including: Board of Governors for the IEEE Signal Processing Society from 1996 to 1998; Co-Founder and Editor-in-Chief for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1996 to 2002; Editorial Board for the THE PROCEEDINGS OF THE IEEE from 1998 to 2004; Series Editor for *Image, Video, and Multimedia Processing* (Morgan and Claypool Publishing Company, 2003); and Founding General Chairman for the First IEEE International Conference on Image Processing, Austin, in 1994. Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.

Christophe Charrier received the M.S. degree from the Nantes University of Science and Technology, Nantes, France, in 1993, and the Ph.D. degree from the University Jean Monnet of Saint-Etienne, France, in 1998. Since 2001, he has been an Associate Professor with the Communications, Networks and Services Department, Cherbourg Institute of Technology, France. From 1998 and 2001, he was a Research Assistant with the Laboratory of Radiocommunications and Signal Processing, Laval University, Quebec, QC, Canada. In 2008, he was a Visiting Scholar with the LIVE Laboratory, University of Texas at Austin. From 2009 to 2011, he was an Invited Professor with the Computer Department, University of Sherbrooke, Canada. His current research interests include digital image and video coding, processing and quality assessment, and computational vision.