Detectability and Annoyance Value of MPEG-2 Artifacts Inserted into Uncompressed Video Sequences

Michael S. Moore*, John Foley, and Sanjit K. Mitra

Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106 USA

ABSTRACT

Many approaches have been proposed recently for estimating the perceived fidelity of digitally compressed and reconstructed video clips. Several metrics rely on the accurate estimation of detection thresholds, basically assuming that perceived quality depends significantly on the error threshold. To test that assumption, we designed an experiment to measure the actual detection threshold of some MPEG-2 artifacts in typical video sequences. In each of the test clips, we briefly replaced a region with a test stimulus created using an MPEG-2 encoder. The location, time, and strength of the artifact were varied between test videos. At the end of each clip, each subject provided a detection response and, for the detected artifacts, location and annoyance information. From the data, we determined the detection threshold for each artifact. Using the thresholds, we computed a perceptually weighted error metric. In this paper, we describe the experiment in more detail, summarize the experimental results including the threshold calculations, and compare the weighted error measure to the output of a commercial fidelity metric. Finally, we discuss our conclusions on the validity of predicting quality from threshold measurements.

Keywords: video, quality, MPEG-2, metrics, annoyance

1. INTRODUCTION

The digital transmission of images and video offers many advantages over the existing analog methods. The resulting bit streams can be enhanced, transmitted with little or no error, translated across different standards, and displayed on a variety of devices. This flexibility is both a blessing and a curse. Digital data can be transformed in many complex ways. Different methods should be used for different applications and in different, and dynamically changing, situations. To choose an approach, a method for calculating the resulting quality is required.

In many applications, the transformed video is destined for human consumption and humans will decide if the operation was successful. Therefore, human perception should be taken into account when picking an algorithm. Human perception comes into play when trying to establish the degree to which a video can be compressed,[1] figuring out how much data can be hidden in an image,[2] deciding whether visual enhancements have provided an actual benefit,[3] and choosing a set of image features to use for segmentation and indexing.[4] In this paper, we consider video that has been compressed using a lossy compression algorithm. Because data is deliberately thrown away, compression frequently results in reduced quality.

Research in the Image Processing Laboratory at the University of California, Santa Barbara has in the past been mostly concerned with the development of algorithms for image and video processing. For example, work has been done on video compression standards, wavelet compression schemes, image noise removal filters, and automated video segmentation schemes. However, measurement of the performance of various algorithms has been limited to a few objective measures such as the mean absolute error (MAE), mean square error (MSE), and peak signal-to-noise ratio (PSNR) supplemented by limited subjective evaluation. Although this approach is fairly standard in published literature, it suffers from a frustrating problem. Sometimes, images with better objective scores look subjectively worse than images with worse objective scores. Algorithms are designed to minimize or maximize an objective measure. As the algorithms improve and converge faster to a particular goal, it becomes more important that the objective measure actually represent perceived quality.

We had multiple goals for this experiment. Perhaps most importantly, we hoped to get a reasonably good set of annoyance value measurements for a set of test sequences of varying quality. Even if no more specific goals were

*Correspondence: Email: msmoore@iplab.ece.ucsb..edu, WWW: http://iplserv.ece.ucsb.edu/users/msmoore

met, the annoyance value data would allow us to confirm some basic findings (for example, that MSE is not a good measure of subjective quality), compare existing metrics, search for the strengths and weaknesses of the various approaches, and try out some of our own ideas.

As a second major goal, we wanted to test one of the underlying assumptions used in some fidelity metrics, that perceived quality depends on the detection threshold for an error pattern. The detection threshold depends on the context in which the error occurs. In many cases, efforts are being made to fully flesh out a model for detecting patterns in context as a function of several variables.[1,5] The hope seems to be that an accurate realization of the detection function will lead to better correlation with subjective scores. We do not have a complete general model of error thresholds. Consequently, we tested this hypothesis by directly measuring error thresholds for compression artifacts. If the thresholds capture all of the context dependent variation in reported quality, then a function of the video error and the threshold should be able to accurately predict the annoyance values.

2. BACKGROUND

In a review of current literature, we found that quality metrics can generally be categorized into one of two types. Fidelity metrics measure the perceivable differences between two video sources.[1,6,7] Attribute measures attempt to determine the strength of visual features in images, such as the sharpness or color correctness of an image.[8–10]

The most common approach to quality measurement is the fidelity metric. Fidelity metrics compare a transformed image to an uncorrupted or reference image. The MSE or PSNR are simple fidelity metrics. However, as mentioned before, these metrics are inadequate in that they sometimes give bad objective scores to subjectively good images. One possible reason for this behavior is that they include imperceptible differences. Consequently, attempts to create improved fidelity metrics take into account the threshold of error detection at each point in the reference image. The differences between pixels in the transformed and reference images are weighted by some function of the threshold at each pixel location. For example, each difference may be divided by the threshold and any result less than one set to zero.[1] The output of this procedure is a map of the perceivable differences between the images. To get a single number for each image or frame in a video, the weighted differences are pooled.

The detection thresholds used by fidelity metrics vary. In most cases, the thresholds are estimated using psychophysical models for contrast sensitivity. Based on evidence from psychophysical experiments, contrast sensitivity is thought to vary with many image characteristics, such as spatial frequency, temporal frequency, orientation, mean lumninance, color, and size.[11] In many cases, models have been proposed to describe the dependence of contrast sensitivity on physical variables, such as spatial frequency. However, the models were generally developed to fit experiments with only one or two of the relevant independent variables. The various fidelity metrics vary in the subset of variables they use, the models for each variable, and the methods for combining the effects across variables to create a single combined contrast sensitivity.

