Blind Prediction of Natural Video Quality and H.264 Applications

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

***Abstract*—We propose two H.264 bit rate prediction and selection applications based on features extracted for our blind video quality assessment (VQA) algorithm, Video BLIINDS. We describe the blind VQA algorithm briefly, and we show that is correlates highly with human judgments of quality. We then demonstrate two applications: an H.264 bit rate predictor and an H.264 bit rate selector.**

***Keywords*-Video quality assessment, discrete cosine transform, natural scene statistics, motion modeling, H.264 compression.**

I. Introduction

The tumescent increase in video content that is being transmitted over wired and especially wireless networks, as a consequence of the rise in mobile device purchases (smart phones with increasingly larger screens, tablets, PDAs, laptops), has led to a tremendous increase in network traffic. Consequently, providers have sought to manage this traffic so as to ensure multimedia delivery at the available bandwidth capacity while ensuring adequate quality of experience.

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. This necessitates optimizing network and encoder parameters to meet consumer expectations of multimedia experience.

Until recently, there did not exist blind VQA algorithms that consistently and reliably correlate well with human judgments of temporal visual quality. Towards designing such a model, we have developed the Video BLIINDS framework, that utilizes a spatio-temporal model of DCT coefficient statistics to predict vidual quality. In this work, we will briefly review the Video BLIINDS model, and we introduce two applications of H.264 bit rate prediction and bit rate selection that rely on the features extracted by our VQA model. In the first application, *bit rate selection*, an H.264 encoding bit rate is selected given a certain desired visual video quality. In the second application, *bit rate prediction*, the bit rate at which an H.264 encoded video is predicted after its visual quality is assessed.

II. The Video BLIINDS Model

The algorithm is explained in detail in [1]. We shall however, briefly review it here.

Our approach "Natural Scene/Video Statistics" approach to VQA assumes the hypothesis that the human vision system has evolved in response to the physical properties of the natural environment [2], [3], and hence, the study of natural image/video statistics is highly relevant to understanding visual perception.

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. The approach to our VQA design leverages the fact that natural, undistorted videos exhibit statistical regularities that distinguish them from distorted videos where these regularities are destroyed, and that deviations from the NVS model, caused by the inflicting distortions, can be used to predict the perceptual quality of videos. Specifically, we propose an NVS model of DCT coefficients of frame-differences, since differencing consecutive frames reduces the high redundancy in videos, and captures the instantaneous (frame-wise) change along the temporal dimension.

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, capturing 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, namely the generalized Gaussian density.

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 to weight the parameters derived from the spatio-temporal NVS model of DCT coefficients. Features extracted under the

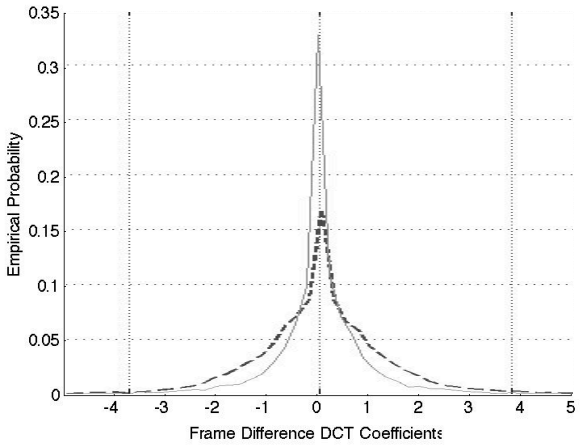

Fig. 1: Empirical probability distribution of frame-difference DCT coefficients of pristine and distorted videos. Dashed line: pristine video. Solid line: distorted video.

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.

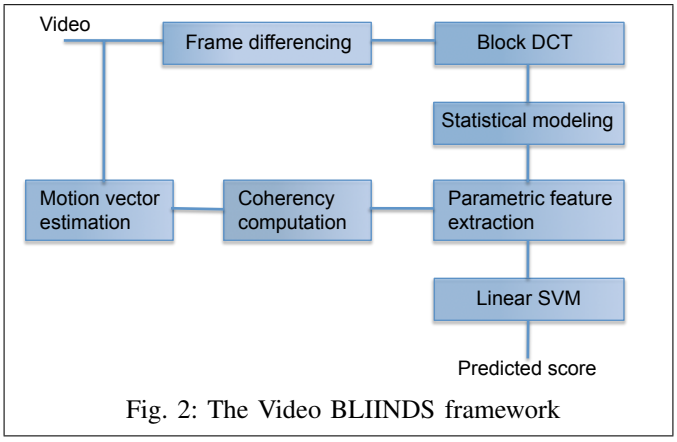

Fig. 2: The Video BLIINDS framework

A. 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 [4] 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 [5] was used to conduct quality score prediction.

