H.264 VISUALLY LOSSLESS COMPRESSIBILITY INDEX: PSYCHOPHYSICS AND ALGORITHM DESIGN

Anush Krishna Moorthy and Alan Conrad Bovik

Dept. of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas - 78712, USA.

ABSTRACT

Although the term 'visually lossless' (VL) has been used liberally in the video compression literature, there does not seem to be a systematic evaluation of what it means for a video to be compressed visually lossless-ly. Here, we undertake a psychovisual study to infer the visually lossless threshold for H.264 compression of videos spanning a wide range of contents. Based on results from this study, we then propose a *compressibility index* which provides a measure of the appropriate bit-rate for VL H.264 compression of a video given texture (i.e., spatial activity) and motion (i.e., temporal activity) information. This compressibility index has been made available online at [1] in order to facilitate practical application of the research presented here and to further research in the area of VL compression.

1. INTRODUCTION

Lossy compression of visual stimuli such that the loss induced by the compression algorithm is not perceived by a human is referred to as visually lossless (VL) compression [2]. Visually lossless compression has received substantial attention from the medical imaging community [3, 4, 5, 6, 7], and to some extent in a general setting where perceptual thresholds or quantization matrices are developed for compression of visual signals in DCT/wavelet domain at the visually lossless threshold [8, 9]. Further, although research in medical imaging has focused on lossless compression of images using JPEG or JPEG2000 [5, 6], visually lossless compression of videos has not been explored as much [7]. Hence, our goal is to understand VL compression of videos and to develop a measure of compressibility that will aid in the VL compression of videos. Since the H.264 AVC currently enjoys industry acceptance and is the latest standard proposed by the video coding experts group (VCEG), we evaluate VL H.264 video compression.

Previous work in the area of VL compression of images has focused on medical images such as radiograms [4, 5]. The general approach followed is to compress medical images over a range of bit-rates and to present these images to experts in the field, who rate the similarity of the compressed image with respect to the uncompressed original. Such ratings are then used to decide a particular bit-rate for VL compression. In [7], a similar study was conducted for H.264 compressed bronchoscopy videos and a bit-rate for VL H.264 compression was inferred. The authors state that for videos with high motion, the VL bit-rate is 172 kBps, while for those with low motion, this rate is 108 kBps. Although the work in [7] is one of the few for H.264 compressed videos, wide-scale utilization of the results presented there are hindered by many factors. First, the study was specifically for medical videos and hence the bit-rates proposed do not correspond to a general setting. Second, it is not clear if two bronchoscopy videos can be referred to as different 'contents' - thereby limiting its applicability. Third, the resolution of these videos were 256 $\times$ 256 with a duration between 7-8 seconds, which again negates the possibility of it being used in a more general higher resolution setting for. Finally, even though the authors mention that the VL thresholds listed above were for 'high motion' and 'low motion' videos, no analysis is undertaken to algorithmically substantiate this claim. Here, we propose a method for VL H.264 compression of videos at resolutions larger than those considered in [7], where the application is not limited to medical videos, and the videos span a wide range of contents. As we shall see, such a general setting produces not only VL thresholds, but also allows for the development of a compressibility index for VL H.264 compression.

Our approach to understanding VL compression and the development of a compressibility index for VL compression is as follows. First, we conduct a single-stimulus 2-alternate forced choice (2-AFC) psychovisual study [10] in which human observers are shown a compressed video and its original uncompressed reference and are asked to choose that video which has better perceived quality. Subject responses are then analyzed using statistical techniques to produce a hypothesis on the lossless-ness of the video. These responses are then utilized in conjunction with algorithmically extracted measures of temporal activity and texture information (spatial activity) along with the corresponding bit-rates in order to produce a

