A Comparison of the Detectability and Annoyance Value of Embedded MPEG-2 Artifacts of Different Type, Size and Duration

Michael S. Moore[a], John M. Foley[b], and Sanjit K. Mitra[a]

[a]Department of Electrical and Computer Engineering and [b]Department of Psychology
University of California, Santa Barbara
Santa Barbara, CA 93106 USA

ABSTRACT

One fundamental problem in predicting the subjective quality of a degraded video is that the perceived quality depends on the properties of the video itself, or the *context* of the degradation. In this paper, we present the results for a series of experiments designed to measure the detection thresholds and annoyance values of small regions of MPEG-2 artifacts inserted into mostly uncorrupted video sequences. In previous work, we found that the detection threshold contains much, but not all, of the information needed to remove the dependence of quality on the context. In this paper, we report the result of two experiments. In one experiment, we varied the type of MPEG-2 artifacts inserted into the test sequences. In the other experiment, we varied the location, size, and duration of the corrupted regions in the test sequences. From each set of data, we estimated detection thresholds and fitted the parameters of a quality function. The experimental results demonstrated that, under a wide set of test conditions, the detection threshold is still very useful for the estimation of quality as context varies. In fact, the detection threshold was the only factor necessary to model the changes in the quality function parameters with artifact type. The experimental data showed that the detection threshold and the quality function parameters do depend on the size and duration of the degraded region. However, the effects of size and duration are minor relative to the effect of artifact location.

Keywords: video, quality, MPEG-2, metrics, annoyance, appearance, duration, size, location

1. INTRODUCTION

The use of digital video is growing. Digital video data can be transformed in complex ways. Compression techniques are often used to reduce the cost of storage and transmission of digital video. Compression generally results in information loss and may introduce visible defects into video. If the defects are visible, they may be annoying to human observers.

We can control both the strength and the subjective appearance of digital video defects by using different compression algorithms or by changing the algorithm parameters. The goal in selecting a compression algorithm is to minimize both the visibility of defects and the annoyance caused by any visible defects. To achieve this goal, we need measures of detectability and annoyance. At the present time there are no physical measurements that we can make to accurately determine the detectability and annoyance value of artifacts. They have to be determined by psychophysical experimentation. Our goal is to create a model that will allow us to predict detectability and annoyance for a typical observer from the physical properties of video sequences.

There are two main approaches to creating such a model. In the first approach, several general models of perceived video fidelity have been proposed.[1-7] These models compare the output of a video transformation to the original input and compute a measure of fidelity. Recent fidelity measures incorporate models of the human visual system (HVS). Ideally, these models will provide a general fidelity measure that applies to all systems, regardless of the video content and the transformation process. In the second approach, defect-specific models have been proposed.[8-10] These models compute the strength of common types of defects, such as blockiness or blurriness. In many cases, the types of defects depend strongly on the transformation process. As a result, defect measures are typically designed to go with a particular transformation process. In this paper, we investigate the general HVS model approach.

Many models of the low-level HVS include filtering and masking stages that were developed using threshold experiments. However, while models based on threshold experiments may be used to predict the visibility of defects, the best way to use the models to predict quality is not obvious. Last year, we reported the results of an experiment designed to measure *both* the detection threshold and the annoyance value of MPEG-2 defects inserted into otherwise unchanged digital videos[11]. In that experiment, our main goal was to determine the usefulness of the detection threshold for estimating video quality.

In our experiment, the defect detection threshold was closely related to the annoyance that it produced. Our test subjects consistently reported small annoyance values at threshold. Therefore, if we know the threshold in terms of an objective

measure, we can 'fix' the lower end of an annoyance versus objective measure function. However, to completely model the annoyance results, we need to be able to find another point on the curve, preferably near the annoyance saturation point. Alternatively, it may be sufficient to predict the steepness of the curve. Our experimental results suggested that the slope of the curve does vary from sequence to sequence and that it does not significantly depend on the detection threshold.

In this paper, we report several variations of the original experiment. In each variation, we measured both the detection threshold and annoyance value for our test sequences. We have varied the type of MPEG-2 defects injected and the duration, size, and location of the corrupted region in each test sequence. We had three goals. First, we needed to verify the usefulness of the detection threshold for predicting quality under a wider set of test conditions. Second, we wanted to see how the detection thresholds varied with these changes. Finally, we were looking for another variable that would reduce the remaining variation in the predicted quality by predicting another point on (or the steepness of) the annoyance value curve.

In the next section, we will briefly review the experimental paradigm used in our experiments. The basic experiment is described in more detail elsewhere,[11] but a summary is provided here to understand the experiments described in this paper. Section 3 describes the two new experiments and analyzes the results. The final section summarizes the principal findings of these experiments.

2. METHODS

The normal approach to subjective quality testing is to degrade an entire 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.). For MPEG-2 video compression, the errors introduced into the reconstructed video are rarely spread evenly. The amount of error and its appearance can vary from region to region. Some recent experiments have started to look at limited regions of a compressed signal, either spatially in images,[14] temporally in video,[15] or both.[11]

In previous work,[11] we used an experimental paradigm that measured the annoyance value of brief, spatially limited MPEG-2 artifacts in video. We degraded an isolated region of the video clip for a short time interval with the rest of the video clip left in its original state. The test subjects were asked to search each video clip for defective regions and to indicate the annoyance value caused by any defects or impairments that were seen. We used the same paradigm for the two experiments described in Section 3. In this section, we summarize the important aspects of the test sequence generation procedure and outline the methods used for data analysis.

2.1. Test Sequences

To generate test video clips, we started with a set of seven original video clips of assumed high quality: Amusement park, Bus, Cheerleader, Flower-garden, Football, Hockey, and Susie. The video clips were all five seconds long and contained scenes that we thought were typical of normal television. For each experiment, a subset of the available original sequences was chosen. In general, the number of original sequences should be maximized. The phenomenological appearance of the defects, the visibility of the defects, and the resulting annoyance all depend on the original sequence. However, no more than one hundred sequences can be shown during a 30-minute test session. Therefore, the number of originals must be weighed against the number of variations of each original that we want to test. We selected five original sequences for the first experiment (Bus, Cheerleader, Flower-garden, Football, and Hockey) and three original sequences for the second experiment (Amusement park, Football, and Susie).

After the original videos were selected, each original was broken into defect zones. A *defect zone* is defined to be the specific section (time interval and spatial region) of a video where defects were inserted. Both the detection threshold and the perceived annoyance depend on the size and position of the zone. In our original experiment,[11] we picked regions that were of approximately constant size (about one-third of the video frame) but with varying orientations and locations. Figure 1 shows the regions chosen for one of the sequences (Flower-garden). For each region, a corresponding time interval was chosen. The first and fifth seconds of each video are never altered to avoid end effects. In our previous experiment, random one-second intervals were chosen within the available three-second window. In the experiments reported here, we varied both the size and the duration of the defect regions over a greater range.