Although some testing has been done to experimentally determine the detection threshold with respect to multiple independent variables (such as spatial and temporal frequency), for the most part, the effort to develop the multivariable threshold functions has outpaced experimental research. To compensate, fidelity metric parameters may be tuned based on subjective quality experiments. In other words, separate lower level models of pattern detection threshold are calibrated using psychophysical experiment results. But the overall, combined model is calibrated to match the scores measured in subjective quality experiments.

In general, the perceptually-based fidelity metrics correlate with subjective quality judgements better than the pure error fidelity metrics, like the MSE or PSNR. Fidelity metrics require a reference input. A fidelity metric provides a measure of image quality only if the reference input can be assumed to be perfect. In addition, the reference and transformed inputs must be perfectly registered, so that the same time and location are being compared. Otherwise the accumulation of small differences from misregistration will result in erroneously poor quality scores. These requirements are not very restrictive in laboratory research environments, but may be impractical in other applications.

3. METHODS

3.1. Test Sequences

The normal approach to subjective quality testing is to degrade a video by a variable amount and ask the test subjects for a quality rating.[12,13] There are variations in the questions asked of the viewers (single stimulus rating,

dual stimulus orderings, etc.). However, the process is usually applied to the entire video. For MPEG-2 video compression, the errors introduced into the reconstructed video are rarely spread evenly. The amount of error and even its appearance can change from region to region. Some recent experiments have started to look at limited regions of an compressed signal, either spatially in images[14] or temporally in video.[12,15]

We designed an experiment that would experimentally measure the detection threshold of brief, spatially limited MPEG-2 artifacts in videos. We degraded an isolated region of the video clip for short time interval. The rest of the video clip was left in its original state. The test subjects were asked to search each video clip for defective regions and to indicate where the artifacts were seen. The regions and time intervals where defects appeared were varied to prevent the test subjects from learning the locations. Also, we suspected that different spatial regions would have different thresholds, for reasons related to both MPEG-2 (the amount of corruption varies with region characteristics) and human factors (for example, attention effects).

To generate the test video clips, we used five original video clips of assumed high quality. The five video clips were chosen because they all had the same length (five seconds) and contained scenes that we thought were typical of normal television. Specifically, there are videos of a bus moving through traffic (Bus), cheerleaders performing on a playing field (Cheerleaders), houses and flowers flowing by as seen out of a car window (Flower), a football player running with the ball (Football), and part of a hockey game (Hockey).

Each original test sequence was compressed using an MPEG-2 codec implementation and then reconstructed. We used the MPEG-2 codec that is available for public download from the MPEG Software Simulation Group, version 1.2. The source code was modified slightly to read and write videos in the Tektronix format required by our display equipment. The codec comes with a sample parameter file for compression of NTSC video. This parameter file was used as a template for the parameter files for the test sequences. The only parameters that were changed were those relating to the input file format (format type, size, number of frames, etc.) and the bit-rate goal in bits per second. Other codecs are commercially available. These codecs generally perform better, in terms of quality at a given bit-rate, than the public codec. However, the source code is available for the public codec and that let us rapidly adapt the program to our needs. In addition, our intent was to create highly visible errors and optimizing the codec did not seem necessary. However, if better codecs produce qualitatively different errors, then we may not be able to generalize our conclusions to those codecs. We do not believe this to be the case.

For the experiment, the MPEG-2 compressed video was created using a bit-rate goal of one megabit per second (Mbps). This bit-rate produces very poor video quality, unsuitable for most applications. The low bit-rate removed most high frequency components from the reconstructed video, which resulted in a blurred appearance. By mixing the original and reconstructed videos, these components were gradually added back. So the sequences ranged from normal (high quality) to high-frequency deficient (low quality).

Different artifact regions were chosen for each original sequence. Each video frame was broken into three regions of roughly equal size. In two videos, the originals had good natural boundaries that divided the frame into approximate thirds. In the Bus video, the regions were above the bus, the bus, and below the bus. In the Flower video, the regions were the sky, the houses, and the garden. The Flower regions have the largest size disparity, as the garden takes up almost half of the frame and the other two regions split the remainder. For the other videos, no natural boundaries were obvious, so the frames were simply split into exact thirds. The Cheerleader frames were split into horizontal bands, and the Football and Hockey frames were split into vertical bands.

For all of the test sequences, the time interval selected for the artifact was one second. However, the one second interval could occur in either the second, third, or fourth seconds of the five second video clip. Any changes to the first and last seconds were avoided.

The combination of three regions with three time intervals would result in nine different sequences for each original. However, we also needed to vary the strength of the artifacts through several different levels, from below threshold to highly visible. Using all possible combinations of strength, time interval, and region would have resulted in a very long experiment. Therefore, only three combinations of time interval and region were selected for the test sequences derived from a single original video. However, the combinations were changed between sequences derived from different originals. For example, the artifact in the center region of the Football-based sequences always appeared in the fourth second. However, the artifact in the center region of the Hockey-based sequences always appeared in the third second. The complete set of test sequences was shown to each test subject in random order. We think that it is fairly unlikely that the subjects realized specific regions were associated with specific time intervals.

For each artifact region, six values of the weighting factor, a, were chosen. Ninety test sequences were created using the five original and reconstructed videos. In this experiment, a never exceeded a value of one, so the worst case artifact was identical to the MPEG-2 video in the selected region. In fact, the same values of a were used in all of the test sequences: 0.25, 0.32, 0.42, 0.57, 0.75, and 1.0.