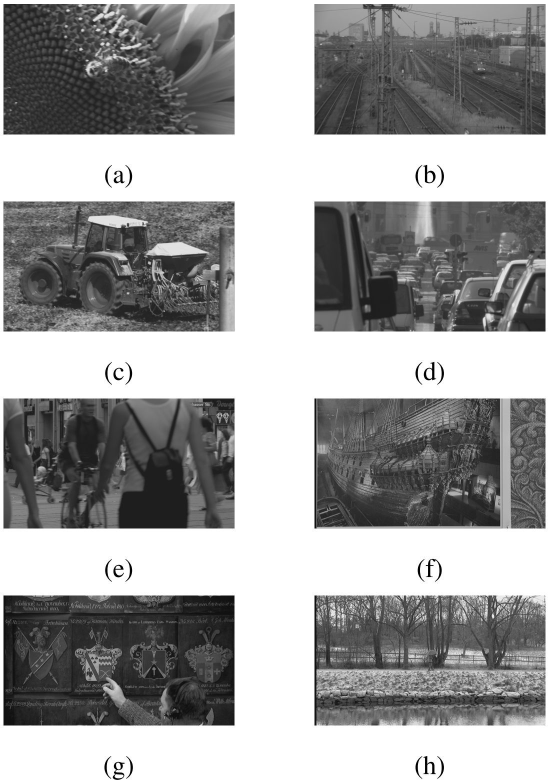

Fig. 1. (a)-(h) Frames of videos used in the study. Videos were sourced from the LIVE video quality assessment database [13].

compressibility index. One could view this approach as a dual to that followed by those who seek to assess the just-noticeable-distortion in videos [11, 12]. The H.264 visually lossless compressibility index (HVLCI) that we have created is now made available online at [1], in order to allow for practical application of the research presented here.

2. ASSESSING THE VISUALLY LOSSLESS THRESHOLD

2.1. The Videos

We used eight reference videos from the LIVE video quality assessment (VQA) database [13], a frame of each of them is seen in Fig. 1. The videos were chosen in order to encompass as wide a range of contents spanning different motion intensities and textural properties. The reader is referred to [13] for a detailed description of these videos. Videos (a)-(e) in Fig. 1 have a frame-rate of 25 fps, while (f)-(h) are at 50 fps - which were temporally down-sampled to 25fps, so as to allow for a generic definition of bit-rate levels for VL compression. All videos have a resolution of 768 × 432 and a duration of 10 seconds.

In order to create a set of compressed videos, each of the above reference videos was compressed using the JM reference software for H.264 encoding [14] with the following bit-rates - 0.5, 0.6198, 0.7684, 0.9526, 1.1810, 1.4640, 1.8150 and 2.2500 Mbps - to produce 40 (5 reference (a)-(e) × 8 bit-rates) + 24 (3 reference (f)-(h) × 8 bit-rates) = 64 compressed videos. The bit-rates were chosen based on a previous study (unpublished) conducted over a smaller set of (different) videos spanning a larger range of bit-rates in order to select the appropriate range for compression. Notice that the bit-rates are uniformly sampled on a log-scale between 0.5 Mbps and 2.25 Mbps. The baseline profile was used for encoding with an I-frame period of 16 and with R-D optimization enabled. A fixed number of macroblocks (36) were used per slice, with 3 slice-groups per frame and a dispersed flexible macroblock ordering (FMO) mode. The only parameter that was varied across videos was the bit-rate.

2.2. The Study

A single-stimulus two-alternate forced choice (2-AFC) task was conducted in order to measure the visually lossless (VL) threshold. In one interval, the reference video and one of the corresponding compressed videos were displayed one after the other on the center of a screen with a black background on a calibrated monitor in a room lit by artificial lights as per recommendations [15]. The viewing distance was 3 times the height of the video. At the end of each interval, the subject was asked to select which one of the two videos he/she thought had higher quality - the first video or the second. Apart from intervals consisting of compressed-reference presentations; for each content, the subject also viewed a reference-reference pair. The presence of such a pair was unknown to the subject and this reference-reference pair allowed for a statistical analysis based on hypothesis testing for detecting the visually lossless thresholds. Fig. 2 provides an illustration of the study.

Forty-five such presentation intervals corresponding to the 40 compressed videos + 5 reference videos formed one session of viewing, while the 24 compressed + 3 reference videos formed a second session; each of which lasted less than 30 minutes to reduce subject fatigue. The order of reference and compressed video in each interval was randomized for each presentation, and the sequence of compressed videos seen by the subject was randomized across subjects in order to reduce potential bias.