Within a defect zone, we inserted corrupted video that was derived from an MPEG-2 encoded and reconstructed version of the same video. There are many ways to corrupt a video. We chose to use MPEG-2 artifacts in our experiments because of their practical importance. However, MPEG-2 artifacts are difficult to categorize. Depending on the original video content and the encoder bit-rate goal, several different types of artifacts may appear.[16] We controlled the content, to some extent, through our choice of defect region.

We also controlled the artifact strength. Artifact strength is often controlled by varying the encoder bit-rate goal. However, changing the bit-rate goal also tends to change the error pattern and the phenomenological appearance of the MPEG-2 artifacts. The detection threshold and annoyance values depend on both the strength *and* the appearance of the artifacts.

(a) (b) (c)

Figure 1. The three regions selected from one of our video sequences are shown in reverse video. They were (a) the sky, (b) the houses, and (c) the garden. In each test sequence, only one of these regions was corrupted.

To avoid changing the artifact appearance, we created a single degraded sequence for each original sequence using a single bit-rate goal. The bit-rate was chosen to produce highly annoying defects with similar appearance in all of the defect zones. The difference between the original and the degraded sequences constituted an *error pattern*. The error pattern was viewed in the *context* of the original sequence. By scaling an error pattern, we created multiple test sequences with different levels of impairment without changing either the context or the artifact appearance. This method created sets of test sequences that differed only in the *contrast* of the error pattern.

To create a test sequence, both the original and the reconstructed digital video signals were transformed to the linear light domain using a gamma approximation. The difference between the two transformed video signals was scaled and added to the original video. Specifically, for a specific position *(x,y)*, frame *t*, and color component *c*:

$$T_i(x,y,t,c) = \begin{cases} I(x,y,t,c) + a_i\,(M(x,y,t,c) - I(x,y,t,c)), & \text{in the defect zone, and} \\ I(x,y,t,c), & \text{elsewhere.} \end{cases} \quad (1)$$

I(x,y,t,c) was the original video, *M(x,y,t,c)* was the reconstructed video, and a_i was a weighting factor used to control the error pattern contrast. A set of test sequences $T_i(x,y,t,c)$ that share the same defect zone was created by varying a_i. The values of a_i were selected to vary the contrast of the error pattern from below the detection threshold to visible-and-highly-annoying.

In the original experiment, the encoder bit-rate goal was chosen to produce strong blocky and/or blurry artifacts in the reconstructed videos. The weighting factor was varied between zero and one. Reducing the weighting factor effectively removed the block pattern and added back the details lost to blurring. In the first experiment reported in this paper, we used a less restrictive bit-rate goal (7.5 Mbps) to produce mild quantization noise artifacts in the reconstructed video. These were described by subjects as "fuzzy" rather than "blocky" or "blurry". To create a full range of test sequences, the weighting factors were allowed to increase above one for this experiment. This increased the contrast of the error pattern beyond the level found in the reconstructed video.

In summary, a set of test sequences was generated for every experiment. Each test sequence consisted of mostly uncorrupted video with a portion of a few frames replaced with impaired video. In each experimental set, several defect zones were chosen for a set of original sequences. The contrast of the error pattern in each defect zone was by changing a scaling factor. From experiment to experiment, we varied the appearance, duration, size, and location of the defects.

2.2. Data Analysis

For every test sequence, the number of subjects who saw defects and the annoyance values for those defects were recorded during an experiment. We first converted the raw data into estimates of the probability of detection and mean annoyance values along with associated 95% confidence intervals for each sequence. Next, we calculated a simple objective difference measure for each test sequence. Finally, the probability of detection and the mean annoyance values were plotted relative to the difference measure. Each individual curve was fitted using a least-squared-error algorithm.

The detection data was binary. A subject either did or did not report seeing defects or impairments in a test sequence. Assuming that the trials were independent of each other, the detection data should conform to the binomial distribution. We calculated the probabilities and 95% confidence intervals using the `binofit` function in MATLAB.

For the annoyance data, our goal was to find an 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] These procedures were used in the experiments reported here.

No annoyance value was recorded if a subject did not see any artifacts in a test sequence. In our original data analysis,[11] we assumed an annoyance value of zero in this situation. These zeroes were included in the mean annoyance and confidence

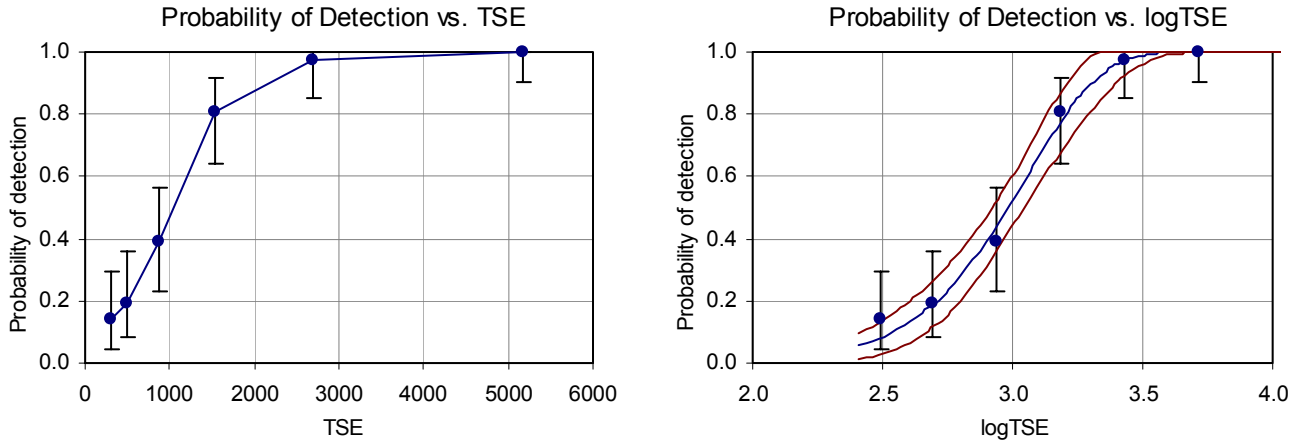

Figure 2. Probability of detection data from our first experiment. Plot (b) also shows the 95% confidence interval for the least-squares Weibull curve fit. This data is from one of our test sequences (Bus –Top error region).

interval calculations. However, these values have been excluded from the mean annoyance and confidence interval calculations in this paper. Although the test subject cannot not be annoyed by a flaw that was not seen, that does not imply that the subject would have had zero annoyance if the flaw was seen. Therefore, the mean annoyance values and confidence intervals were calculated using only the subset of test subjects that actually saw the artifacts.

