BLIND QUALITY ASSESSMENT OF VIDEOS USING A MODEL OF NATURAL SCENE STATISTICS AND MOTION COHERENCY

Michele A. Saad

The University of Texas at Austin
Department of Electrical and Computer Engineering

Alan C. Bovik

The University of Texas at Austin
Department of Electrical and Computer Engineering

ABSTRACT

We propose a no-reference algorithm for video quality evaluation. The algorithm relies on a natural scene statistics (NSS) model of video DCT coefficients as well as a temporal model of motion coherency. The proposed framework is tested on the LIVE VQA database, and shown to correlate well with human visual judgments of quality.

Index Terms— No-reference video quality assessment, discrete cosine transform, natural scene statistics, motion coherency.

1. INTRODUCTION

The tremendous increase in personal digital assistants (PDAs), smart phones, and tablets among consumers in the last decade, has lead to an enormous increase in video traffic over both wired and wireless networks. This increase has consequently led to bandwidth and capacity challenges while catering for consumers' rising demands for video over wired and wireless networks and maintaining a high quality of visual experience. The need for reliable automatic, perceptual video quality assessment methods is hence necessary.

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, which we have dubbed Video BLIINDS, 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.

2. NATURAL VIDEO STATISTICS FRAMEWORK

We 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 [1], [2], and hence, the study of natural image statistics is highly relevant to understanding visual perception.

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.

Figure 1 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. This motivates VQA models that use statistical differences between the DCT coefficients of frame differences in pristine and distorted videos.

The new blind VQA model is summarized in Fig. 2. 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 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.

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

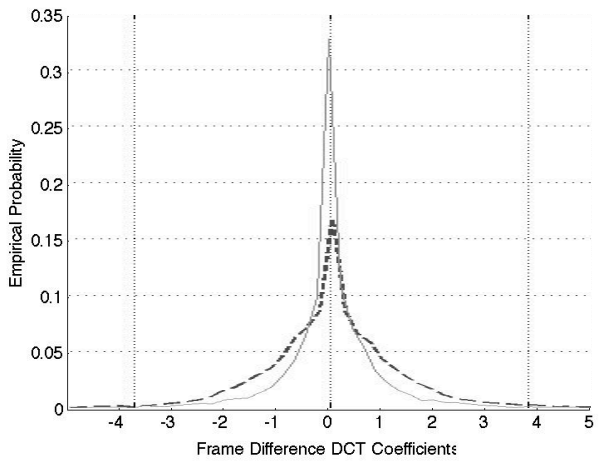

Fig. 1. Empirical probability distribution of frame-difference DCT coefficients of pristine and distorted videos. Dashed line: pristine video. Solid line: distorted video.

and use to weight the parameters derived from the spatio-temporal NVS model of DCT coefficients. 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.

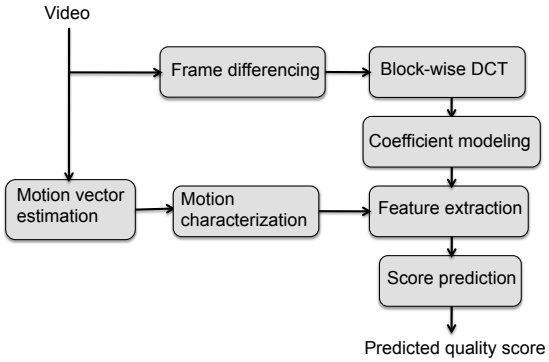

Fig. 2. Blind VQA framework

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 by weighting features derived under the NVS model of the previous step accordingly.

3. MOTION COHERENCY MODEL

We characterize a video's temporal content using a 2D *structure tensor* model applied to a video's computed motion vectors. 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 [3].

The motion coherence tensor summarizes the predominant motion directions over local neighborhoods. The 2D motion coherence tensor at a given pixel is given by:

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

where

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

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

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

where λ_1 and λ_2 are the eigenvalues of the motion coherence tensor.

Our model accounts for the magnitude of global motion. This is computed simply as the mode of the motion vectors between every two consecutive frames. The mean of the mode is computed across a video sequence and used as a feature during the score prediction phase.

4. NVS MODEL

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 [4], [2], [5].

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.

4.1. 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 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. 3 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

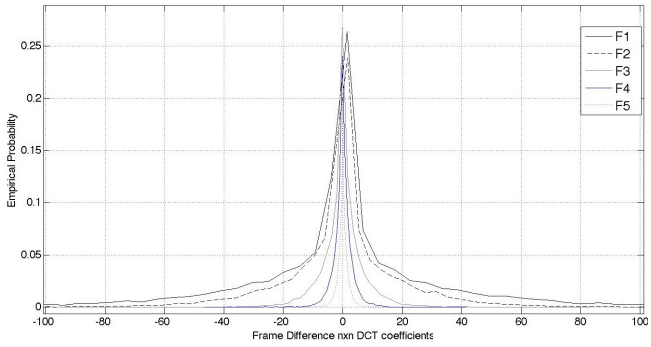

Fig. 3. Empirical distribution of DCT coefficients at 5 different frequencies from an $n \times n$ DCT decomposition of a frame-difference.

the use of a family of distributions that encompasses a range of tail behaviors. The 1-D generalized Gaussian density is a good fit to these coefficient histograms:

$$f(x|\alpha, \beta, \gamma) = \alpha e^{-(\beta|x-\mu|)^{\gamma}}, \tag{4}$$

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

$$\alpha = \frac{\beta\gamma}{2\Gamma(1/\gamma)}, \tag{5}$$

$$\beta = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}, \tag{6}$$

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

$$\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt. \tag{7}$$

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 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. This characterization is typically different between natural videos and distorted ones. The Video BLIINDS model aims to capture this statistical disparity and quantify it for perceptual video quality score prediction.