Fifteen subjects participated in the first session, while eight subjects participated in the second. Most of them were researchers in the area of image and video processing, however, they were unaware of the intent of the study. The subjects were not tested for acuity of vision but a verbal confirmation of (corrected) vision was obtained. The subjects were briefed before the study as follows : *You will be shown two videos on your screen one after the other. At the end of this presentation, you will be asked which video you thought*

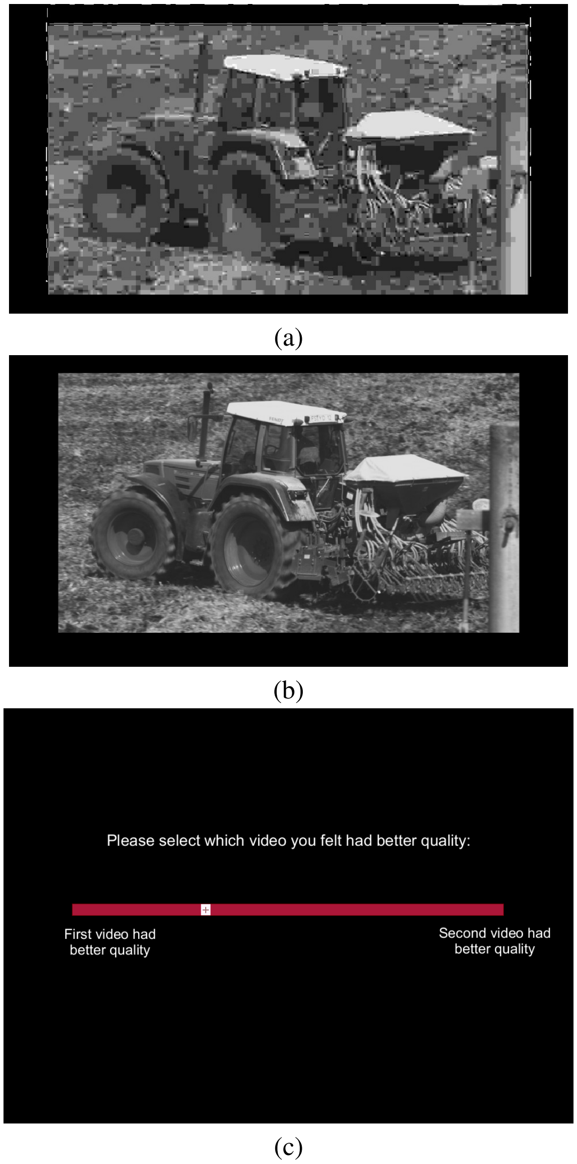
(a)
(b)
Please select which video you felt had better quality:
First video had better quality
Second video had better quality
(c)

Fig. 2. Study setup for determining VL thresholds. One interval consisted of two videos shown one after the other (here, compressed first (a), followed by reference (b)), after which the subject was asked to rate which video he/she thought had better quality (c).

had better quality - the first one or the second . You have to choose one of these two options. Once you make a choice the next set of videos will be played out and so on. Each subject underwent a short training session consisting of 3 pairs of presentations (the training videos were different from those in the actual study) in order to ensure that the subject was comfortable with the task. Once the actual study began, the subject was alone in the room and was not permitted to leave the room until he/she completed the study.

2.3. Assessing the visually lossless threshold

Once a set of (binary) preferences from each subject was collected, a statistical analysis followed. For each compressed video from each session a Wilcoxon rank sum test for equal medians was carried out to judge if the distribution of the scores assigned to the compressed video across subjects (in the compressed-reference case) had a median value equal to the distribution of scores assigned to the reference video in the reference-reference case [16]. The principle here being that, for VL compressed videos the distribution of the binary preference should match that of the reference-reference case, since the VL video would be perceived equivalent to the reference video.

The null hypothesis was that the two distributions (compressed video scores and reference video scores) come from distributions with equal medians. The results of such an analysis carried out at the 95% confidence level are seen in Table 1 for each of the videos from Fig. 1. A ‘0’ in the table indicates that the null hypothesis cannot be rejected at the 95% confidence level, and hence implies that the compressed video is perceptually identical to the reference video, and the corresponding bit-rate is the compression level for visually lossless compression of that video.

From Table 1, it is clear that for videos (b), (e) and (g), no consensus on the VL bit-rate could be achieved from this study, and hence we do not consider them for further analysis. Video (h) is already at the VL level at a bit-rate of 0.5 Mbps, and hence we consider 0.5 Mbps as the VL threshold for this video. In all videos, the highest bit-rate corresponding to a ‘0’ in Table 1 is used as the VL bit-rate.