In this paper, we use the total squared error as the objective measure of the difference between the original and test sequences. The total squared error (TSE) is defined as:

$$TSE_i = \sum_x \sum_y \sum_t \sum_c \left[I(x,y,t,c) - T_i(x,y,t,c) \right]^2 , \tag{2}$$

where I and T are the original and test video signals, respectively, in the linear light domain. The TSE is closely related to the mean-square-error (MSE) and peak-signal-to-noise-ratio (PSNR):

$$MSE_i = \frac{1}{N} TSE_i \text{ and } PSNR_i = 10 \log_{10} \left(\frac{N \cdot 255^2}{TSE_i} \right),$$

where N is the number of pixels per frame times the number of frames times the number of color components and a peak pixel value of 255 is assumed. We used TSE because we wanted a measure of total error. Both the MSE and PSNR are measures of error averaged over pixels. As a result, these measures do not necessarily change with the size or duration of a defect zone.

Figure 2(a) shows the detection probability plotted against TSE for a test sequence in our original experiment. The function is positively accelerated at the bottom left and negatively accelerated at the upper right. If the detection probabilities are plotted against the $\log_{10}$ of the TSE, as shown in Figure 2(b), the data can be fitted with a symmetric function. We used the Weibull function:

$$\hat{P}(x) = 1 - 2^{-(Sx)^k} , \tag{3}$$

where $\hat{P}(x)$ is the detection probability and x is the $\log_{10}$TSE. The parameter S (sensitivity) is equal to the inverse of x at a detection probability of 0.5. The product of S and k is proportional to the first derivative of the function at a detection probability of 0.5, or the *midpoint slope* for the Weibull function. The measured detection probabilities were fitted with Eq. (3) using the nonlinear least-squares algorithm implemented in the MATLAB function `nlinfit`.

Figure 3(a) shows the mean annoyance values plotted against TSE for a test sequence in our original experiment. Once again, the data has a small positive acceleration at the bottom left and a large negative acceleration at the upper right. If the mean annoyance values are plotted against $\log_{10}$TSE, as in Figure 3(b), the data can be fitted with the logistic function:

$$y = y_{\min} + \frac{(y_{\max} - y_{\min})}{1 + \exp(-(x - \bar{x})/|\beta|)} , \tag{4}$$

where y is the predicted annoyance value and x is the $\log_{10}$TSE. The parameters y_{min} and y_{max} establish the limits on the annoyance value range. The parameter $\bar{x}$ translates the curve in the x-direction and the parameter β controls the slope of the

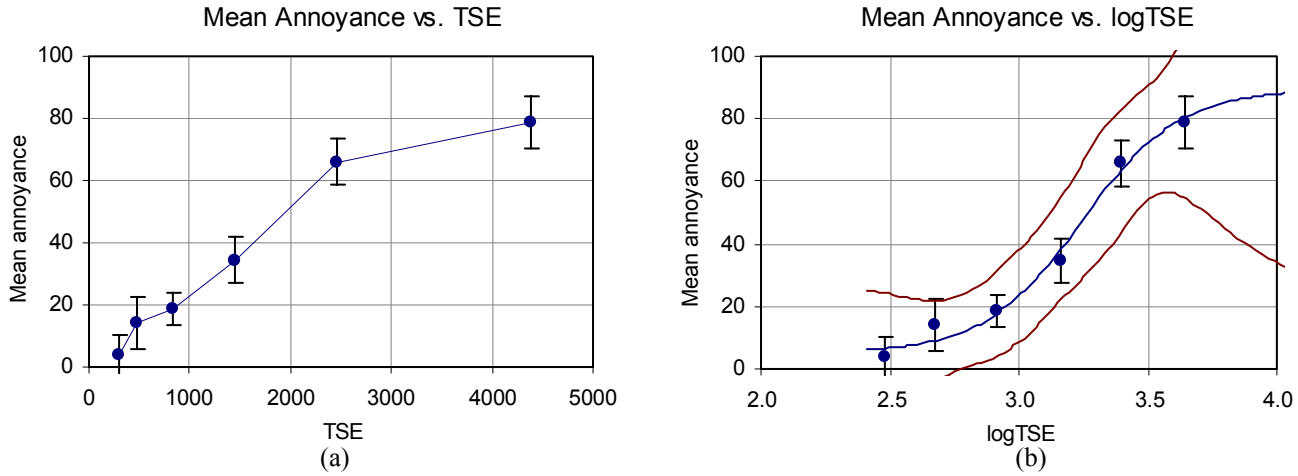

Figure 3. Mean annoyance value data from our first experiment. Plot (b) also shows the 95% confidence interval for the least-squares logistic curve fit. This data is from one of our test sequences (Hockey – Middle error region).

curve. The mean annoyance values were also fitted with Eq. (4) using the Matlab function `nlinfit`. However, in this case, the differences between the measured and predicted values were weighted by the inverse of the measurement variance. This procedure was adopted by the Video Quality Experts Group (VQEG) to assign greater weight to the data points that were most accurately measured[17].

3. EXPERIMENTS

In this paper, we report two new experiments. In the first experiment, we used "fuzzy" defects created as described in Section 2.1. In the second experiment, we varied the size and duration of the error regions in the test sequences to see how these parameters affect the detection probabilities and annoyance values.

In this section, we discuss the results from these experiments. The sections are broken down by experiment. In each section, the effect of the changes on both the detection probabilities and the annoyance values are considered. We analyze the changes in terms of the curve fit parameters. If we could predict the curve fit parameters reliably, then we could predict the detection probabilities and annoyance values using the simple TSE measure. Although we cannot do that yet, these experiments were intended to provide a starting point.

3.1. Experiment 1: Fuzzy Artifacts

Our test sequences are created by combining two other sequences: a high-quality, original sequence and a low-quality, MPEG-2 compressed sequence. In our original experiment,[11] the low-quality sequence was created by compressing the original to a very high degree. The test subjects reported that the resulting sequences had a high degree of blocky/blurry artifacts.

However, MPEG-2 compression does not only produce blocky and blurry artifacts.[16] If the original is not highly compressed, the blocky artifacts may be reduced to the point where they disappear and many image details are preserved. However, there may still be some visible noise, mostly due to variations in encoder performance from one frame to another. This noise looks different from the impairment introduced in the original experiment. The noise is frequently described differently in published literature, where it may be described as edge busyness, mosquito noise, or general quantization noise. Our test subjects described it as "fuzzy".