The following process was used to combine the original and reconstructed videos into a test video clip:

1. The original video, $I(x, y, t)$, was copied to create the basis for the test video, $T(x, y, t)$. Note that at each spatio-temporal location (x, y, t), the functions $I(x, y, t)$ and $T(x, y, t)$ actually contain the luminance and two color differences (Y', C'_R, and C'_B). The remaining processing was only done in the time interval, $t_1 \leq t \leq t_2$, and region, $\forall x, y \in \Re$, for a particular test sequence.

2. The frames were stored in a gamma corrected format (ITU BT.601 digital component video). Therefore, the pixel value for a particular frame and location was converted to an approximation of the linear light domain, using an assumed gamma value of 2.5. The conversion was applied to both original frame values, $I(x, y, t)$, and MPEG-2 image values, $M(x, y, t)$. The same transformation was used for all three color components.

$$\begin{array}{rcl} L_I(x, y, t) & = & I(x, y, t)^{2.5} \\ L_M(x, y, t) & = & M(x, y, t)^{2.5} \end{array} \tag{1}$$

3. The linear light values from the original frame and the MPEG-2 compressed and reconstructed frames were linearly combined using the weighting constant a.

$$L_C(x, y, t) = (1 - a)L_I(x, y, t) + aL_M(x, y, t) \tag{2}$$

4. The result was returned to the gamma-corrected domain using the inverse transformation.

$$C(x, y, t) = L_C(x, y, t)^{0.4} \tag{3}$$

5. The result was clipped to the valid range of 0 to 255 and truncated to form an 8-bit integer.

$$T(x, y, t) = \begin{cases} 0, & \text{if } C(x, y, t) < 0, \\ \text{floor}(C(x, y, t)), & \text{if } 0 \leq C(x, y, t) \leq 255, \\ 255, & \text{if } 255 < C(x, y, t). \end{cases} \tag{4}$$

In summary, a set of test sequences was generated. The sequences consisted of mostly uncorrupted video with one-third of a frame replaced with an artifact for a one-second interval. The artifact was generated by linearly combining the original video with a video reconstructed after MPEG-2 compression. At each location, several sequences were created with different artifact strengths. The full set consisted of ninety-five videos (five original sequences times three spatio-temporal regions times six artifact strengths plus the five original sequences).

3.2. Apparatus

The test sequences were stored on the hard disk of an NEC Express Server. Each sequence requires just over 100 megabytes of storage and the entire test set for one experiment needs almost 10 gigabytes. The sequences were stored in the Tektronix rawTekTV format, which is essentially the ITU BT.601 standard 4:2:2, Y:C_R:C_B format, with proprietary file and frame headers. The server used a single 18-gigabyte SCSI hard drive to store the test sequences. With this configuration, each test sequence could be loaded for display in six to eight seconds.

Each video was displayed using a subset of the PC cards normally provided with the Tektronix PQA-200 picture quality analyzer. A generator card was used to locally store the video and stream it out in a serial digital (SDI) component format. The length of the test sequences was limited to five seconds by the amount of memory on the generator card. The output of the generator card was sent to a decoder card that converted the SDI signal into an analog NTSC S-video output. The analog output was then displayed on a Sony PVM-1343 monitor.

In addition to storing the video sequences, the server was also used to run the experiment and collect data. A TCL script was written to run under the software provided with the Tektronix PQA-200. The script recorded each subject's name, displayed the sample video clips, ran the practice experiment, and ran the actual experiment. During

the experiment, the videos were displayed in the order specified by one of ten prepared lists. Each list contained a different random ordering of the test sequence set. Because of limitations in the display software, each sequence was displayed twice in immediate succession. After the second repetition finished, the script asked a series of questions and recorded the subject's responses in a subject-specific data file.

The experiment was run with one test subject at a time. Each subject was seated in front of the computer keyboard at the end of a table. Directly ahead of the subject was the Sony video monitor, located at or slightly below eye height for most subjects. The subjects were initially positioned at a distance of four screen heights (80 cm) from the video monitor. However, the subjects were not prevented from moving during the experiment, and sometimes ended up leaning forward to the table edge (approximately 65 cm). The computer monitor was located to the right of the keyboard and closer to the test subject.

The test sequences appeared on the video monitor directly in front of the viewer. After each sequence, questions appeared on the computer monitor to the right. This setup required the subject to switch attention from the video monitor to the computer display. However, because the same questions were asked after each sequence, the subjects soon memorized the questions and rarely looked at the computer monitor. When an impairment was detected, the subject had six to eight seconds to answer the annoyance and location questions while the computer loaded the next video clip. This interval was long enough for the test subject to answer the questions and return attention to the video monitor.

3.3. Procedure

Our test subjects were drawn from a pool of students in the introductory psychology class at UCSB. The students are thought to be relatively naive of MPEG-2 compression artifacts and the associated terminology. They were asked to wear any vision correction devices (glasses or contacts) that they would normally wear to watch television. The students were tested one at a time.

The course of each experimental session went through four stages. In the first stage, the subject was verbally given instructions. In the second stage, ten sample test sequences (five original and five created with $a = 1$) were shown to the subject to establish the range of quality. Afterwards, five practice trials were performed. Finally, the actual experiment was performed with the complete set of test sequences.

The verbal instructions were written to emphasize several points. First, the subjects were told that we were not concerned with the content of the videos but that we wanted them to report any defects or impairments that they saw in the imagery. Second, we wanted the subjects to search for defects. They were asked to initially look at the center of the screen, but to move their eyes around and look for defects or impairments.

We instructed the test subjects to give us annoyance values for the worst defect or impairment that they saw in each test sequence. Although we only introduced one defect ourselves, we could not rule out the possibility of either pre-existing defects or that the inserted defect would look like multiple defects. We asked the subjects to give us a value proportional to the annoyance caused by the defect or impairment. Once again, they were asked not to judge the entire clip, but rather to judge just the defect.