3. ALGORITHMIC ANALYSIS & COMPRESSIBILITY

One could simply use the results of the above study and set a bit-rate for VL compression of videos as in [7] for videos with ‘low motion’ or ‘high motion’. However, a simple human classification into videos with high and low motion will not suffice if the final goal were to create a compressibility index which analyzes the video and produces a measure of visual lossless-ness. Hence, we analyze each of the videos from Fig. 1 for two parameters which are related to the complexity of the content and hence influence the bit-rates needed for VL compression - (1) the amount of texture (i.e., a measure of spatial activity) and (2) the amount of motion (i.e., a measure of temporal activity). We realize that there may be other important parameters that relate to the compressibility of a video, however, owing to the preliminary nature of the study and the limited size of the dataset under consideration, we confine ourselves to these two important factors. Future work will involve further analysis of possible factors which

Video	0.50	0.62	0.77	0.95	1.18	1.46	1.81	2.25
Video a	1	1	1	1	0	0	0	0
Video b	1	1	1	1	1	1	1	1
Video c	1	1	1	1	0	1	0	0
Video d	1	1	0	1	1	0	0	0
Video e	1	1	1	1	1	1	1	1
Video f	1	0	1	0	1	0	0	0
Video g	1	1	1	1	1	1	1	1
Video h	0	0	0	0	0	0	0	0

Table 1. Results of 2-AFC task. The first column lists videos corresponding to those in fig. 1. The rest of the columns are the result of a Wilcoxon rank sum test across bit-rates. A '0' in the table indicates that the null hypothesis (the two distributions have the same median) cannot be rejected at the 95% confidence level, and the corresponding bit-rate is the compression level for visually lossless compression of that video.

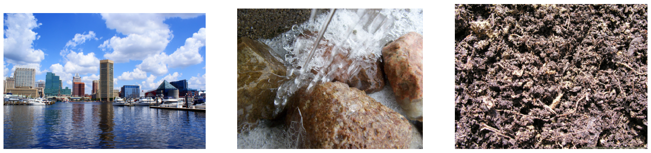

Fig. 3. (Left-to-right) Images with increasing amount of texture information - mean kurtosis of 17.20, 8.53 and 3.83 respectively. Images with greater texture have lower kurtosis.

influence compressibility[1].

Texture analysis has been a widely researched field and many models have been proposed to analyze and create textures [20, 21]. Our goal is not to classify or identify textures, but to simply provide a measure of the 'spatial activity' in a video using textural information. For this purpose, we utilize steerable filters to perform a wavelet decomposition [22]. Steerable filters have been previous used for texture analysis [20] and are attractive for our purpose owing to their increased orientation selectivity. In order to form a measure of textural information in a frame, the frame is decomposed using steerable filters over 3 scales and 8 orientations. Our research has lead us to believe that the kurtosis of coefficient distributions in each of the subbands is a good measure of the activity in a frame [18, 19]. In order to demonstrate that the simple kurtosis of subband coefficients captures spatial activity, in Fig. 3, we plot three images with increasing amounts of texture/activity and the mean kurtosis value across the 24 subbands for each of these images. It should be clear that kurtosis is negatively correlated with image activity and that the mean kurtosis across subbands is a good measure of the activity in an image.

Thus, in order to measure spatial activity in an frame of the video, the kurtosis of subband coefficients is averaged across subbands. The total spatial activity measure for a video is the median activity across frames. We have experimented with other statistical measures such as the mean and the coefficient of variation, however, we choose to use the median. The study of the optimal pooling strategy across frames is another interesting direction of research.

Temporal activity is measured using a modification of the technique described in [17][2]. Briefly, an absolute difference of adjacent frames is first performed, and the resulting sequence is decomposed into 4 pixel × 4 line × 0.2 second spatio-temporal (S-T) regions, whose standard deviation is computed. Perceptibility thresholds are then applied [17], and the mean value of the thresholded standard deviation is an indicator of the temporal activity across the frame. The temporal activity of the video is measured by the simple mean across frames. Both temporal and spatial activities are computed only on the luminance channel and we do not use any color information in our model.

In Fig. 4, we plot the videos in Fig. 1, as a function of their spatial and temporal activities, in order to demonstrate that the videos span a decent area of this space.