Because we used the same original sequences and the same error regions for this experiment as the original experiment,[11] the only difference between the two sets of sequences was the specific error pattern in each sequence. The context, which is defined by the combination of the original sequence and the defect zone, was not changed. Therefore any significant differences between the detection probabilities and the annoyance values can be attributed to the error pattern type (blocky/blurry or fuzzy).

This experiment was performed using 34 test subjects. Both detection and annoyance data were gathered. The training and practice sets were chosen from among the test sequences used in this experiment. Therefore, the annoyance value scale may be different for this experiment than for the original experiment.

Figure 4(a) shows the detection probabilities and Weibull curve fits for one error region in a set of sequences derived from the Cheerleader original sequence. Both the curves for the original experiment and for this experiment are plotted. This

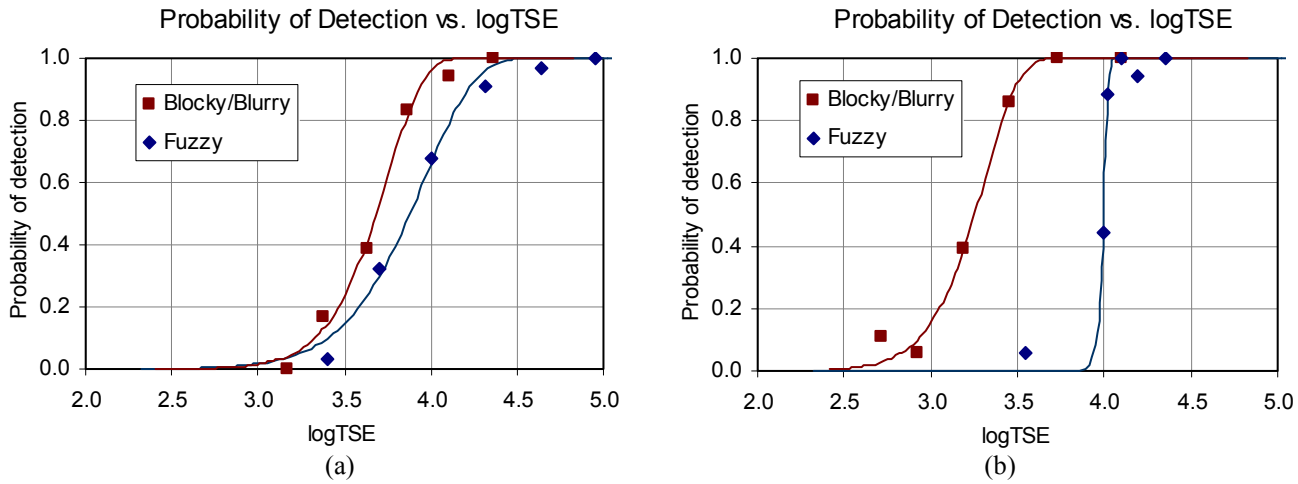

Figure 4. Probability of detection versus the total squared error for the (a) Cheerleader-Top and (b) Hockey-Right test sequence sets. The relationship between the curves in (a) is more typical than the relationship between the curves in (b).

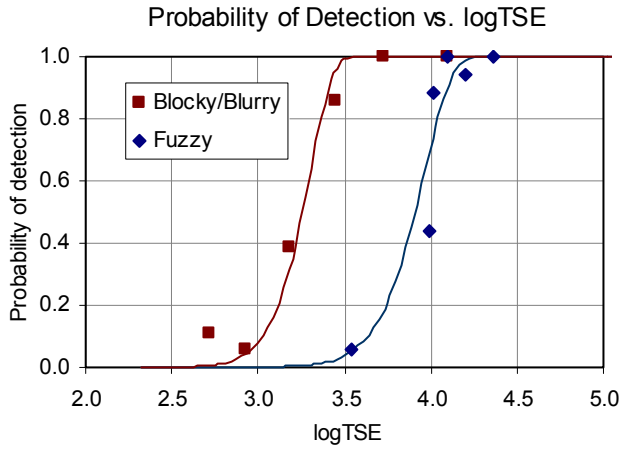

Figure 5. Curve fits for the Hockey-Right sequence where k is constrained to be the same for both curves.

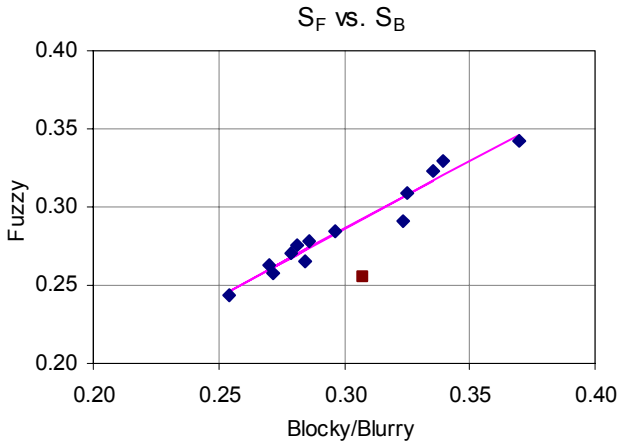

Figure 6. Curve fit S parameter values for the two sets of data. The block data point represents the Hockey-Right defect zone and was excluded from the linear curve fit.

set of curves is typical. In every set of test sequences, the threshold for the blocky/blurry sequences is smaller than the threshold for the fuzzy sequences. With the exception of the Hockey-Right set of sequences, the shift is small and consistent. The detection probability curves for Hockey-Right are plotted in Figure 4(b).

We performed two additional sets of curve fits for the experimental data. In one set of curve fits, the constant k was constrained to be the same for each error region. In essence, k became a function of context while S was allowed to vary with both context and error pattern type. The second set of curve fits constrained the midpoint slope to be the same for each error region. In this case, the product Sk became a function of context while S was allowed to vary. The results were essentially equivalent, although the constant k fits were slightly better. Figure 5 shows an example of a fit with a constant value of k for the Hockey-Right set.

Excluding the Hockey-Right error region, the S parameters of the curve fits between the two experiments were highly correlated (linear Pearson correlation coefficient of 0.95). Figure 6 plots the parameters from the two experiments. The linear fit to the data points is:

$$S_F = 0.86 S_B + 0.03 \text{ for the constant } k \text{ curve fits,} \quad (5)$$

where S_F is from the fuzzy test set and S_B is from the blocky/blurry test set. The linear correlation between the S parameters was improved by holding k constant. Thus, with the exception of one set of sequences, our subjects were less sensitive to the fuzzy defects by approximately a constant factor.