To set the scale for the annoyance values, we used the ten sample video clips (stage two). The series consisted of five pairs of clips. Each pair consisted of an original clip followed by one of the three test sequences for that clip with the maximum amount of error added. However, the subjects were only told that the clips would alternate between low and high annoyance. They were not told that one set of clips was original or defect-free, as there could in fact have been defects in the original clips.

The subjects were asked to set a value of 100 for the worst defect they saw in the sample clips. During the experiment, any clip just as bad should have been scored at 100, any half as bad scored at 50, and any twice as bad scored at 200. Although we tried to include the worst five test sequences in the sample set, we acknowledged the fact that the subjects might find some of the other tests clips to be worse and specifically instructed them to go above 100 in that case.

After the sample clips, the test subjects were told how to enter their responses during the test. Essentially, the Y and N keys were used to answer yes/no questions, and the numeric keypad was used to enter annoyance values and locations. Location data was entered by dividing the screen into nine equal size regions, corresponding to the arrangement of keys on the numeric keypad. For example, a defect that was seen down the right side of the screen would be entered as locations 369 while a defect only seen in the center would be entered as location 5.

The next step was a short practice run of the experiment. The practice run included five trials. The subjects were shown the test sequences and asked questions, just like in the actual experiment. However, data were not recorded for the practice trials. The main purpose of the practice trials was to give the subject confidence that he or she understood the instructions. The practice also served to give the subjects an idea of the pace of the experiment. The first few responses were typically slow, but got much faster by the end of the practice trials.

The subjects were asked two questions after the main part of the experiment was over. First, they were asked to describe the defects that they saw. Second, they were asked if there was more than one type of defect. Generally, more than one type of defect was seen. We made an effort to not provide descriptive words to the subjects, because we wanted to know how naive viewers would describe MPEG-2 artifacts. However, describing the defects was a difficult task, and we frequently had to move beyond the two initial questions to flesh out a subject's response.

4. DATA ANALYSIS

4.1. Initial Processing

The data from each subject were stored in individual data files. After the experiment was completed, the individual data files were combined into a series of spreadsheets. The data were sorted by subject and test sequence. Therefore, we ended up with spreadsheets containing matrices of the detection responses, the annoyance values, and the detected locations. For every artifact that was not detected, we assumed an annoyance value of zero. We also summarized the answers to the qualitative questions.

Although the test subjects were requested to use a range for the annoyance values such that the worst of the five sample videos would be assigned a value of 100, only sixteen of the thirty-two test subjects actually gave the sample videos a score of 100 when the sequences reappeared in the actual experiment. To match the values for these sequences across subjects, we rescaled the annoyance values for the other sixteen subjects so that their highest annoyance value for a sample video equalled 100. These rescaled annoyance values are used in the rest of this paper. Note, however, that the rescaling did not limit the highest annoyance value to 100. Seven subjects gave us annoyance values for non-sample video clips that were higher than any of the sample videos. After rescaling, these annoyance values all exceeded 100.

We also needed an objective measure of the differences between the corrupted test sequences and the original test sequences. We chose to calculate the squared error, or error energy, between the sequences, summed spatially over each frame and temporally across frames. The error energy was calculated using the following formula:

$$EE_i = \sum_t \sum_x \sum_y \left[\left(\frac{I(x,y,t)}{255} \right)^{2.5} - \left(\frac{T_i(x,y,t)}{255} \right)^{2.5} \right]^2 . \tag{5}$$

As can be seen, the pixel values were weighted to a range of [0, 1] and then transformed to the linear light domain. The difference between the linear light values was squared and summed over all spatial locations and frames. This resulted in the error energy for test sequence i. Although it is not apparent in the formula, the same summation was applied to the luminance and the color difference values (Y, C_R, and C_B) and summed to create the overall error energy. Because the number of samples for each color difference is half of the number of luminance samples (4:2:2 format), each color component error is effectively weighted by one-half compared to the luminance error.

4.2. Detection Consistency

The experiments have provided us with several kinds of data. The detection and annoyance value data is quantitative. Using this data, we will compare different ways of predicting annoyance. However, before we put too much stock in the answers, we need to know how good our data set is.

The detection data is not perfectly consistent in the way that it depends on the strength of the artifacts. We would expect that, outside of a narrow range around the threshold, subjects would detect very few artifacts below threshold and nearly all of the artifacts above threshold.

Although this is true for many subjects and artifacts, it is not true for all subjects and all artifacts. There are many possible explanations for the few deviations from the expected results. For example, perception is thought to be a noisy process resulting a probability distribution for detection at each artifact strength. In this case, occasional

Table 1. Detection consistency.

Missed Artifacts		Detections in Originals	
Missed after	Number	Sequence	Number
1 was seen	80	Bus	0
2 were seen	13	Cheerleader	0
3 were seen	5	Flower	4
4 were seen	2	Football	1
> 4 were seen	0	Hockey	4

detections at low strength or non-detections at high strengths are expected and can be attributed to the nature of human perception.

However, there are other possible explanations for deviations from the expected results. For example, a motor error might result in an incorrect detection entry. A lapse in attention, distractions, or unexpected complication in entering data could also cause a subject to miss an artifact. Some of the detection inconsistency may be due to errors such as these. Our methods do not allow us to directly estimate the sources of detection inconsistency.

However, there are a few checks we can do. We can compare a subject's performance on a test sequence with other similar sequences. The likelihood of a non-detection should decrease rapidly as the number of detections at lower strengths increases. Also, defects were recorded in some test sequences did not contain any introduced artifacts. By examining the locations of these defects, we might be able to determine whether or not the unexpected detections were errors.

