Visually Lossless H.264 Compression of Natural Videos

Anish Mittal, Anush K. Moorthy and Alan C. Bovik

Laboratory for Image and Video Engineering (LIVE),
The University of Texas at Austin,
Austin, Texas, USA.
Email: mittal.anish@gmail.com

We performed a systematic evaluation of visually lossless (VL) threshold selection for H.264/AVC (Advanced Video Coding) compressed natural videos spanning a wide range of content and motion. A psychovisual study was conducted using a 2AFC (alternative forced choice) task design, where by a series of reference vs. compressed video pairs were displayed to the subjects, where bit rates were varied to achieve a spread in the amount of compression. A statistical analysis was conducted on this data to estimate the VL threshold. Based on the visual thresholds estimated from the observed human ratings, we learn a mapping from 'perceptually relevant' statistical video features that capture visual lossless-ness, to statistically-determined VL threshold. Using this VL threshold, we derive an H.264 Compressibility Index. This new Compressibility Index is shown to correlate well with human subjective judgments of VL thresholds. We have also made the code for compressibility index available online [1] for its use in practical applications and facilitate future research in this area.

Keywords: Visually lossless compression, psycho-visual study, visually lossless threshold, distorted video, H.264/AVC, compressibility index

Received 01 March 2012

1. INTRODUCTION

A simple interpretation of the human visual system is that it functions akin to a camera, capturing the world around us and transmitting the data onto the brain which in turn produces streams of visual stimuli that form a continuous narrative. However, seeing in itself is a deception. While the human visual system is similar to a camera, and, the human eye is indeed a lens-sensor-processor arrangement, human vision goes beyond the simplistic recreation of the world that current cameras offers. The human visual system (HVS) is a sophisticated machine where millions of neurons work in unison to accomplish complicated tasks such as depth perception, object recognition, motion computation and so on. The fact that most humans fail to realize the sophisticated processes of data winnowing, organization, inference, and discovering underlying our visual interpretation of the world is testimony to the efficiency and speed by which the HVS processes and interprets visual data. As the vision scientist, D. D. Hoffman put it, *Vision is more like Sherlock Holmes than movie cameras* [2].

While the human brain devotes a significant amount of processing to visual information (30%, as against 8% and 3% for tactile and auditory information respectively), the HVS cannot record all possible shapes, colors, edges, motions, objects, and all of the features of the visual world. Interpreting a continuum of light and information using a large number of discrete neurons and photoreceptors has necessitated the evolution of intelligent sampling and estimation strategies, so that the loss of information does not lead to significant handicaps. An illuminating example is that of human color perception. The human retina subsamples the entire color spectrum using just three broad types of color receptor cone cells, which implies that different light spectra produce the exact same sensory response in humans [3, 4]. These *metamers* have been extensively studied and demonstrate that while the visual system is adept at using the subsampled information, the HVS can be decieved by certain spatial arrangements of the stimulus.

There exist many examples of visual stimuli (and illusions) that reveal the ways in which HVS achieves efficient processing. The metameric behavior that the HVS design exhibits has been exploited by engineers for image and video processing. The relatively higher sensitivity of the HVS to luminance information than chrominance (color) information has been exploited in the design of video compression algorithms, where more bits are assigned to each piece of luminance

information (on average) than to corresponding pieces in the co-located color channels. Different visual stimuli can evoke the same percept when the visual system fuses information from different cues to build robust perception. Recent experiments on cue conflicts [5] have shown that the visual system estimates a particular scene parameter as a weighted average of each cue and hence, multiple cue-conflict stimuli can evoke the same perception. This phenomenon has found applications in stereo image coding [6] where humans use both monocular and binocular cues to perceive the 3D world. When the left and right eyes provide monocular images to the brain that contain different spatial frequency characteristics (as may occur when one eye is 'weaker' than the other), the HVS tends to use the higher resolution information. This is an example of what is referred to as binocular rivalry.

In the same vein, the HVS perceptual mechanisms may fail to differentiate between two similar signals, when the two signals are present in similar spectral, temporal (in case of video) or spatial locations [7]. Indeed one of the signals may be rendered less visible (or invisible) by the other, which is visual masking. An important example is masking of distortion by video content, which is an important focus of this article.

Visually lossless compression refers to lossy compression of images or videos that does not produce distortions that are perceived by an average human [8]. While there has been substantial interest in visually lossless (VL) compression of medical image data, the field of general-purpose VL compression of videos has not been deeply studied. In this article, we conduct a deep examination of the question of VL compression of videos compressed using the H.264/AVC [9] encoder.

We first review prior work in the area of VL compression, and discuss the drawbacks of the techniques proposed. Then we describe the H.264/AVC standard and human study that we conducted to reveal the factors that lead to visual lossless-ness of H.264 compressed videos. The results of the study are statistically evaluated using statistical techniques and inferences are made regarding the predictability of lossless-ness of compression and its relationship to the content. We then describe a simple algorithm that extracts statistical motion and image based features from source videos, and using this information predicts visually lossless thresholds, assuming the source video is to be compressed using the H.264 encoder. We also identify limitations of the approach and describe avenues of future work in the area of visually lossless compression.