III. Video BLIINDS Experiments and Results

The algorithm was evaluated on the LIVE VQA database [6]. 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 training it and testing it on each distortion type in isolation (i.e. making it distortion-aware), then we mixed the distortions together and applied the method on the mixture (i.e. making it distortion unaware). 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 ground truth scores of the database for every possible combination of train/test split. We report the median SROCCs in Table I, 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 [7], Video RRED [8] and two leading full-reference VQA indices MOVIE [9] and ST-MAD [10]. We computed the median Spearman rank order correlation coefficient between the predicted DMOS and the subjective DMOS scores of the database. The results are shown in Table I. 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.

IV. Practical Applications

The SROCC results show that the Video BLIINDS features are well suited for predicting the visual quality of videos compressed using H.264 compression. Thus, we now show how the features can be used in two useful applications involving H.264 compression.

The first application addresses the problem: Given an uncompressed video, how much can it be compressed to achieve a desired level of quality (expressed as 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. The second application addresses the problem: Given a video compressed by H.264, can the bit-rate at which it has been compressed be predicted? We demonstrate that the Video BLIINDS features can be used to address these questions.

Distortion	Full/Reduced-Reference VQA						Blind VQA
	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 I: Median SROCC correlations on every possible combination of train/test set splits (subjective DMOS vs predicted DMOS). 80% of content used for training.

In the first application, which we dubb 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 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. 3.

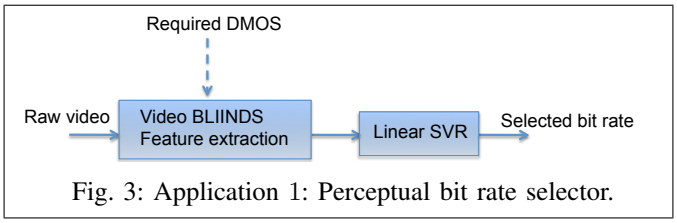

Fig. 3: Application 1: Perceptual bit rate selector.

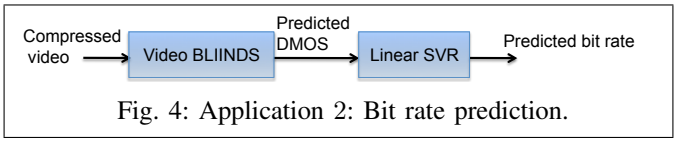

Fig. 4: Application 2: Bit rate prediction.

The second application which we dubb the *Video BLIINDS Bit Rate Predictor*, aims to predict the rate at which a video has already been compressed, using Video BLIINDS quality features. This process is summarized in Fig. 4.

The above two applications assume a particular choice of the H.264 encoder parameters. These are specified in [12]. I.e. given a particular configuration of the H.264 encoder parameters, it is possible to derive a mapping from visual quality to bit rate. This is inherent to the encoder parameters used on the videos comprising the training set from which the mapping was derived. The same assumption applies for both applications.

The applications were tested on the H.264 compressed portion of the LIVE VQA database. The details of the H.264 encoding parameters can be found in [6]. 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 selected 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. 5.

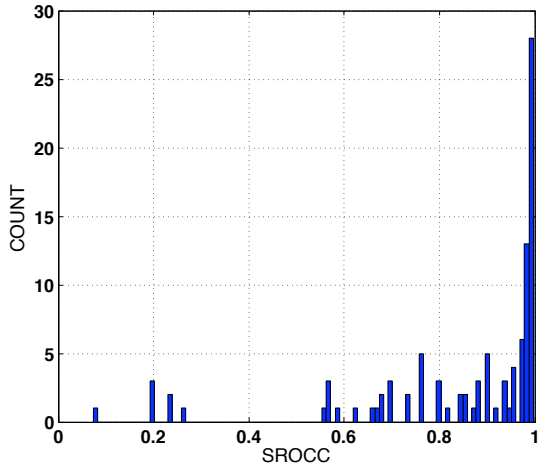

Fig. 5: Application 1: Histogram of SROCC between predicted and actual bit rates over 100 iterations of train/test splits.

It is important to note that although we might expect subjective video quality to vary monotonically with compression level, this relationship need not be strictly monotonic. For example, the perceived quality of a video might remain constant over a wide range of bit-rates. 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.

There is a concentration of SROCC values between 0.8 and 1, with a few outliers below 0.5. The performance of Application 1 depends on the cumulative error of first predicting the visual quality of the video, and then using the predicted score to predict the bit rate at which the video was 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.6.