For the first check, we counted the number of times the test subjects did not see an artifact when they had seen the same artifact at one or more weaker strengths. The data was broken down by the number of weaker strengths that a subject saw the artifact. The results are shown in Table 1. The largest number of missed artifacts came after a single detection of an artifact. Most of these missed artifacts are probably due to the weakness of the artifacts near threshold and perceptual noise. However, as the number of weaker artifacts that were detected increases, the likelihood that the subject should have seen the artifact increases and inconsistencies are more likely to be due to other factors. In the experiment, seven artifacts were missed when at least three weaker artifacts were seen. This number is fairly small when compared to the total number of observations of corrupted sequences (3,240 observations). Even if all of the missed artifacts after earlier detection are counted as experimental errors instead of attributed to detection noise, the error rate is only about 3%.

Table 1 also lists the number of artifacts that were detected in the original video clips. Not all of these reported defects may be inconsistencies. If the original videos contained defects, they would be passed along to any derived test sequences. However, it should be possible to locate the defects in the original by using the location information. In fact, only two original videos (Flower and Hockey) had a significant number of detected defects. In the Flower sequence, the extra defects were always seen in the garden region, but varied in location within this area. In the Hockey sequence, no concentration of errors was found. The data indicates that the garden region in the original Flower sequence may contain defects. The other sequences do not apparently contain defects, so the detections in these clips are probably due to perceptual noise or are real motor errors. However, even if we include the Flower sequence results, then the error rate is only 5%.

4.3. Mean Annoyance Value Calculation

For the annoyance data, our goal was to find a accurate mean annoyance value. To achieve this goal, we need to minimize the variability in the data. We also need a measure of the accuracy of the estimated mean annoyance value. Recommendation ITU-R BT.500 includes procedures for both screening subjects and calculating confidence intervals on a mean.[12]

Reducing the variability in the data usually require the exclusion of subjects. This is a fairly drastic step. The Rec. 500 procedure excludes subjects who recorded annoyance values significantly above and below the mean. The assumption is that such a subject contributes little to the estimated mean while adding variablity.

Two subjects from the first experiment were excluded from the mean-of-subject (MOS) calculation. One subject did not follow the experiment instructions and provided annoyance values over the range of 0 to 1.6 instead of the

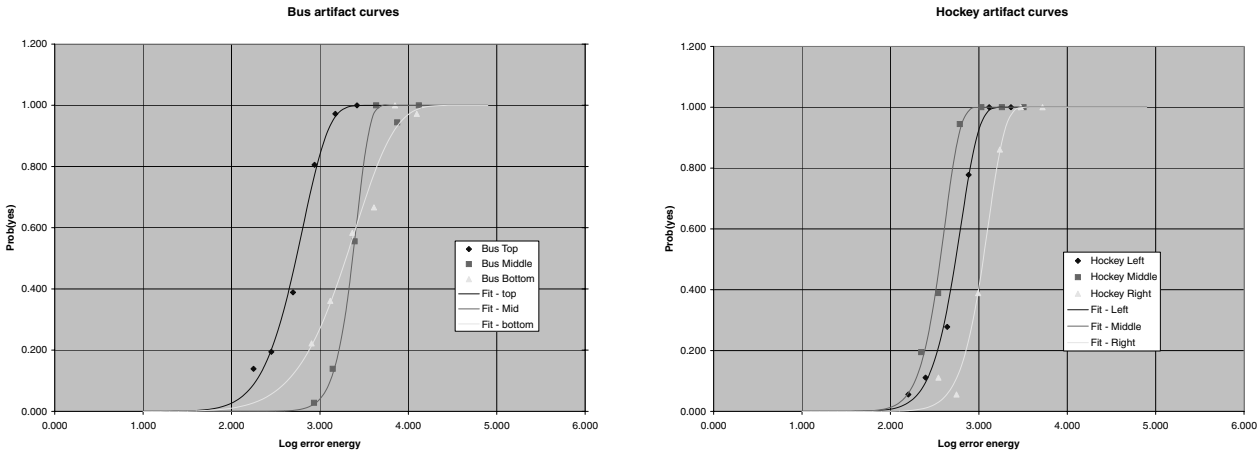


Figure 1. Probability vs. log error energy for the (a) Bus and (b) Hockey sequences. The points are our data values, the lines represent the curve fits.

full range (0 to 100+). We attempted to use the scaling technique, but the final results were internally inconsistent and did not correlate well with the other subjective scores. A second subject failed the screening test for variability recommended in ITU Rec. 500, and was removed from the data set.

Both of the procedures included in Rec. 500 assume that the annoyance values are normally distributed. To check this assumption, the β_2 test is recommended. If the β_2 values are in the range of [2, 4], we can approximate the distribution with a normal distribution. In fact, the β_2 values for 61% of our test sequences were in the range of [2, 4], and 75% were in the range [1.5, 4.5]. Most of the non-normal values were in the near-threshold test sequences. The closer the sequences were to threshold, the fewer annoyance values we received, and the less likely the distribution is going to be normal. For the screening procedure, we used the non-normal approximation for the non-normal test sequences. However, all of the error bars that we calculated are based on the normal assumption. For the higher annoyance values, this assumption is mostly valid.

5. RESULTS

5.1. Threshold Estimation

One of our goals for this experiment was to measure the error energy detection threshold for each of our artifacts. To do that we needed a probability of detection for each artifact as a function of error energy. We defined the threshold as the error energy such that the artifact was seen by 50% of our subjects. In other words, if we created a sequence with the artifact at a strength equal to the detection threshold, 50% of the people shown the sequence should see the artifact.