2. PREVIOUS WORK

Substantial research has been devoted towards achieving better video compression via such diverse mechanisms as spatial and temporal subsampling while encoding, and learning based interpolation schemes for decoding; to deliver acceptable levels of visual quality at the receiver [10, 11, 12, 13]. However, visually lossless (VL) compression of images and videos has received attention largely from researchers in the medical imaging community [14, 15, 16, 17, 18], where maintenance of government mandated records has led to an explosion of image data. When compressing medical images (for example, radiograms) it is imperative to be able to faithfully retain all of the visual information that might be needed to make accurate diagnoses. Such diagnostically lossless compression for images has generally been studied in the context of the JPEG and JPEG2000 compression standards, whereby user studies are conducted and acceptable compression bit-rates are decided based on diagnostic analysis. VL studies are generally performed using an AFC [19], where subjects are asked to view two images – one compressed and the other uncompressed – and asked which of the two images they thought had better quality. In case a decision cannot be made, the subject is asked to randomly pick one of the images. An appropriate statistical analysis of the results then leads to determination of the VL threshold (ideally at chance).

Researchers in the image and video processing community have explored the related area of just noticeable distortion (JND) for images and videos [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. JND studies involve quantifying (either through human studies, such as those in VL compression, or through algorithms) the amount of distortion that can be added to a 'clean' image, before it becomes apparent. In human studies that are designed to produce an estimate of the JND, images and videos are compressed and the subjects are asked to search for distortions in them. In case the subject is unable to perceive any distortions (whether present or not) the chosen compression level is deemed to be below the JND threshold.

JND studies have been used to propose JND models for compression which guarantee that the distortion is not perceived. Such models are defined in either the DCT, wavelet or spatial domain [25, 26, 27, 28, 29, 30, 31, 32, 22, 23, 24] and the models have been used to design perceptually-optimal quantization matrices or to perform perceptually-relevant watermarking. Specifically, two JND models have proposed [33], [34] recently for H.264/AVC standard.

The JND studies and the VL studies have a lot in common, since both seek to define thresholds beyond which distortions become imperceptible to the observer. Yet while both are similar, these inquiries do not address the same problem and indeed may be considered as duals of each other. The JND studies require the subject to actively scan the image/video for the presence of distortion, while the VL studies simply ask the subject to rate which image/video s/he thought had better quality. Since the user is not actively scanning for distortions in the latter case, one would hypothesize that VL thresholds may be higher

(i.e., may allow for higher compression) than the JND thresholds. Further, picking the best quality video is usually an easier task than conducting an active scan for distortions, since the user may respond to non-compression related anomalies (e.g., lens aberration, motion jitter etc.) which can affect JND curves.

Much current research on lossless compression has dealt with the compression of images using JPEG/JPEG2000 encoders [18, 15]. However, the field of visually lossless compression of videos have seen only a limited amount of work, even in the medical field [14].

In [14], a 2-AFC task [19] was designed in which bronchoscopy videos compressed over a wide range of bit-rates using the H.264 codec [9] were shown to subjects along with the uncompressed original. Based on the user ratings, the VL threshold was found to be 172 Kbps for high motion video but only 108 Kbps for the low motion one. Although the study systematically tackled the problem of VL compression of medical videos, results of the work have a number of drawbacks which make it difficult to use the study results to predict VL compression levels in a general purpose video setting.

First, the study was conducted specifically for medical videos and one may argue that the achieved bit-rates reflect the rates for *diagnostic lossless-ness* rather than *visual lossless-ness*. Second, the study incorporated only two distinct videos, both of which were videos of bronchoscopy, rendering any definitive conclusions regarding the effect of content impossible. Third, the videos used in the study were of low resolution (256×256) and spanned only 7-8 seconds in duration. With the recent explosion of high video resolutions, the results may have reduced relevance. Fourthly, while the authors classify the videos as low and high motion, they do not undertake any analysis to support the claim of motion, nor is an algorithm proposed to quantify the VL threshold for future videos. Lastly, the specific image modality produces videos that are quite different from common consumer grade optical videos, with very different statistical properties and responses in compression.

To address these concerns and to provide a general purpose tool that predicts the VL thresholds of commonly encountered types of videos, we conducted a VL compression study using a 2-AFC task. The videos used were of high resolution and spanned a wide variety of contents. The industry standard H.264/AVC explained in next section was used for compression. In the remainder of this article we describe the user study, and propose an algorithm that we have derived to evaluate the visually lossless threshold of H.264 compressed videos.

3. H.264/AVC STANDARD

H.264/AVC is a standard for video coding developed by ITU-T Video Coding Experts Group (VCEG) together with the International Organization for Standardization (ISO) and the Moving Picture Experts Group (MPEG)[35]. The standardization of the first version of H.264/AVC was completed in May 2003 and has been rapidly gaining popularity since then due to its superior compression efficiency. It has been designed to be flexible enough for use in varied applications, including low and high bit rates, low and high resolution video, RTP/IP packet networks, broadcast, DVD storage, and ITU-T multimedia telephony systems. It is currently used by streaming internet sources, such as YouTube, Vimeo, and iTunes; web software such as Microsoft Silverlight and Adobe Flash Player; and codec standards for Blu-ray Discs.

It is a 'family of standards' where different profiles have been developed to cater to different applications. Several new features have been added to this standard including higher quality video coding by using increased sample bit depth precision, adaptive switching between 4×4 and 8×8 integer transforms, higher-resolution color information and additional color spaces support, encoder-specified perceptually relevant quantization weighting matrices and efficient inter-picture lossless coding.

Scalable Video Coding (SVC) and Multiview Video Coding (MVC) are other major features in the standard. The aim of SVC was to provide flexibility of encoding a high-quality bitstream using a set of bit streams which can be individually decoded using the H.264/AVC design. The subset streams can be derived by dropping other packets from the larger bitstream where each substream can represent a lower spatial resolution/temporal resolution/video quality. This enables video decoding at different bit rates depending on available bandwidth. On the other hand, MVC provides the option to generate bitstreams that can represent more than one view of a video. This functionality has obvious applications in stereoscopic 3D video coding.