The Hockey-Right test set is a significant outlier in the sensitivity plot shown in Figure 6. The red block in Figure 6 represents this defect zone. The change in threshold with respect to error pattern type for the sequences derived from the

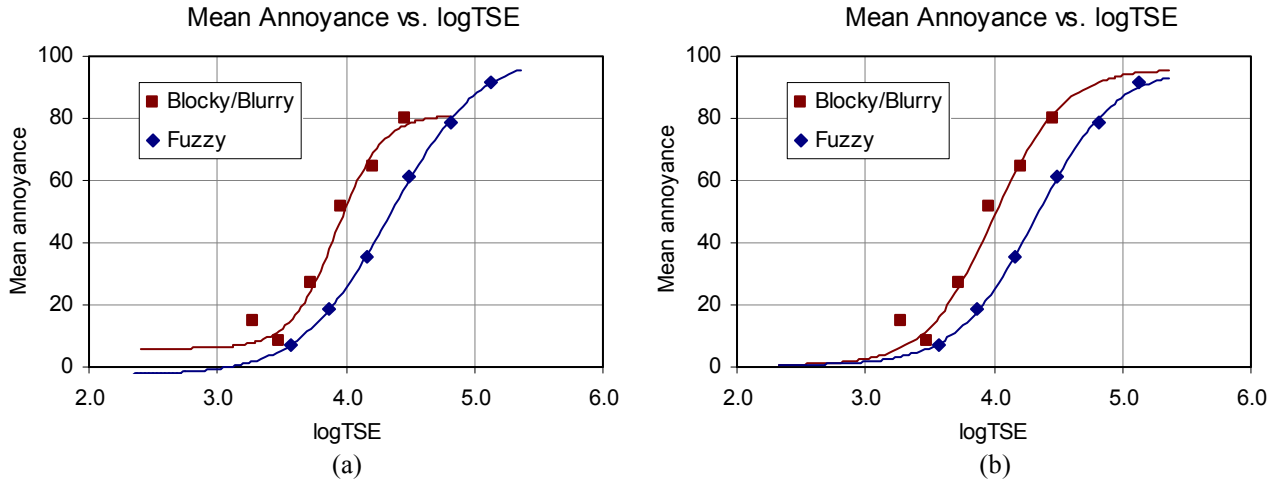

Figure 7. Mean annoyance versus log total square error for the Cheerleader-Middle sequence set: (a) has independent curve fits while the curves in (b) were constrained to have the same annoyance range.

Hockey sequence in the right defect zone was unusually high compared to the change for other defect zones. The only difference between the two sets of test sequences was the error pattern type. The context must have interacted with the error patterns much differently in this defect zone to account for the unusual change in threshold.

The Hockey-Right defect zone was unique in one significant way. The Hockey sequence contains a scene cut from one camera to another. The Hockey-Right error region was temporally located just after this scene cut. The temporal discontinuity may have suppressed the visibility of the error patterns differentially. The blocky/blurry error pattern is mostly characterized by spatial changes while the fuzzy error pattern is characterized by rapid temporal changes (busyness). The temporal discontinuity may have masked the temporal (fuzzy) error pattern to a much greater extent than the spatial (blocky/blurry) error pattern. However, we have not yet tested this hypothesis.

Figure 7(a) shows the mean annoyance values and the unconstrained logistic curve fits for one defect zone in a set of sequences derived from the Cheerleader original sequence. Both the curves for the blocky/blurry sequences and for the fuzzy sequences are plotted. This pair of curves is typical. As for detection, in every set of test sequences, the curve for the blocky/blurry sequences is to the left of the curve for the fuzzy sequences. With the exception of the Hockey-Right set of sequences, the shift is small and the slopes of the functions are similar. The curves look similar, but unlike for the detection results, there is no consistent relationship between the *unconstrained* curve fits parameters for the two experiments.

However, some of the curve fit parameters cannot be precisely estimated from our data. For example, if the test sequence set does not contain sequences at or near the annoyance saturation point, y_{max} cannot be accurately estimated. The data points may be accurately fitted with a number of curves with widely varying values of y_{max}. To compensate for this fact, we guessed that the annoyance range parameters y_{max} and y_{min} did not depend on the error pattern type if the two sequences are otherwise identical. We performed another set of logistic curve fits. This time, the annoyance range was allowed to vary from defect zone to defect zone, but not from experiment to experiment. In other words, we assumed that the annoyance range only depends on context. The other parameters, $\bar{x}$ and β, were allowed to vary from experiment to experiment. The same set of data points resulted in the curve fits in Figure 7(b) using this procedure.

Constraining the annoyance range parameters reduced the variability of these parameters between defect zones. The parameters were essentially being fitted with two sets of data. One data set frequently compensated for the deficiencies of the other. Several relationships were revealed among the other parameters. First, there was a very strong linear relationship for the values of $\bar{x}$ between the experiments for all of the sequence sets, including Hockey-Right. The Pearson correlation coefficient was 0.96. Second, excluding the Hockey-Right data point, the values of β were identical between the experiments with a correlation coefficient of 0.85. Once again, Hockey-Right was an outlier. The exact parameter fits between the two experiments are:

$$\bar{x}_F = 1.07\bar{x}_B + 0.16 \text{ and } \beta_F = 0.99\beta_B + 0.03, \quad (6)$$

where the subscript F refers to the fuzzy data set and the subscript B refers to the blocky/blurry data set.

One of the purposes of this experiment was to see if the relationship between $\bar{x}$ and the detection threshold would be preserved for different error patterns. The thresholds from our original experiment[11] were well correlated with $\bar{x}$. The Pearson correlation coefficient for this relationship was 0.89. It was this relationship that we used to develop the perceived error

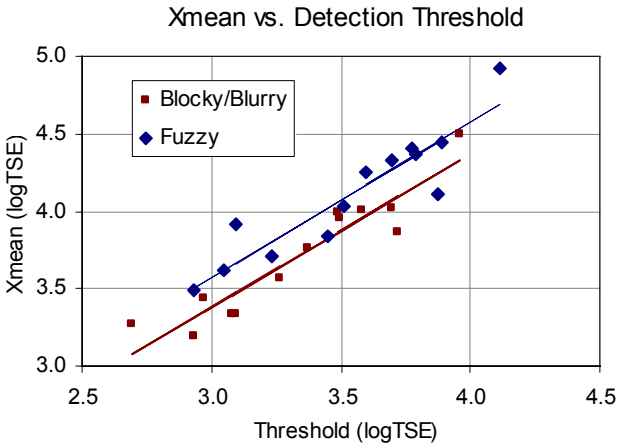

Figure 8. $\bar{x}$ versus the detection threshold for both experiments. The slopes of both curves are approximately one.