We estimated the probability of detecting each artifact by counting the number of people who detected the artifact and dividing by the number of observations (36). No subjects were excluded. We used the logarithm of the error energy (Eq. 5) for each artifact as the independent variable. The probability as a function of logarithmic error energy was fitted using the Weibull function.[11] The Weibull function has an S-shape similar to our data and is defined as

$$P(x) = 1 - 2^{-(Sx)^k}, \tag{6}$$

where $P(x)$ is the probability of detection, x is the logarithmic error energy, $1/s$ is the 50% threshold in logarithmic error energy, and k is a constant that determines the slope of the transition. Figure 1 shows two sets of curve fits. The curves correspond to each of the three artifact regions derived from the Bus and Hockey video clips.

This equation assumes that the probability of detection goes to zero as the artifact strength goes to zero. However, the detection probability was non-zero for a few of our original sequences. For the purpose of this curve fitting, we

Table 2. Detection threshold curve fit parameters and the threshold error energies.

	Curve Fit Parameters		Threshold	
Test Sequence	S	k	Error energy	PSNR
Bus - Top	0.3661	10	539.25	42.4
Bus - Middle	0.2967	24	2346.68	35.9
Bus - Bottom	0.3035	8	1972.25	38.0
Cheerleader - Top	0.2831	16	3406.11	34.9
Cheerleader - Middle	0.2867	16	3075.69	35.3
Cheerleader - Bottom	0.2734	13	4549.47	33.6
Flower garden - Sky	0.2783	11	3918.32	34.0
Flower garden - Houses	0.2847	15	3250.78	34.0
Flower garden - Flowers	0.2608	14	6825.29	32.5
Football - Left	0.3141	19	1525.32	38.3
Football - Middle	0.3418	20	841.90	40.9
Football - Right	0.3382	18	905.92	40.6
Hockey - Left	0.3651	15	547.36	42.4
Hockey - Middle	0.3900	16	366.79	44.8
Hockey - Right	0.3279	17	1121.01	39.7

assumed that these detections were errors and did not include them. As discussed in Section 4.2, this assumption seems good for our test sequences with the possible exception of the Flower-based video clips.

Table 2 summarizes the curve fit parameters found for each artifact. The table also includes the 50% detection threshold in terms of error energy and PSNR. The threshold PSNR values match well with our experience in using PSNR to evaluate image filtering algorithms. This range of PSNR values normally corresponds to fairly good results or, in other words, little visible noise. The detection threshold error energy for each artifact varies considerably, even when compared to other artifacts derived from the same original video. The slope parameter k also varies with artifact. However, the amount of variation in k is different for groups based on different original sequences.

5.2. Perceived Error Energy Ratio

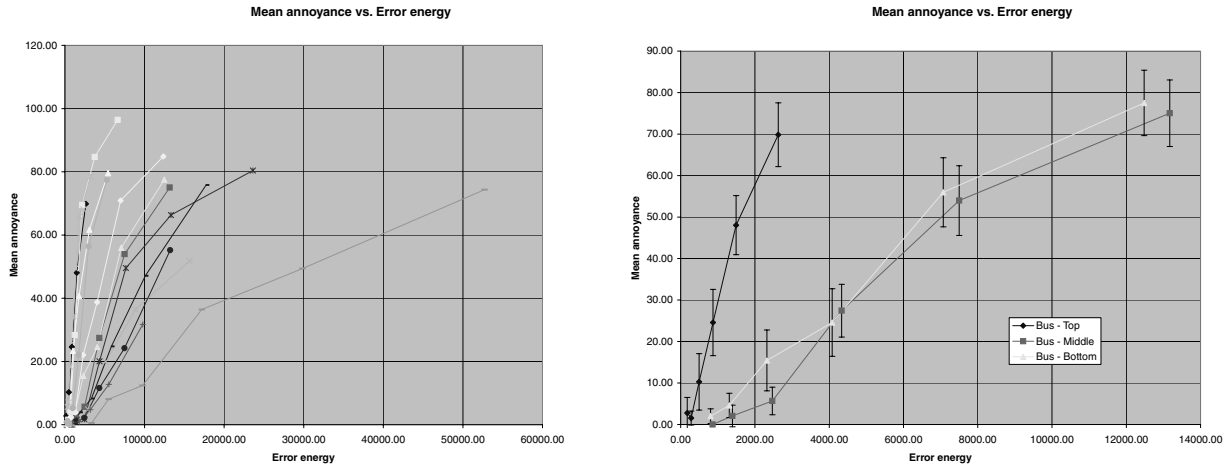
Figure 2 contains plots of the mean annoyance values versus the error energies. The lines connect the data points for sequences with the same artifact location, time interval, and original video sequence but with increasing artifact strength. For clarity, the overall figure in (a) does not contain the annoyance errors bars. Part (b) graphs the data for all of the sequences derived from the original Bus. To some extent, the error energy predicts the mean annoyance value. As the error energy increases, the mean annoyance increases. However, the rate of increase changes from artifact to artifact. For plot (a) in Figure 2, the overall correlation is 0.52.

From section 5.1, we have an estimate of the 50% threshold error energy for each of the artifacts that we introduced. The threshold error energies also vary between artifacts. If the thresholds contain all of the context dependent information needed to predict the mean annoyance, then we can bring the lines together by substituting the x-axis values with a function of the error energy and the threshold error energy for each artifact.

As an initial attempt along these lines, we divided the error energy for each sequence by the threshold error energy. Figure 3 shows the overall results and also the results for just the Bus sequence, with error bars. The correlation for the complete set of data is 0.94.