4.2. NVS Features

Each $n \times n$ matrix of shape-parameters per difference frame is partitioned into three sub-bands as depicted in Fig. 4, 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.

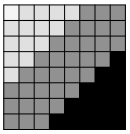

Fig. 4. Frequency band partition of frame differences. Top left: low frequency. Bottom right: high frequency

We then compute a percentile average of the shape parameters per band: The mean of the highest 10% of the shape parameters (γ) per band is computed. Thus for each frame-difference, the following statistical features are computed: 1) tenth-percentile low frequency band shape parameter, 2) tenth-percentile mid-band shape parameter, and 3) tenth-percentile high frequency band shape parameter.

The percentile averages are then weighted by the motion coherency measure C described in Section 3. Weighting by motion coherency is a simple and direct way to account for the extent to which the coherency of the motion affects the visibility of distortion in moving scenes.

4.3. Temporal Variation of 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 frame is also computed. 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

$$T_i = |D_{i+1} - D_i|, \tag{8}$$

where D_{i+1} and D_i are the average DC coefficients at frames indexed $i+1$ and i respectively. The mean of the highest 10%

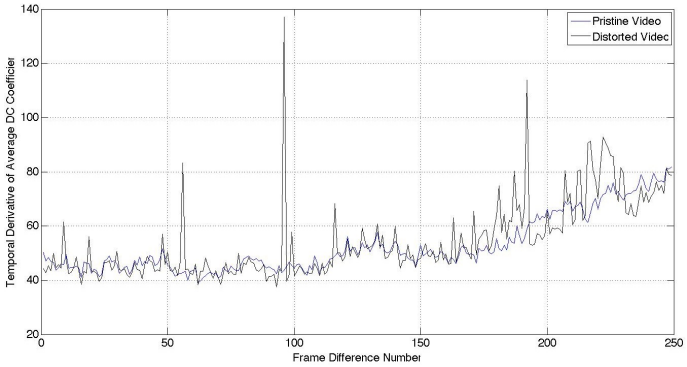

Fig. 5. Plot of the temporal derivative of mean DC coefficients for a pristine and a distorted video.

of the absolute discrete temporal derivatives is computed as a feature for prediction along with the other extracted features.

4.4. 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. We address the question of accounting for the temporal scale of the process by generating temporal scores in two ways: 1) by generating scores on an instantaneous (frame) basis, and 2) by integrating quality scores over 10 second intervals.

Since DMOS scores on VQA databases are usually only reported for complete video segments (10 seconds), we used the MS-SSIM index [6] applied on a frame basis against the reference video as a proxy for human scores. In this way it is possible to train the SVR to generate frame quality scores. Subjective DMOS scores were used to train another SVR to predict quality scores over 10 second video intervals.

In both cases, a linear kernel SVR based on the implementation in [7] was used to conduct quality score prediction.

5. EXPERIMENTS AND RESULTS

The algorithm was evaluated on the publicly available LIVE VQA database [8]. 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. We report the median SROCCs in Table 1, where we compare a number of models including full-reference PSNR and SSIM image quality indices. We also compare against two top-performing reduced reference VQA approaches VQM [9], Video RRED [10] and two leading full-reference VQA indices MOVIE [11] and ST-MAD [12]. Our approach outperforms PSNR, SSIM, and VQM, and is competitive with the performance of the RR-VQA RRED and the FR-VQA MOVIE and ST-MAD models. 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.

6. CONCLUSION

We have proposed a model-based, general (non-distortion specific) approach to NR-IQA using a minimal number of features extracted entirely from the DCT-domain which is also computationally convenient. We have shown that the new BLIINDS-II algorithm can be easily trained and it employs a simple probabilistic model for prediction. The method correlates highly with human visual perception of quality, and outperforms the *full-reference* PSNR measure and the recent no-reference BIQI index, and approaches the performance of the *full-reference* SSIM index.

7. REFERENCES

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

[2] R. Blake and R. Sekuler, *Perception*, McGraw Hill, 5th edition, 2006.

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

[4] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neurosci.*, vol. 5, pp. 598 – 604, May 2002.

Distortion	PSNR	SSIM	VQM	STMAD	MOVIE	RRED	Video-BLIINDS
MPEG-2	0.667	0.786	0.828	0.9484	0.9286	0.809	0.882
H.264	0.714	0.762	0.828	0.9286	0.9048	0.885	0.851
Wireless	0.680	0.714	0.714	0.7976	0.800	0.771	0.802
IP	0.660	0.600	0.770	0.7143	0.788	0.771	0.826
ALL	0.671	0.650	0.7451	0.825	0.807	0.826	0.821

Table 1. Median SROCC correlations on every possible combination of train/test set splits (subjective DMOS vs predicted DMOS). 80% of content used for training.

[5] 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, November 2011.

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

[7] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "Kernlab – an S4 package for kernel methods in R," *J. Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004, http://www.jstatsoft.org/v11/i09/.

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

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

[10] R. Soundararajan and A.C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circ. Syst. Video Technol.*, 2012, (to appear).

[11] 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, February 2010.

[12] P.V. Vu, C.T. Vu, and D.M. Chandler, "A spatio-ttemporal most apparent distortion model for video quality assessment," in *IEEE Int'l Conf. Image Process.*, 2011, pp. 2505–2508.