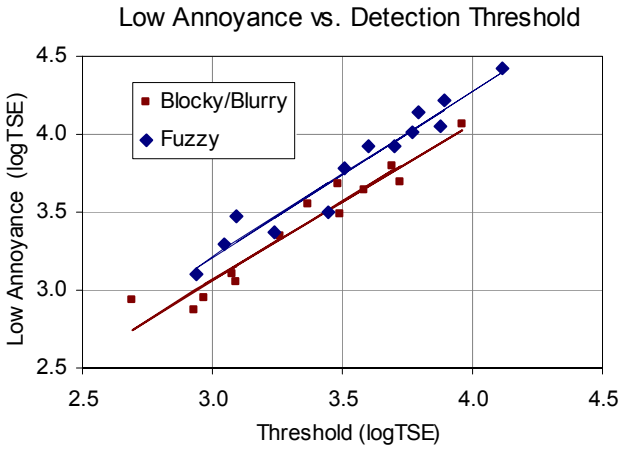

Figure 9. The $\log_{10}$ TSE for an annoyance value of 20 versus the detection threshold. Both curves intercept near zero.

energy ratio (PEER) measure.[11] In the fuzzy error pattern experiment, the Pearson correlation coefficient using all of the test sequences was 0.85. The following linear curve fits for both functions are shown in Figure 8:

$$\bar{x}_B = 0.99T_B + 0.40 \text{ and } \bar{x}_F = 1.00T_F + 0.57, \quad (7)$$

where T_B and T_F are the detection thresholds in terms of $\log_{10}$TSE for the blocky/blurry and fuzzy error pattern types, respectively.

As discussed in Section 1, the best use of the detection threshold is to fix the lower section of the annoyance value curve. The parameter $\bar{x}$ is the $\log_{10}$TSE value for the middle of the annoyance value range. Given the annoyance function parameters, we can invert Eq. (4) and calculate the $\log_{10}$TSE value for a low annoyance value, such as ten or twenty. For both experiments, the Pearson correlation coefficient increased when lower annoyance values were used instead of $\bar{x}$. Using an annoyance value of twenty, the data points in Figure 9 were calculated. The original experiment data had a Pearson correlation coefficient of 0.93 while the fuzzy experiment data had a correlation coefficient of 0.95. The following curve fits are plotted in Figure 9:

$$x_B(20) = 1.01T_B + 0.03 \text{ and } x_F(20) = 1.06T_F + 0.03. \quad (8)$$

Although the relationships represented by Eq. (8) are stronger than those in Eq. (7), the relationships in Eq. (7) are easier to use. These relationships imply that value of $\bar{x}$ should simply be the detection threshold plus an error pattern specific offset.

The comparison between the results for the blocky/blurry and fuzzy sequences can be summarized as follows: First, blocky/blurry artifacts have a lower threshold, in terms of TSE, than fuzzy artifacts. In other words, humans have a higher sensitivity to these types of artifacts. Second, the annoyance value functions are affected by the higher sensitivity to blocky/blurry artifacts – the annoyance caused by these artifacts starts increasing at a lower TSE than for fuzzy artifacts. Third, for detection probability curve fits, the value of k is independent of the error pattern. Fourth, for annoyance value curve fits, the parameters y_{max}, y_{min}, and β are independent of the error pattern. All of these parameters do change with the video context. Finally, for both experiments, the detection threshold is a good predictor of the value of $\bar{x}$.

If the only annoyance function parameter that changes with artifact appearance is $\bar{x}$, $\bar{x}$ can be predicted using the detection threshold and a constant, and the detection thresholds for different types of artifacts have a simple relationship like the one in Eq. (5), then modeling different types of artifacts may be greatly simplified. However, we need to run other experiments with new error patterns to verify these conclusions.

3.2. Experiment 2: Effects of Duration and Size

In the first two experiments, the size and duration of all of the defect zones were approximately the same. Each defect zone encompassed approximately one-third of the frame area and lasted for one-second. Five original sequences were used with three defects zones per original for a total of fifteen zones. The error pattern in each zone was presented at six different contrast levels, ranging from near threshold to highly annoying. Each experiment was run with 32 to 36 test subjects.

In the Experiment 2, we varied the duration, size, and location of the defect zones within the experiment. Three original sequences were chosen (Amusement park, Football, and Susie). Each original sequence was partitioned into 21 defect zones: fourteen that varied in size and location but not duration and seven that varied in duration but not size or location. The blocky/blurry error pattern type was used, as in the original experiment.[11] Each error pattern was shown to 20 test subjects at

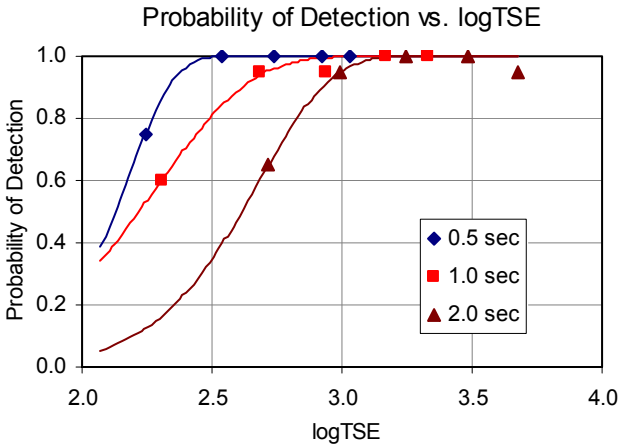

Figure 10. Detection probabilities and fits for the duration experiment with test sequences based on the Susie original.

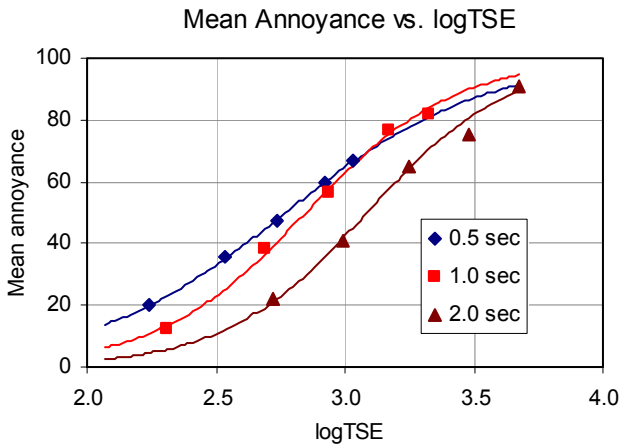

Figure 11. Mean annoyance values and fits for the duration experiment with test sequences based on the Susie original.

five different contrast levels. The results for the defect zones that varied in duration are presented in Section 3.2.1. The other defect zones are discussed in Section 3.2.2.

3.2.1. Duration

Identical duration defect zones were chosen for each original video. The spatial region was square and centered in the video frame. It covered an area equal to one-third of the frame area. Seven time intervals were chosen: four intervals of one-half second duration, two intervals of one-second duration, and one interval of two-second duration. All of the time intervals fell within the middle two seconds of the five-second video sequence.