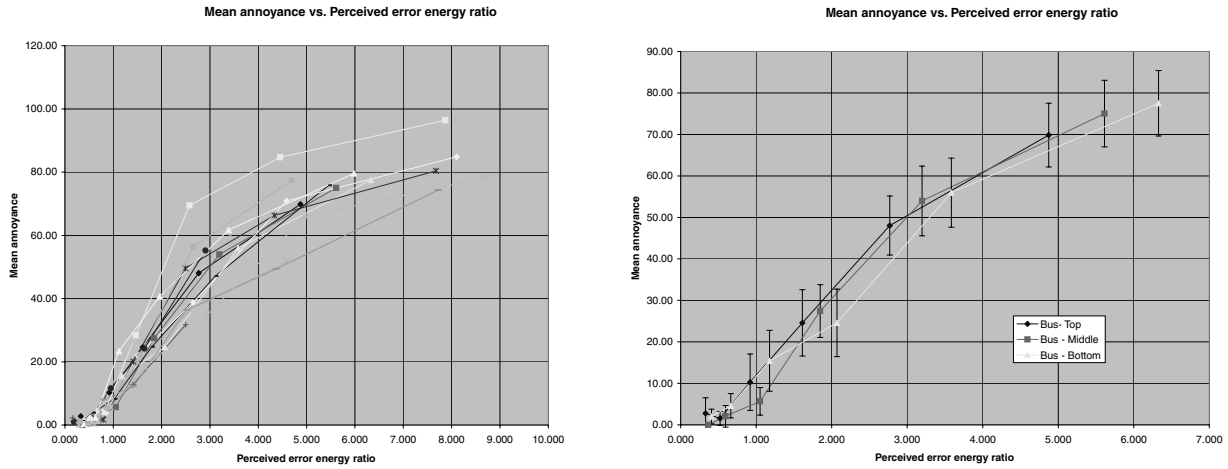
Weighting the error energy by the threshold error does bring the curves for all of the artifacts closer together. In other words, some of the content or location-based errors have been removed. The curves appear to be saturating as the perceived error energy ratio increases. So it probably is not valid to assume that the annoyance is a linear function of the perceived error energy ratio. However, as a first approximation, it is interesting that the correlation coefficient (0.938) shows a large increase over the raw error energy.

The PQA-200 includes a fidelity metric based on the Sarnoff Corporation JND algorithm.[16] We ran our test sequences through the PQA-200 implementation using the default settings. The overall results and the results for



(a) (b)

Figure 2. Plot of the mean annoyance values versus the total error energy for the (a) whole sequence set (15 artifacts) and (b) Bus sequences (3 artifacts). The correlation of the entire data set is 0.519.



(a) (b)

Figure 3. Plot of the mean annoyance values versus the perceived error energy ratio for the (a) whole sequence set (15 artifacts) and (b) Bus sequences (3 artifacts). The correlation of the entire data set is 0.938.

the Bus sequences are shown in Figure 4. The correlation for the complete set of data with the JND measure is 0.87.

The Sarnoff JND output is also plainly better than the raw error energy output. For the Bus set of test sequences, the JND algorithm removes much of the artifact specific variation, although not to quite the same degree as the perceived error energy ratio approach. However, as can be seen in the overall sequence plot, there are a few artifacts for which the JND algorithm overpredicts the annoyance (gave a high response when our viewers gave low responses and vice versa). This resulted in the smaller correlation coefficient when compared to the perceived error energy ratio.

5.3. Artifact Descriptions

MPEG-2 artifacts are commonly broken down into different types in the literature. For example, MPEG-2 video can be described as blurry or blocky. To find out how naive subjects would describe the artifacts we inserted into

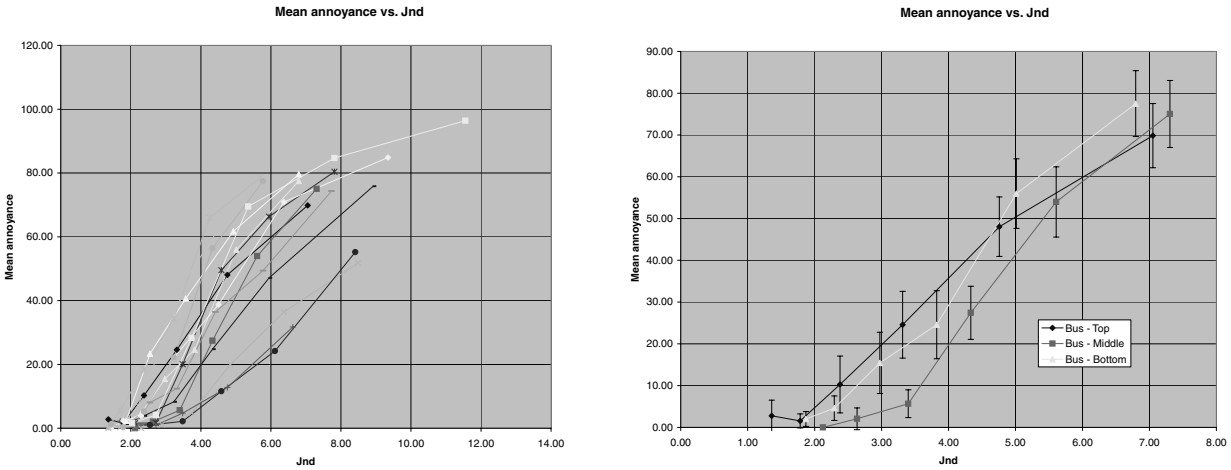


Figure 4. Plot of the mean annoyance values versus Sarnoff JND for the (a) whole sequence set (15 artifacts) and (b) Bus sequences (3 artifacts). The correlation of the data set is 0.871.

the videos, we asked each subject to describe the defects and if there was more than one type of defect. Because of time constraints, we only asked the questions at the end of the experiment. Therefore, we cannot associate the descriptions with particular artifacts (although the subjects sometime mentioned examples).