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 quality to bit rate only sees the features extracted from the pristine videos

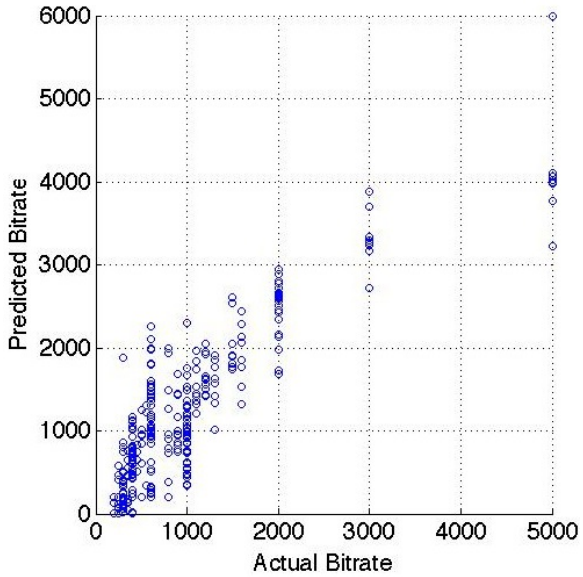

Fig. 6: Application 1: Scatter plot of predicted versus actual bit rates.

of the database and not from the compressed videos. The histogram of the obtained SROCC values is shown in Fig. 8. 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.7.

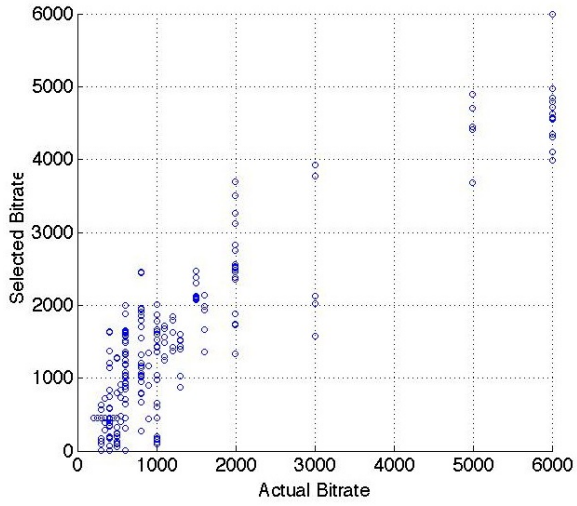

Fig. 7: Application 1: Scatter plot of selected versus actual bit rates.

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

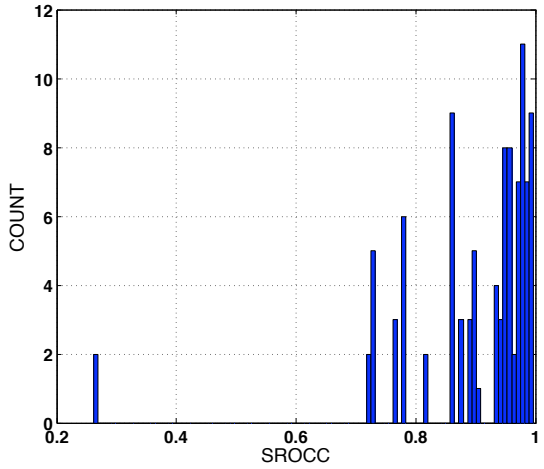

Fig. 8: Application 1: Histogram of SROCC between selected and actual bit rates over 100 iterations of train/test splits.

used in Video BLIINDS can be exploited for other perceptual optimization problems, such as tracking, denoising, deblocking, and so on.

V. Conclusion

We have described a natural scene statistic model-based approach to the no-reference/blind video quality assessment problem. The new Video BLIINDS model uses a small number of computationally convenient DCT-domain features. The method correlates highly with human visual judgments of quality. Additionally, we demonstrated two interesting applications of the Video BLIINDS features.

References

[1] M.A. Saad and A.C. Bovik, "Blind quality assessment of natural videos using motion coherency," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, November 2012.

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

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

[4] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity image quality assessment," in *Asilomar Conf. Sig, Syst., and Comp.*, November 2003, vol. 2, pp. 1398–1402.

[5] 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/.

[6] 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.

[7] 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.

[8] 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).

[9] 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.

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

[11] 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 Proc.*, vol. 21, no. 8, pp. 3339 – 3352, August 2012.

[12] *H.264/MPEG-4 AVC reference software manual*, 2007, Available: http://iphome.hhi.de/suehring/tml/JMX072).pdf.