4. ASSESSMENT OF VISUALLY LOSSLESS THRESHOLD

4.1. The Videos

To create a set of distorted videos for the study, we used 8 uncompressed raw source videos of high quality that are freely available for researchers from the Technical University of Munich [36]. This content has been previously used as source material in the LIVE Video Quality Assessment (VQA) Database [37], which is currently a popular database used to test the performance of VQA algorithms [38, 39, 40, 41, 42, 43].

The videos used in the study are all naturalistic, meaning they were captured with an ordinary camera under ordinary conditions with as little distortions as possible. These natural videos are also free of synthetic content, such as graphics, animations, text overlays or computer modified data. The features that we extract

from these videos derive from perceptually relevant natural scene statistic (NSS) models. These kinds of models have been used to both understand and explain certain properties of visual perception,and to model distortion processes in images and videos for engineering quality assessment algorithms [44, 45, 46, 47, 48, 49, 50, 51].

The videos were shot using a professional cinematic camera and are recorded in YUV420P format, without audio. While the relationship between audio and video has been studied [52, 53, 54], and certain conclusions may be made from such work, we do not attempt to study the interaction between the two here, choosing instead, to study the visual lossless-ness of videos only, to isolate the effects of compression on visual experience. Given the development of similar models for audio, along with the models of audio-visual interaction, audio-video lossless compression measures are a logical evolution of this work.

The videos were down sampled from 1280×720 to a resolution of 768×432, using MATLAB's `imresize` function which uses bicubic interpolation. This was done due to resource constraints with respect to the graphics card and the processor usage. Figure 1 shows one of the frames from each reference video used in the study, and a short description of the scenes follows:

(a) Sunflower - Still camera, shows a bee moving over a sunflower in close-up
(b) Station - Still camera, shows a railway track, a train and some people walking across the track
(c) Tractor - Camera pan, shows a tractor moving across some fields
(d) Rush hour - Still camera, shows rush hour traffic on a street
(e) Pedestrian area - Still camera, shows some people walking about in a street intersection
(f) Mobile & Calendar - Camera pan, toy train moving horizontally with a calendar moving vertically in the background
(g) Shields - Camera pans at first, then becomes still and zooms in; shows a person walking across a display pointing at it
(h) Park run - Camera pan, a person running across a park

Amongst the source videos, the first five videos ((a)-(e)) had a frame rate of 25 fps, while the remaining three videos had a frame rate of 50 fps. In order to ensure uniformity of content, these three videos were temporally down sampled to a frame-rate of 25 fps by dropping every other frame.

In order to create the distorted videos, we used the JM reference encoder [55] for H.264/AVC compression [9]. Each reference video was compressed using 8 different bit rates, sampled uniformly on a log scale: 0.5, 0.6198, 0.7684, 0.9526, 1.1810, 1.4640, 1.18150 and 2.2500 Mbps. The bit rate range selection for this study was done based on a small (unpublished) study conducted over a smaller set of (different) videos compressed over a larger set of bit rates. The baseline profile of the H.264 encoder was used for compression with an I-frame period of 16 and with the rate-distortion optimization option enabled. 3 slice groups were used per frame with 36 macro blocks per slice and a dispersed flexible macro block ordering (FMO) mode. While varying these parameters and studying their effects on visual quality remain of interest, in the present work, we restricted ourselves to studying the visually lossless threshold as a function of the bit-rate.

Thus, a total of 40 (5 reference (a)-(e) $\times$ 8 bit-rates) + 24 (3 reference (f)-(h) $\times$ 8 bit rates) = 64 compressed (distorted) videos were created, which were then used for the human study.

4.2. The Study

A single-stimulus, two-alternative forced choice (2-AFC) task was conducted at the LIVE lab, The University of Texas at Austin, with voluntary recruitment. Subjects taking part in the study were mostly male graduate students with age varying from 20 to 30 years. In a the 2-AFC task, the subject has to pick one of two choices, and does not have an option of picking neither; hence the moniker 'forced'.

During each interval of the study, the subject was shown a pair of videos separated by a brief pause - the reference video and one of its compressed versions. Multiple such presentations formed one session of the study. The order of presentation of the videos was randomized such that no two consecutive intervals consisted of the same content. Further, within each interval, the order in which the reference and the distorted videos were shown was randomized as well. Randomization serves to limit bias inherent to human opinion. Apart from the compressed-reference pairs, the subjects also viewed reference-reference pairs in the session, albeit without his/her knowledge of their presence. The distribution of rating that the subject gives these reference-reference pairs serve as thresholds, which may be used for comparison with the scores given for the reference-compressed pairs.

The videos were displayed on a 21" calibrated CRT monitor, set at a resolution of 1024×768 in a well-lit room. The study environment was in line with recommendations [56]. The videos were displayed on the center of the screen with a black border surrounding the video content and the subjects viewed the videos from a distance of three times the height of the video [56]. Since the videos had a frame-rate of 25 fps and the monitor refresh rate was set at 50 Hz, each frame of the video was displayed twice.