Figure 10 shows the detection probability data and Weibull curve fits for three of the seven time intervals derived from the Susie video sequence. The figure illustrates a major problem with this set of data. In previous experiments, the test sequences were chosen so that at least one test sequence was below threshold (0.5). To ensure that this is true, one must have an accurate initial estimate of the detection threshold prior to generating the test sequences. Our initial estimates for the duration sequence thresholds were poor. As a result, nearly all of the duration sequences were above threshold. In a few cases, enough sequences were below threshold to produce a good Weibull curve fit. However, in most cases the data does not provide enough information for a good fit, especially for the slope (k) parameter, and sometimes the best we could do was establish an upper limit for the detection threshold.

Even with these problems, the detection data still illustrate two points. First, the probability of detection in terms of TSE increased for all three sets of data. Second, both the initial threshold and the rate of increase with duration depended on the original sequences. In other words, the initial threshold and the rate of increase depended on the context. Simply scaling the TSE, as is done when calculating the MSE, would not eliminate the threshold variation between the sets derived from different originals.

These conclusions eliminate two of the simplest possibilities for accounting for the effects of duration. If the detection threshold stayed constant as duration increased, then error would be simply integrated over time by summation. In other words, the detection probability of a strong brief defect would be the same as the detection probability of a weaker, briefer defect with the same TSE. This was definitely not true, as the detection threshold increased as duration increased. If the detection threshold increased in proportion with the duration (doubled when the duration doubled), then error should not be integrated over time at all. This was not true either. All of the detection thresholds increased, but at a rate significantly less than the rate at which the duration increased. The best model lies somewhere in-between.

One benefit of having artifacts above threshold was that more subjects provided annoyance values for the test sequences. Figure 11 contains a plot of mean annoyance for the same test sequences as Figure 10. These curves are typical. Although the experimental data is good for estimating $\bar{x}$ and β, most of the error regions did not include enough sequences near threshold and past saturation to estimate y_{max} and y_{min} accurately. In these plots, the curve fit annoyance value range parameters have been fixed to 100 and zero, respectively.

Figure 11 shows two trends that are consistent in the experimental results. First, increasing the duration shifted the curves to the right in terms of the $\log_{10}$TSE. The amount of shift correlated very well with the estimated thresholds. The Pearson correlation coefficient for a linear fit of the detection threshold and the $\log_{10}$TSE for an annoyance value of ten was

0.95 for the complete data set. Second, the midpoint slope of the annoyance value curves increased as the duration increased. The fact that the slope increased was consistent across data sets. However, the rate of increase depended on the context.

The Amusement-Park original sequence contains a scene change at precisely the middle frame of the sequence. The time intervals were arranged symmetrically around the scene change. However, the scene change did not have a consistent effect on either the threshold or the annoyance value curves. For example, the detection threshold was higher for the one-second interval immediately following the scene change, but the detection threshold for the half-second right after the change was lower. The main effect of the scene change seemed to be a greater variability in the annoyance values that were recorded immediately after the change.

3.2.2. Size and Location

As discussed in Section 3.2, the original sequences were divided into 21 defect zones in the third experiment. Fourteen zones had a constant time interval but varied in size and location. Five different contrast levels of blocky/blurry error patterns were created for each original. Over 300 test sequences were derived from the originals using these defect zones. All of the test sequences were shown to approximately 20 test subjects.

All of the size and location defect zones had a one-second duration. The defect zone occurred in the third, or middle, interval in the five-second video sequences. Nine of the fourteen spatial regions covered one-ninth of the frame. The frame was divided into nine equal size rectangles arranged in a tic-tac-toe pattern. Four of the remaining regions covered one-third of the frame. These defect zones were made up of three of the smaller zones. Two defect zones crossed the middle of the screen by combining three smaller zones. One zone was horizontal. The other zone was vertical. Two additional defect zones ran horizontally across the top and bottom of the screen, combining three smaller zones in those areas. The final defect zone included the entire video frame.

The detection probability data for these defect zones were fitted using the procedure described in Section 2.2. For the sequences derived from the Amusement Park original sequence, the data did not support a good curve fit for many of the smaller regions. Unlike the problem with the duration sequences, where too many of the subjects detected the weakest error patterns, here the problem was that few people saw the strongest error patterns. In many ways, this problem was worse. Sometimes, only a lower limit for the detection threshold could be estimated. Because few people detected these artifacts, the annoyance values for the sequences were based on fewer observations. Also, this problem cannot be easily fixed in future experiments. The contrast of the blocky/blurry error patterns used in this experiment cannot be increased much beyond the existing strengths without causing saturation problems. Therefore, there is no good way to make the smaller zones more visible in future experiments.

Most of the data derived from the other two original sequences – Football and Susie – had a detection probability range large enough for good curve fits. From the previous experiments, we expected the detection threshold to be higher for the error regions further away from the center. However, the data did not support this conclusion. The detection threshold varied greatly from zone to zone and with no clear pattern. Context effects such as masking or content were much stronger than any variation simply due to position.

Based on the duration results, we expected the threshold to increase as the size of the spatial region increased. The data did not support this conclusion in its most general form. The variation in threshold for zones in different locations was large enough to swamp out any trend due to variation in the size of the zones. However, if only zones that were co-located were considered, the threshold did increase as the size of the spatial region increased. Figure 12 contains a plot of the detection threshold versus normalized defect zone area for all of the defect zones that were centered in the frame. For the Football sequences, there was a strong correlation between the threshold and the defect zone size (0.96). For the other sequences, the correlation was weaker, but still high. If the threshold were constant as size changed, this would indicate perfect integration of error over space; if the threshold increased in proportion to size, this would indicate no integration of error over space. Clearly, the result was in-between these two extremes.

The mean annoyance value results were similar to the threshold results. From the previous experiments, we expected the curves to shift the same way that the thresholds shifted. For size, this implied that the curves would shift to the right as the defect zone size increased. In fact, the curves did shift in the manner predicted by the thresholds. For the defect zones that were co-located around the center, the annoyance curves did uniformly shift to the right. Figure 13 shows a plot of annoyance curve position, as represented by the $\log_{10}$TSE predicted for an annoyance value of ten, versus the detection thresholds for every curve where the detection threshold and annoyance values were successfully fitted. Three data points (two from Amusement Park and one from Susie) were excluded because the detection threshold could not be estimated from the data. The correlation between low annoyance position and threshold was 0.92.