Finally, in order to produce a compressibility index, we perform a regression between the 2-tuple - $\mathbf{X}$ = (motion, texture) and the visually lossless bit-rate ($\mathbf{y}$) associated with each video obtained from the above rank-sum test. Support-vector regression using a ν-support vector machine (SVM) [23, 24] and a radial-basis function (RBF) kernel is utilized for this purpose. An SVM is utilized in order to abstain from specifying a particular functional form for relationships between $\mathbf{X}$ and $\mathbf{y}$. Since we have limited data, we used leave-out-one validation to test the performance of this approach. We trained the SVM on four out of the five videos with valid VL bit-rates and used this trained SVM to predict the VL threshold of the remaining video. This was repeated five-times, in order to predict the VL thresholds for each of the five videos. The results of such testing are listed in Table 2, where the

[1] We note that we have experimented with some other measures, including optical flow information, spatial activity as defined in [17], and subband features as defined in [18, 19]. While some of these measures are useful as indicators of visual lossless-ness, the measures chosen here were better suited to the task at hand.

[2] The parameter corresponding to temporal activity in [17] is *ct_ati_gain*.

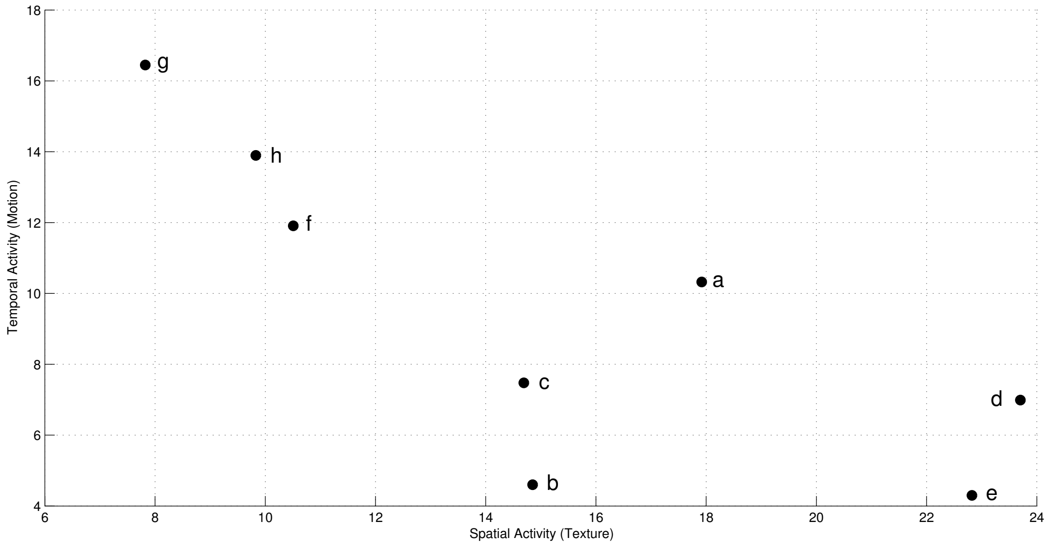

Fig. 4. Plot of spatial (x-axis) activity vs. temporal (y-axis) activity for videos used in this study. Points a-h correspond to videos (a) - (h) from Fig. 1.

Video	a	c	d	f	h
Actual	0.9500	1.4600	1.1800	1.1800	0.5000
Predicted	1.0923	1.2651	1.1981	1.0189	0.9365

Table 2. Ground truth and predicted visually lossless bit-rates for videos (a), (c), (d), (f) and (h) from fig. 1.

ground-truth VL bit-rates and the predicted VL bit-rates are tabulated.

As Table 2 demonstrates, the simple measures of spatial and temporal activity predict the VL threshold with good accuracy – the mean squared error between the actual VL bit-rates and the predicted bit-rates across videos is 0.0153. Since we are unaware of any other such measure for VL H.264 compression, a performance comparison with other objective measures is impossible.

Thus, our final implementation of the compressibility index for visually lossless H.264 compression (HVLCI) - available online at [1] - consists of a support vector machine that regresses measures of motion and texture onto a bit-rate corresponding to the visually lossless threshold for that video.

Although extremely useful as a tool, HVLCI estimates of visually lossless thresholds must be utilized with caution. Owing to the limited amount of data that was used to produce the index and the small number of factors used to analyze the video, it is possible that HVLCI estimates may not always correlate well with human perception across videos although we have not observed this. This, however, will easily be remedied in future work where a study involving a larger set of videos will be undertaken and other factors which contribute to the VL threshold will also be analyzed. In spite of these drawbacks, this work is important as one of the first to systematically study visually lossless compression via H.264 and to produce a practical tool - HVLCI - that estimates the visually lossless threshold using a small set of parameters.