The videos were displayed using MATLAB and the XGL toolbox [57]. The XGL toolbox, developed at UT Austin, allows for precise display of controlled stimuli for psychovisual experiments. Using the XGL toolbox, the YUV videos were first fed into the graphics

FIGURE 1. (a)-(h) Sample frames of videos used in the study. Videos were sourced from the LIVE Video Quality Assessment Database [37].

buffer. Playback was performed only after the entire video was loaded onto the buffer, to ensure that the only distortions that the subject would see are those arising from the controlled distortion process (H.264 compression), as opposed to any occurring during playback (such as stuttering). At the end of each interval, a choice was made using a mouse and the scores were collected using specially designed software for this purpose. Figure 2 illustrates the study setup.

The study was divided in two sessions: 40 compressed/reference + 5 reference/reference video pairs were shown in one session and 24 compressed/reference + 3 reference/reference video pairs were shown in the other. Both sessions lasted for less than 30 minutes to limit subject fatigue [56]. Fifteen subjects participated in the first session while eight subjects took part in the second. The subjects were unaware of the purpose of the study. No visual acuity test was performed, but a verbal confirmation of (corrected) vision was obtained from the subjects. This follows our philosophy of conducting image and video quality related human studies: to report the result of studies using a set of individuals likely to be representative of those viewing consumer videos.

All subjects were given same instructions to ensure uniformity: *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 had better quality - the first or the second. You have to choose one of the two options. Once you make a choice the next set of videos will be played and so on.*

In order to familiarize the subject with the task, and to ensure a minimum level of comfort with the study environment, a small training session was conducted at the start of each session. The training session consisted of three intervals, each consisting of a pair of videos. The videos used in the training session differed from those used in the actual study. The subject was allowed to ask questions during the training session but once the test session started, s/he was alone in the room and was asked not to leave the room until the session was completed.

4.3. Analysis of Subjective Opinion

Once the (binary) preferences were collected from each user, we performed a statistical analysis to estimate the visually lossless threshold for each of the videos in the study. We used the Wilcoxon sum rank test for equal medians [58] to judge whether the distribution of scores assigned to the compressed video (in the compressed-reference case) had a median value equal to that of the scores assigned to the reference video. Ideally, the subject would have

(a) (b)

(c)

FIGURE 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).

randomly chosen the 'better' reference video from the reference/reference presentations, leading to an even symmetric distribution. However, to account for human bias we directly compared the compressed/reference results with the reference/reference results. In case the distributions match (in the statistical sense of their medians matching) then the compressed video is presumed visually lossless, else the subject is deemed able to perceive the distortions. Fig. 3 shows the distribution of (binary) preferences collected from users for the rush hour video where the distorted video was obtained by compressing at 0.6198 Mbps. '-1' indicates the (binary) preference where subjects rated second video as having better quality than first video whereas '1' indicates vice-versa. (a) shows the distribution of preferences for reference-reference pair whereas (b) shows the distribution for distorted-reference pair. We can observe from the distributions that most of the subjects were able to correctly judge the distorted video from the distorted-reference pair whereas judgment was random for reference-reference pair.

The null hypothesis was that the two distributions (compressed/reference video scores and reference/reference video scores) come from distributions with equal medians. The results of such an analysis carried out at the 95% confidence level are given in Ta-

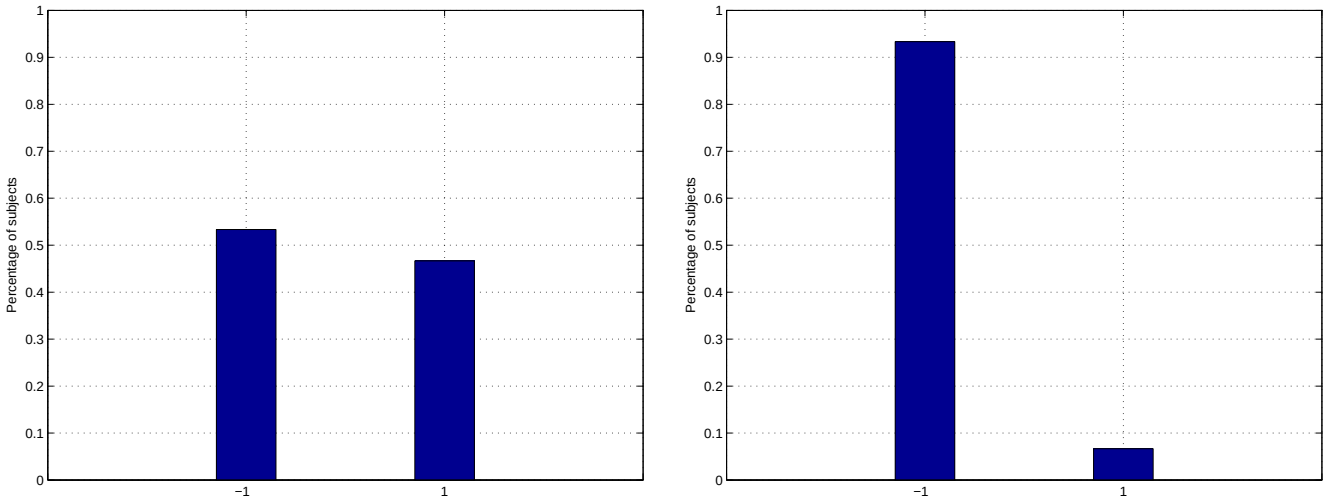

FIGURE 3. Shows the distribution of (binary) preferences collected from users for the rush hour video where the ddistorted video was obtained by compressing at 0.6198 Mbps. (a) shows the distribution of preferences for reference-reference pair whereas (b) shows the distribution for distorted-reference pair. '−1' indicates the (binary) preference where subjects rated second video as having better quality than first video whereas '1' indicates vice-versa.

ble 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, which implies that the compressed video is perceptually identical to the reference video.