Many subjects mentioned that the defects differed in position, size and orientation. They were asked if there were any differences in type apart from these spatial differences, and they were scored yes only if they described other differences.

The three most common descriptions of the defects were blurry (or fuzzy), squares (or boxes, blocks, grids, mosaics, or large pixels), and lines (probably the defect edges). Most of the viewers identified two of these three features. Blur and squares or synonyms for these were the most often identified. These types were often seen in the same defect, although some subjects reported that they had seen defects that had just one of these features.

A substantial minority indicated that there was more than one type of defect. Their elaborations showed that some had characterized defects as blurry, blocky, or both, while others said that there was more than one type on the basis of which feature dominated the percept without claiming that there were defects that had only one feature.

In summary, the defects were generally described as blur and/or squares. Most defects were characterized by both of these features, but one of them might dominate. Some subjects reported defects that manifested just one of these features.

6. CONCLUSION

Our main goal was to develop an experiment and obtain reliable subjective quality results. This experiment has accomplished that goal.

The data confirms that the error energy is not a good predictor of annoyance values. Although the resulting curves are monotonically increasing, the rate of increase depends on the artifact. The variation can be a result of the artifact locations or of the characteristics of the regions. In future work, we will attempt to isolate the factors which varied between artifacts derived from a single original sequence and see which factors contributed to the change in annoyance value.

The detection data were used to estimate the 50% detection thresholds for our artifacts. Weighting the error energy by the threshold error energy produced a significantly better predictor than the raw error energy. The threshold does seem to incorporate much of the content or artifact specific information. However, there are some exceptions, and examination of the exceptions may lead to a method for enhancing fidelity metrics.

The commercial fidelity metric also performed significantly better than the raw error energy. However, the metric did not quite perform as well as the perceived error energy ratio. The differences may be due to the approach - the JND algorithm weights by the estimated threshold on a pixel-by-pixel basis, while we measured the overall artifact error detection threshold. This difference in performance suggests that there may be ways to improve the algorithm. Once again, examination of the artifacts where the algorithm most noticeably failed should provide some clues for improvement.

A significant number of test subjects described multiple artifact types with different appearances. Artifacts which differ qualitatively in their appearance may produce different levels of annoyance even though they have the same error threshold and the same error energy. To examine this possibility, we will do similar experiments using qualitatively different artifacts.

ACKNOWLEDGMENTS

This work was supported in part by a University of California MICRO grant with matching support from Lucent Technologies, National Semiconductor Corporation, Raytheon Missile Systems, Tektronix Corporation, and Xerox Corporation.

REFERENCES

1. C. J. van den Branden Lambrecht and M. Kunt, "Characterization of human visual sensitivity for video imaging applications," *Signal Processing* **67**, pp. 255–269, 1998.
2. J. J. Chae, D. Mukherjee, and B. S. Manjunath, "A robust data hiding technique using multidimensional lattices," in *Proceedings of the IEEE International Forum on Research and Technology*, (Santa Barbara, CA), 1998.
3. S. Thurnhofer and S. K. Mitra, "A general framework for quadratic volterra filters for edge enhancement," *IEEE Transactions on Image Processing* **5**, pp. 950–963, 1996.
4. W. Y. Ma and B. S. Manjunath, "A texture thesaurus for browsing large aerial photographs," *Journal of the American Society for Information Science* **49**, pp. 633–648, 1998.
5. S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing* **78**, pp. 231–253, October 1999.
6. J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision models for target detection and recognition*, E. Peli, ed., World Scientific Publishing, Singapore, 1995.
7. E. M. Yeh, A. C. Kokaram, and N. G. Kingsbury, "Psychovisual measurement and distortion metrics for image sequences," in *European Signal Processing Conference*, (Island of Rhodes, Greece), 1998.
8. ANSI T1.801.03-1996, *Digital transport of one-way video signals - parameters for objective performance assessment*, 1996.
9. J. B. Martens and V. Kayargadde, "Image quality prediction in a multidimensional perceptual space," in *IEEE International Conference on Image Processing*, (Lausanne, Switzerland), 1996.
10. E. A. Fedorovskaya, H. de Ridder, and F. J. J. Blommaert, "Chroma variations and perceived quality of color images of natural scenes," *Color research and application* **22**(2), pp. 96–110, 1997.
11. N. V. S. Graham, *Visual pattern analyzers*, Oxford University Press, New York, 1989.
12. ITU Recommendation BT.500-8, *Methodology for the subjective assessment of the quality of television pictures*, 1998.
13. W. J. Tam and L. B. Stelmach, "Perceived quality of MPEG-2 steroscopic sequences," in *SPIE Conference on Human Vision and Electronic Imaging II*, vol. 3016, (San Jose, CA), 1997.
14. C. C. Taylor, Z. Pizlo, and J. P. Allebach, "Perceptually relevant image fidelity," in *SPIE Conference on Human Vision and Electronic Imaging III*, vol. 3299, (San Jose, CA), 1998.
15. R. Hamberg and H. de Ridder, "Continuous assessment of time-varying image quality," in *SPIE Conference on Human Vision and Electronic Imaging II*, vol. 3016, (San Jose, CA), 1997.
16. J. Lubin, *Sarnoff JND Vision Model*. Constribution to the IEEE Compression and Processing Subcommitte, G-2.1.6, 1997.