4. CONCLUSION

We conducted one of the first systematic psychovisual studies to estimate the visually lossless threshold for H.264 compression. The videos were then analyzed and quantified based on the activity measures of texture (i.e, spatial activity) and motion (i.e., temporal activity) in order to produce an index for visually lossless H.264 compression - HVLCI. The proposed index has been made available online [1] in order to allow for practical application and research in the area of visually lossless compression. Future work will involve increasing the number of videos in the psychovisual study and analyzing various other parameters that may influence lossless compression in order to produce a more robust version of HVLCI.

5. REFERENCES

[1] A. K. Moorthy and A. C. Bovik, “H.264 Visually Lossless Compressibility Index (HVLCI), Software release,” *http://live.ece.utexas.edu/research/quality/hvlci.zip*, 2010.

[2] L. J. Karam, “Lossless image compression,” in *The Essential Gudie to Image Processing*, Al Bovik, Ed., pp. 385–417. Elsevier Academic Press, 2009.

[3] O. Kocsis, L. Costaridou, L. Varaki, E. Likaki, C. Kalogeropoulou, S. Skiadopoulos, and G. Panayiotakis, “Visually lossless threshold determination for microcalcification detection in wavelet compressed mammograms,” *European Radiology*, vol. 13, no. 10, pp. 2390–2396, 2003.

[4] R.M. Slone, D.H. Foos, B.R. Whiting, E. Muka, D.A. Rubin, T.K. Pilgram, K.S. Kohm, S.S. Young, P. Ho, and D.D. Hendrickson, “Assessment of Visually Lossless Irreversible Image Compression: Comparison of Three Methods by Using an Image-Comparison Workstation1,” *Radiology*, vol. 215, no. 2, pp. 543, 2000.

[5] R.M. Slone, E. Muka, and T.K. Pilgram, “Irreversible JPEG Compression of Digital Chest Radiographs for Primary Interpretation: Assessment of Visually Lossless Threshold1,” *Radiology*, vol. 228, no. 2, pp. 425, 2003.

[6] K.H. Lee, Y.H. Kim, B.H. Kim, K.J. Kim, T.J. Kim, H.J. Kim, and S. Hahn, “Irreversible JPEG 2000 compression of abdominal CT for primary interpretation: assessment of visually lossless threshold,” *European radiology*, vol. 17, no. 6, pp. 1529–1534, 2007.

[7] A. Przelaskowski and R. Jozwiak, “Compression of Bronchoscopy Video: Coding Usefulness and Efficiency Assessment,” *Information Technologies in Biomedicine*, p. 208, 2008.

[8] A.B. Watson, “DCT quantization matrices visually optimized for individual images,” in *Proceedings of SPIE*, 1993, vol. 1913, pp. 202–216.

[9] A.B. Watson, G.Y. Yang, J.A. Solomon, and J. Villasenor, “Visibility of wavelet quantization noise,” *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164–1175, 1997.

[10] R. Sekuler and R. Blake, *Perception*, McGraw Hill, 2002.

[11] A.B. Watson and L. Kreslake, “Measurement of visual impairment scales for digital video,” *Proceedings of SPIE*, vol. 4299, pp. 79–89, 2001.

[12] X. K. Yang, W. S. Ling, Z. K. Lu, E. P. Ong, and S. S. Yao, “Just noticeable distortion model and its applications in video coding,” *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 662–680, 2005.

[13] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[14] H.264/mpeg-4 avc reference software manual, “http://iphome.hhi.de/suehring/tml/ jmreferencesoftwaremanual(jvt-x072).pdf,” 2007.

[15] International Telecommuncation Union, “BT-500-11: Methodology for the subjective assessment of the quality of television pictures,” .

[16] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, CRC Pr I Llc, 2004.

[17] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, , no. 3, pp. 312–313, Sept. 2004.

[18] A. K. Moorthy and A. C. Bovik, “A two-step framework for constructing blind image quality indices,” *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 587–599, May 2010.

[19] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE Transactions on Image Processing*, submitted.

[20] J. Portilla and E.P. Simoncelli, “A parametric texture model based on joint statistics of complex wavelet coefficients,” *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[21] A. C. Bovik, M. Clark, and W. S. Geisler, “Multichannel texture analysis using localized spatial filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 55–73, 1990.

[22] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, “Shiftable multiscale transforms,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, 1992.

[23] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, “New support vector algorithms,” *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[24] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2000.