As Table 1 indicates, for videos (b), (e) and (g), no consensus could be reached on the VL threshold. While this might be due to limited number of subjects who participated in the study, it is also possible that the bit rates used did not entirely span the range from visually lossless to completely perceptible for these contents. In any case, we do not consider these videos in the subsequent algorithm analysis.

In all videos, the highest bit-rate corresponding to a '1' in Table 1 is used as the VL bit-rate. However, 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.

5. ALGORITHM ANALYSIS

Having statistically analyzed the visually lossless threshold, we could simply have stated the visually lossless bit-rates as in [14], based on an arbitrary classification based on motion or spatial activity. However, a simple human classification may be insufficient to address the general scenario where a previously unseen video is to be analyzed to automatically extract a VL threshold. Hence, as a preliminary step, we analyze the videos by extracting measures of (1) spatial activity and (2) temporal activity, which we then use to create an algorithm that is capable of predicting the VL threshold for a previously unseen video. Spatial activity is a measure of the amount of variation in the scene, while temporal activity is a measure of the amount of motion. Both of these are related to perceptual masking and hence, to distortion visibility.

5.1. Spatial Activity & Variation

We define a simple measure of spatial activity that is based on the steerable pyramid decomposition [59]. The steerable pyramid decomposition is an overcomplete wavelet transform that has been widely used for describing the statistics of natural images [60], for image quality assessment [47, 48, 49] and for texture analysis [61]. The steerable pyramid decomposition allows for a multi-scale multi-orientation decomposition similar to that which is hypothesized to occur in area V1 of the primary visual cortex [62].

Each frame of the video is decomposed using the steerable pyramid decomposition over 3 scales and 8 orientations. Previously, we have demonstrated that the kurtosis (the ratio of the fourth moment to the squared second moment) is a good measure of spatial activity in an image [48, 49]. Band pass filter coefficients of a low activity sub band have larger kurtosis value attributed to most coefficients having near zero values causing probability distribution to be peaky. The measure of activity is simply the mean kurtosis value across sub bands. In order to demonstrate that the kurtosis is indeed a good measure of activity, in Fig. 4, we plot three images with increasing amounts of texture/activity and the mean kurtosis value across the 24 sub bands for each of these images. Clearly, kurtosis is negatively correlated with image activity and our measure of mean kurtosis across sub bands is a good descriptor of spatial activity in an image.

We compute the mean kurtosis across the sub bands as a measure of spatial activity and compute the median activity across frames to produce an overall

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. In all videos, the highest bit-rate corresponding to a '1' is used as the VL bit-rate.

FIGURE 4. (Left-to-right) Images with increasing amount of spatial activity(energy) - mean kurtosis of 17.20, 8.53 and 3.83 respectively. Images with greater texture have lower kurtosis.

measure of spatial activity of the video. While we have experimented with other temporal aggregation (pooling) strategies such as the mean and the coefficient of variation, for quality assessment design, in this experiment we are not interested in evolving behavioral responses, hence the sample median provides a robust estimate of distortion visibility.

5.2. Temporal Activity & Motion

To quantify the amount of motion in the video, we modified the technique in [63] to measure temporal activity. First, an absolute difference between adjacent frames is computed and the resulting difference sequence is analyzed over spatio-temporal blocks of size 4 pixel $\times$ 4 line $\times$ 0.2 seconds. The standard deviation of each such block gives a measure of local temporal activity, which is then thresholded using the perceptual thresholds of [63] in order to account for human perception. The mean of these thresholded values across the frame is a measure of the overall temporal activity of the frame. The temporal activity of the video is then the mean across luminance frames.

Although proposed temporal activity features model motion information by computing statistics of spatio-temporal block based frame differences, we believe that a richer set of temporal features can be obtained by using models of motion perception. Psychophysical study by Stocker and Simoncelli [64] on human visual speed perception suggests that accuracy of visual speed perception is greatly reduced in the presence of large frame motion. In other words, distortion visibility gets reduced in frames with large ego motion. This statistical model has recently been used by Wang and Li in [39] in developing their video quality assessment algorithm. Also, Suchow and Alvarez [65] recently showed a visual illusion which indicates that motion silences awareness of visual change. Large coherent motions present in the frame tend to mask the distortions. Going forward, we aim to incorporate these models in our temporal features.

5.3. Algorithm

To provide a visual illustration of the temporal and spatial activities of the videos used in the study, Fig. 5 plots the videos in Fig. 1 as a function of their spatial and temporal activities. As the figure indicates, the videos span a wide range of spatial and temporal activity.

To ascertain how well each feature individually correlate with VL threshold, we compute Spearman rank ordered correlation coefficient (SROCC) between actual VL-bit rate as given in Table 2 and our spatial & temporal activity features. Spatial activity has a correlation of 0.36 with VL bit-rate whereas temporal activity has a correlation of 0.67, which emphasizes the fact that temporal activity is a more salient cue in videos. However, we would like the reader to note that correlation numbers were computed using only 5 videos

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.

for which VL bit-rates are available. This might not exactly reflect how features correlate with VL bit-rate given the small number of videos used for computation. We hope that with our future studies, a higher quantity of data will allow for a better understanding of such correlations.

Having extracted temporal and spatial activity measures we map these measures onto a visually lossless bit-rate (obtained from the table above). This mapping is performed using a support vector regression model in which the 2-tuple - $\mathbf{X}$ = (spatial activity, temporal activity) is used as the input and the VL bit-rate is used as the target output. A ν-support vector machine (SVM) [66, 67] with a radial basis function (RBF) kernel was used. SVM has also been used successfully in quality assessment domain in the past [49]. We used a leave-out-one validation to test the performance of the approach. We trained the SVM on four of the five videos with valid VL bit-rates and used this trained SVM to predict the VL threshold on the remaining videos. The learning and prediction framework is shown in Fig. 6.