From the duration experiment results, we expected the slope of the mean annoyance values curves to increase as the size of the error region increased. In fact, the data support this hypothesis, although the change in slope is very small compared to

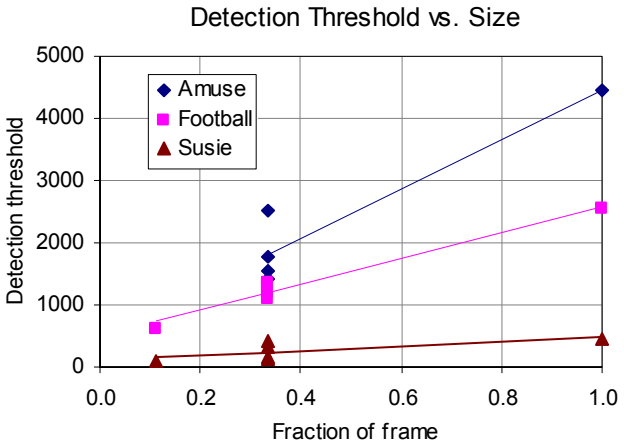

Figure 12. Detection threshold TSE as a function of error region size.

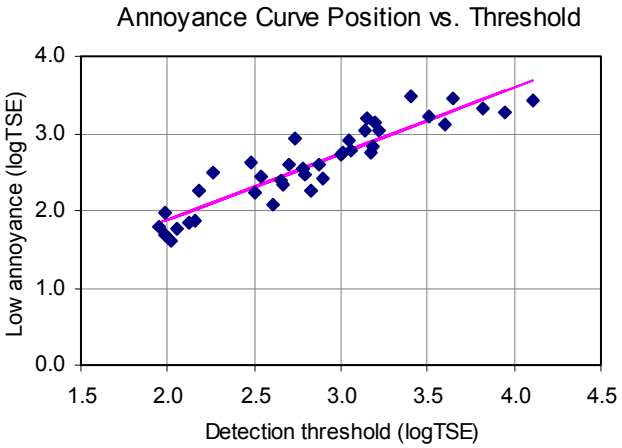

Figure 13. Mean annoyance curve position for an annoyance of ten versus the estimated detection thresholds.

the variation in slope between artifacts in different locations. This was true even for error regions that were co-located around the center.

Among the small regions, both the curve location and the slope of the annoyance value curves varied greatly from region to region and with no clear pattern. It was impossible to predict either value simply based on the error region position. Context effects such as masking or content were much stronger than any variation simply due to position.

4. CONCLUSIONS

The main goal of these experiments was to examine the effect of error pattern type, duration, and size in our stimuli on the detection thresholds and mean annoyance value functions. Our ultimate goal is to be able to predict the detection thresholds and the annoyance values directly from the stimuli. We cannot yet do that. However, from these experiments we have formed the following conclusions:

- When nothing but the type of error pattern is changed, the detection threshold is changed by an approximately constant factor (Eq. 5).
- The $\log_{10}$TSE at the midpoint of the annoyance function is linearly related to the threshold (Eq. 7). Therefore, the effect of changing error pattern types is to shift the annoyance function by an approximately constant factor (Eq. 6).
- Threshold increases with duration. Threshold increases with size when the defect zones are centered on the same point. Consequently, the effect of increasing duration or size is to shift the annoyance function to higher values of $\log_{10}$TSE. These increases are not proportional to duration or size.
- Context has a large effect on thresholds and consequently on annoyance. It affects the form as well as the midpoint location of the detection and annoyance functions. The effect of context remains to be understood.

ACKNOWLEDGMENTS

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

REFERENCES

1. S. Daly, "A visual model for optimizing the design of image processing algorithms," in *Proceedings of 1st International Conference on Image Processing*, vol. 2, pp. 16-20, Austin, TX, USA, 1994.
2. J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision models for target detection and recognition*, E. Peli, World Scientific Publishing, Singapore, 1995.
3. A. J. Ahumada, Jr., "Simplified vision models for image-quality assessment," in *Proceedings of Digest of Technical Papers. First Edition Proceedings of SID '96*, pp. 397-400, San Diego, CA, USA, 1996.
4. C. J. Van den Branden Lambrecht and J. E. Farrell, "Perceptual quality metric for digitally coded color images," in *Proceedings of Proceedings of EUSIPCO-96*, vol. 2, pp. 1175-8, Trieste, Italy, 1996.
5. A. B. Watson, "Towards a Visual Quality Metric for Digital Video," in *Proceedings of European Signal Processing Conference*, vol. 2, Island of Rhodes, Greece, 1998.

6. S. Winkler, "A perceptual distortion metric for digital color images," in *Proceedings of Proceedings of IPCIP'98 International Conference on Image Processing*, vol. 3, pp. 399-403, Chicago, IL, USA, 1998.
7. E. M. Yeh, A. C. Kokaram, and N. G. Kingsbury, "Psychovisual Measurement and Distortion Metrics for Image Sequences," in *Proceedings of European Signal Processing Conference*, vol. 2, pp. 1061-1064, Island of Rhodes, Greece, 1998.
8. K. T. Tan and M. Ghanbari, "A multi-metric objective picture-quality measurement model for MPEG video*," IEEE Transactions on Circuits and Systems for Video Technology* **10**, pp. 1208-13, 2000.
9. K. T. Tan and M. Ghanbari, "Blockiness detection for MPEG2-coded video," *IEEE Signal Processing Letters* **7**, pp. 213-15, 2000.
10. Z. Wang, A. Bovik, and B. Evans, "Blind Measurement of Blocking Artifacts in Images," in *Proceedings of International Conference on Image Processing (ICIP)*, vol. 3, pp. 981-984, Vancouver, Canada, 2000.
11. M. S. Moore, J. M. Foley, and S. K. Mitra, "Detectability and Annoyance Value of MPEG-2 Artifacts Inserted into Uncompressed Video Sequences," in *Proceedings of Human Vision and Electronic Imaging V*, vol. 3959, pp. 99-110, San Jose, CA, 2000.
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 image quality of MPEG-2 stereoscopic sequences," in *Proceedings of Human Vision and Electronic Imaging II*, vol. 3016, pp. 296-301, San Jose, CA, USA, 1997.
14. C. C. Taylor, Z. Pizlo, and J. P. Allebach, "Perceptually relevant image fidelity," in *Proceedings of Human Vision and Electronic Imaging III*, vol. 3299, pp. 110-18, San Jose, CA, USA, 1998.
15. R. Hamberg and H. de Ridder, "Continuous assessment of time-varying image quality," in *Proceedings of Human Vision and Electronic Imaging II*, vol. 3016, pp. 248-59, San Jose, CA, USA, 1997.
16. ITU Recommendation P.930, "Principles of a reference impairment system for video," 1996.
17. A. M. Rohaly, et. al., "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," Video Quality Experts Group, March, 2000.