This was repeated five-times, in order to predict the VL thresholds for each of the five videos. The results are listed in Table 2, where the ground-truth VL bit-rates and the predicted VL bit-rates are tabulated.

As Table 2 demonstrates, the 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 not possible.

Thus, our implementation of the compressibility index for H.264 compression consists of a support vector machine that regresses measures of spatial and temporal activity onto a bit-rate corresponding to the visually lossless threshold for that video.

The H.264 Visually Lossless Compressibility Index (HVLCI) is the first algorithm that predicts visually lossless threshold bit rates for H.264 compressed videos.

6. CONCLUSION AND FUTURE WORK

The results indicate that HVLCI may be useful as a tool for applications such as automatic perceptual rate control. Given that mobile traffic is expected to double every year till 2015 and percentage of video traffic is going to increase up to two-third and current wireless network infrastructure including base stations, access points, capacity cannot keep up with such growth; maximizing the overall quality of experience through perceptual rate control seems one of the promising solutions. Another related scenario where the application can prove helpful is when multiple users are sharing a congested wireless source subjected to slow wireless capacity and perceived video rate variability. Joint source rate adaptation can be a key to maximize the overall user quality of experience as users watching different contents can be reallocated bandwidth based on VL bit-rates of corresponding videos. Another application where we expect our index to contribute is paving a way for perceptually optimized video compression which would not only exploit spatial and temporal redundancy but also model the distortion visibility as perceived by a human.

The proposed index has been made available online [1] for its utilization in practical scenarios and promote further research in this direction.

Future work could involve the computation of a psychometric function which would represent the *degree* of visual loss at every bit-rate. This could give users the freedom to select an operating point on rate distortion (R-D) curve depending on available bandwidth. Naturally, as our understanding of spatio-temporal distortion perception improves, we will be able to augment and improve this first version of the H.264 Visually Lossless Compressibility Index. We are also conducting a new visually lossless study with videos of frame size 1280×720 given the prevalence of HD content in today's world.

REFERENCES

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

[2] Hoffman, D. D. (2005) Visual Illusions and Perception. *McGraw-Hill Yearbook of Science and Technology.*

[3] Hunt, R. W. G. (2004) *The Reproduction of Colour.* Wiley.

[4] Fairchild, M. D. (2005) *Color appearance models.* Wiley.

[5] Clark, J. J. and Yuille, A. L. (1990) *Data Fusion for Sensory Information Processing Systems.* Springer.

[6] Perkins, M. G. (1992) Data compression of stereopairs. *IEEE Transactions on Communications*, **40**, 684–696.

[7] Jayant, N., Johnston, J., and Safranek, R. (1993) Signal compression based on models of human perception. *Proceedings of the IEEE*, **81**, 1385–1422.

[8] Karam, L. J. (2009) "Lossless Image Compression". *The Essential Guide to Image Processing, Al Bovik Ed.*, pp. 385–417. Elsevier Academic Press.

[9] (2003) *Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IECAVC.*

[10] Gelenbe, E., Sungur, M., Cramer, C., and Gelenbe, P. (1996) Traffic and video quality with adaptive neural compression. *Multimedia Systems*, **4**, 357–369.

[11] Cramer, C., Gelenbe, E., and Bakircloglu, H. (1996) Low bit-rate video compression with neural networks

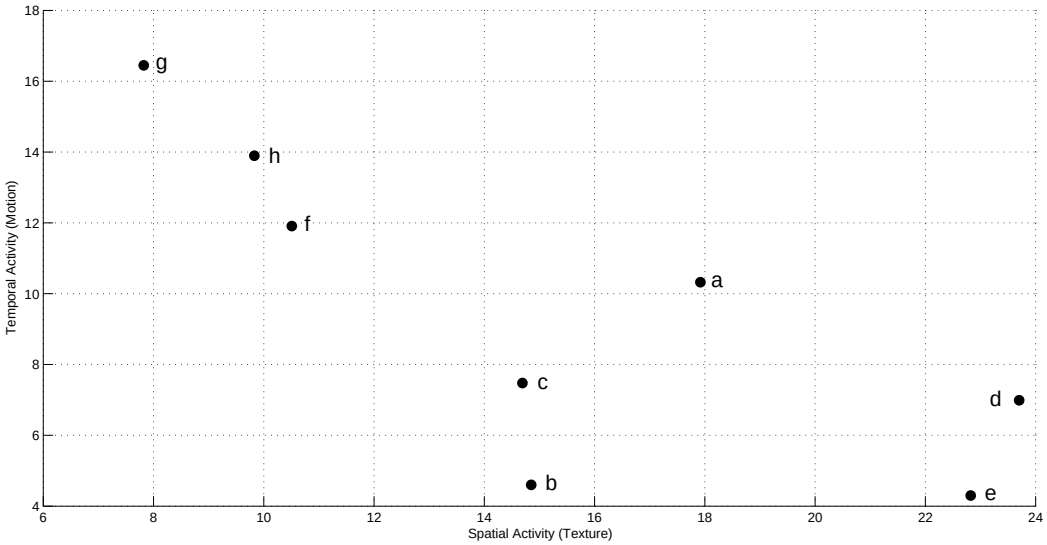

FIGURE 5. Plot of spatial (x-axis) activity vs. temporal (y-axis) activity for videos used in this study. Points a-h correspond to videos (a) - (h) from Fig. 1.

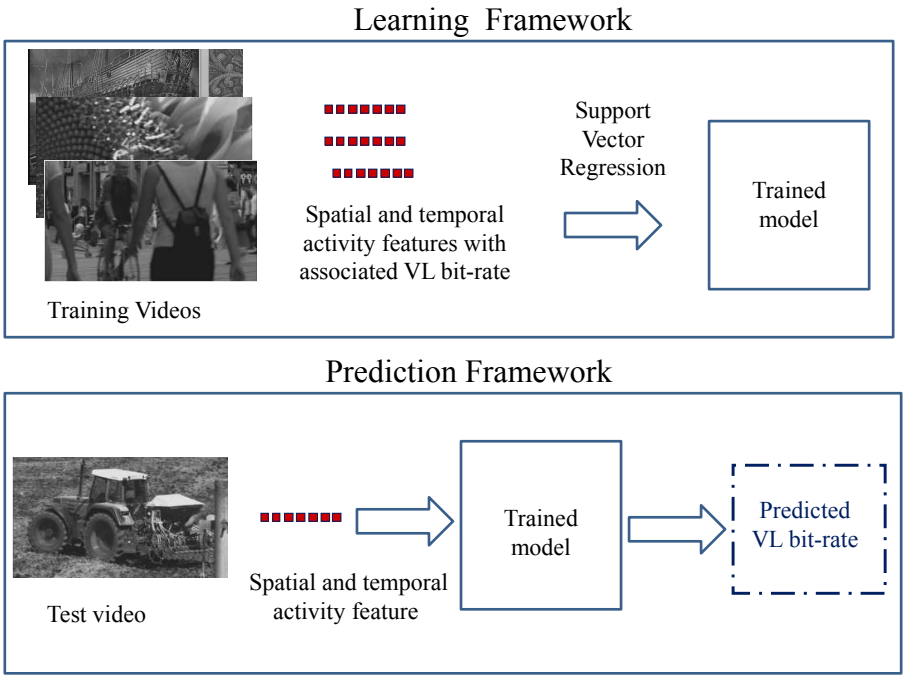

FIGURE 6. Shows the learning framework where mapping of spatial and temporal activity features to VL bit-rate is learned. Trained SVM model is then used to predict the VL bit-rate of the new test video.

and temporal subsampling. *Proceedings of the IEEE*, **84**, 1529–1543.

[12] Feng, Y. and Gelenbe, E. (1998) Adaptive object tracking and video compression. *Networking and Information Systems Journal*, **1**, 371–400.

[13] Cramer, C. and Gelenbe, E. (2000) Video quality and traffic QoS in learning-based subsampled and receiver-interpolated video sequences. *IEEE Journal on Selected Areas in Communications*, **18**, 150–167.

[14] Przelaskowski, A. and Jozwiak, R. (2008) Compression of bronchoscopy video: Coding usefulness and efficiency assessment, . pp. 208–216. Springer.

[15] Slone, R. M., Muka, E., and Pilgram, T. K. (2003) Irreversible JPEG compression of digital chest radiographs for primary interpretation: Assessment of visually lossless threshold. *Radiology*, **228**, 425.

[16] Slone, R. M., Foos, D. H., Whiting, B. R., Muka, E., Rubin, D. A., Pilgram, T. K., Kohm, K. S., Young, S. S., Ho, P., and Hendrickson, D. D. (2000) Assessment of visually lossless irreversible image compression: Comparison of three methods by using an image-comparison workstation1. *Radiology*, **215**, 543–553.

[17] Kocsis, O., Costaridou, L., Varaki, L., Likaki, E., Kalogeropoulou, C., Skiadopoulos, S., and Panayiotakis, G. (2003) Visually lossless threshold determination for microcalcification detection in wavelet compressed mammograms. *European Radiology*, **13**, 2390–2396.

[18] Lee, K. H., Kim, Y. H., Kim, B. H., Kim, K. J., Kim, T. J., Kim, H. J., and Hahn, S. (2007) Irreversible JPEG 2000 compression of abdominal CT for primary interpretation: assessment of visually lossless threshold. *European Radiology*, **17**, 1529–1534.

[19] Busemeyer, J. R. and Townsend, J. T. (1993) Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, **100**, 432.

[20] Watson, A. B. and Kreslake, L. (2001) Measurement of visual impairment scales for digital video. *Proceedings of the SPIE*, pp. 79–89.

[21] Yang, X. K., Ling, W. S., Lu, Z. K., Ong, E. P., and Yao, S. S. (2005) Just noticeable distortion model and its applications in video coding. *Signal processing: Image communication*, **20**, 662–680.

[22] Chin, Y. J. and Berger, T. (1999) A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements. *IEEE Transactions on Circuits and Systems for Video Technology*, **9**, 438–450.

[23] Chou, C. H. and Chen, C. W. (1996) A perceptually optimized 3-D subband codec for video communication over wireless channels. *IEEE Transactions on Circuits and Systems for Video Technology*, **6**, 143–156.

[24] Chou, C. H. and Li, Y. C. (1995) A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on Circuits and Systems for Video Technology*, **5**, 467–476.

[25] Ahumada, A. J. and Peterson, H. A. (1992) Luminance-model-based DCT quantization for color image compression. *SPIE Conf. on Human Vision, Visual Processing, and Digital Display III*, **1666**, 365–374.

[26] Watson, A. B. (1993) DCT quantization matrices visually optimized for individual images. *Proceedings of the SPIE*, pp. 202–216.

[27] Watson, A. B., Yang, G. Y., Solomon, J. A., and Villasenor, J. (1997) Visibility of wavelet quantization noise. *IEEE Transactions on Image Processing*, **6**, 1164–1175.

[28] Bradley, A. P. (1999) A wavelet visible difference predictor. *IEEE Transactions on Image Processing*, **8**, 717–730.

[29] Tong, H. H. Y. and Venetsanopoulos, A. N. (1998) A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking. *International Conference on Image Processing*, pp. 428–432. IEEE.

[30] Tran, T. D. and Safranek, R. (1996) A locally adaptive perceptual masking threshold model for image coding. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1882–1885. IEEE.

[31] Hontsch, I. and Karam, L. J. (2000) Locally adaptive perceptual image coding. *IEEE Transactions on Image Processing*, **9**, 1472–1483.

[32] Hontsch, I. and Karam, L. J. (2002) Adaptive image coding with perceptual distortion control. *IEEE Transactions on Image Processing*, **11**, 213–222.

[33] Chen, Z. and Guillemot, C. (2010) Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model. *IEEE Transactions on Circuits and Systems for Video Technology*, **20**, 806–819.

[34] Naccari, M. and Pereira, F. (2011) Advanced H.264/AVC-based perceptual video coding: Architecture, tools, and assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, **21**, 766–782.

[35] Richardson, I. E. G. (2003) *H.264 and MPEG-4 video compression*. Wiley Online Library.

[36] Technical university of munich. [online]. ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/.

[37] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K. (2010) Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, **19**, 1427–1441.

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

[39] Wang, Z. and Li, Q. (2007) Video quality assessment using a statistical model of human visual speed perception. *JOSA*, **24**, B61–B69.

[40] Ninassi, A., Le Meur, O., Le Callet, P., and Barba, D. (2009) Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, **3**, 253–265.

[41] Watson, A. B., Hu, J., and McGowan III, J. F. (2001) Digital video quality metric based on human vision. *Journal of Electronic imaging*, **10**, 20.

[42] Barkowsky, M., Bialkowski, J., Eskofier, B., Bitto, R., and Kaup, A. (2009) Temporal trajectory aware video quality measure. *IEEE Journal of Selected Topics in Signal Processing*, **3**, 266–279.

[43] Hekstra, A. P., Beerends, J. G., Ledermann, D., De Caluwe, F. E., Kohler, S., Koenen, R. H., Rihs, S., Ehrsam, M., and Schlauss, D. (2002) PVQM–A perceptual video quality measure. *Signal processing: Image communication*, **17**, 781–798.

[44] Sheikh, H. R. and Bovik, A. C. (2006) Image information and visual quality. *IEEE Transactions on Image Processing*, **15**, 430–444.

[45] Sheikh, H. R., Bovik, A. C., and De Veciana, G. (2005) An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, **14**, 2117–2128.

[46] Sheikh, H. R., Bovik, A. C., and Cormack, L. K. (2005) No-reference quality assessment using natural scene

statistics: JPEG2000. *IEEE Transactions on Image Processing*, **14**, 1918–1927.

[47] Soundararajan, R. and Bovik, A. C. (2012) RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, **21**, 517–526.

[48] Moorthy, A. K. and Bovik, A. C. (2010) A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, **17**, 513–516.

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

[50] Saad, M., Bovik, A. C., and Charrier, C. (2011) DCT statistics model-based blind image quality assessment, . 11-14 Sept, pp. 3093–3096.

[51] Wang, Z. and Bovik, A. C. (2011) Reduced and no-reference image quality assessment: The natural scene statistic model approach. *IEEE Signal Processing Magazine*, **28**.

[52] Belmudez, B., Moeller, S., Lewcio, B., Raake, A., and Mehmood, A. (2009) Audio and video channel impact on perceived audio-visual quality in different interactive contexts. *IEEE International Workshop on Multimedia Signal Processing, 5-7 Oct.*

[53] Kühnel, C., Westermann, T., Weiss, B., and Möller, S. (2010) Evaluating multimodal systems: a comparison of established questionnaires and interaction parameters. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, New York, NY, USA NordiCHI '10, pp. 286–294. ACM.

[54] Mehmood, A., Wundsam, A., Uhlig, S., Levin, D., Sarrar, N., and Feldmann, A. (2011) QoE-lab: Towards evaluating quality of experience for future internet conditions. Tridentcom.

[55] (2007). H.264/AVC software coordination. [online]. available. http://iphome.hhi.de/suehring/tml/.

[56] Union, I. T. (2003). BT-500-11: Method for the subjective assessment of the quality of television pictures.

[57] Perry, J. S. (2008). XGL toolbox[online]. http://fi.cvis.psy.utexas.edu/software.shtml.

[58] Sheskin, D. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.

[59] Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J. (1992) Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, **38**, 587–607.

[60] Lyu, S. and Simoncelli, E. P. (2007) Statistical modeling of images with fields of gaussian scale mixtures. *Advances in Neural Information Processing Systems*, **19**, 945.

[61] Portilla, J. and Simoncelli, E. P. (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, **40**, 49–70.

[62] Wandell, B. A. (1995) *Foundations of Vision*. Sinauer Associates.

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

[64] Stocker, A. and Simoncelli, E. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, **9**, 578–585.

[65] Suchow, J. and Alvarez, G. (2011) Motion silences awareness of visual change. *Current Biology* , **?**

[66] Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000) New support vector algorithms. *Neural Computation*, **12**, 1207–1245.

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