Perceptual Contributions of Blocky, Blurry, and Fuzzy Impairments to Overall Annoyance

Mylène C.Q. Farias [a], Michael S. Moore [a], John M. Foley [b], and Sanjit K. Mitra [a*]

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

ABSTRACT

In this work, we used two types of impairments in a psychophysical experiment to measure the overall annoyance and individual strength of three impairment features (fuzzy, blocky, and blurry). The impairments were generated by compressing the original videos with MPEG-2 at two different bitrates: 1.0 and 7.5 Mbps. The heavily compressed videos presented blurry and blocky impairments, while the lightly compressed videos presented 'fuzzy' impairments, using a word provided by our test subjects. These impairments were then linearly combined in different proportions and strengths, generating videos in which all three impairment features are present. Our goal was to determine how these impairment features combine to produce the overall annoyance. A modified Minkowski metric was used to describe the 'combination rule' which relates the strengths of the impairment features to the overall annoyance. For the data set containing all test sequences, the optimal value found for the Minkowski parameter *p* was 1.55. From the data obtained, we also estimated the psychometric and annoyance functions. We found that for blocky-blurry and fuzzy artifacts there is no consistent difference between either the thresholds or mid-annoyance strengths.

Keywords: artifacts, video quality, video, MPEG, compression.

1. INTRODUCTION

An impairment is defined as being a perceived flaw introduced into an image or video during capture, transmission, storage, and/or display, as well as by any image processing algorithm that may be applied along the way (e.g. enhancement, compression, etc.). Impairments can be decomposed into a set of features (perceptual components). Although most of the impairments have more than one feature, it is possible to produce impairments that are relatively pure (artifacts). Digital systems, for example, are known to introduce impairments, which can be very complex in their perceptual description. Examples of impairment features are blurriness, noisiness, ringing, and blockiness.[1] Many video quality models have been proposed, but little work has been done on studying and characterizing the individual artifacts found in digital video applications.[2-5] Psychophysical scaling experiments have shown that the overall annoyance of impairments increases when different artifacts are combined simultaneously.[6] A study of the individual perceived artifacts is necessary since we do not yet have a good understanding of how artifacts depend on the physical properties of the video and how they combine to produce the overall annoyance.

In two previous experiments, test subjects were asked to detect and judge localized impairments present in MPEG-compressed videos.[6] The test sequences used in these experiments were generated by compressing the originals using an MPEG-2 codec with specific bitrate goals. Two of the bitrate goals used in these experiments were 1.0 and 7.5 megabits per second (Mbps), respectively. In these cases, the resulting videos contained impairments that looked different and were des-cribed differently by the test subjects. Highly compressed videos at the extreme bitrate (1.0 Mbps) were very blurry and had visible blocks. We called these errors 'blocky-blurry' impairments. The relaxed bitrate goal (7.5 Mbps) resulted in videos with impairments mainly due to quantization noise.[1] We called these errors 'fuzzy' impairments, using the word provided by our test subjects. In this work, we conducted a study of these two types of impairments in a psychophysical experiment to measure the overall annoyance and individual strength of the three impairment features encountered in MPEG-2 compressed videos (fuzzy, blocky, and blurry) described above. We have also studied how these three impairment features combine to produce the overall annoyance.

[*] Further author information: (Send correspondence to M.C.Q.F.)
M.C.Q.F.: E-mail: mylene@ece.ucsb.edu, M.S.M.: E-mail: msmoore@ece.ucsb.edu,
J.M.F.: E-mail: foley@psych.ucsb.edu, S.K.M.: E-mail: mitra@ece.ucsb.edu.

Human Vision and Electronic Imaging IX, edited by Bernice E. Rogowitz,
Thrasyvoulos N. Pappas, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 5292 © 2004 SPIE and IS&T · 0277-786X/04/$15

This paper is divided as follows. In Section 2 the test sequences generation is described. In Section 3 the details of the psychophysical experiment are given. In Section 4 the data analysis of the gathered data is presented. Finally, in Section 5 the conclusions are drawn.

2. TEST SEQUENCE GENERATION

To generate the test video sequences, we first selected a set of five original video sequences of assumed high quality. Since only around 100 sequences can be shown to the test subjects during a forty-minute session, the number of originals was limited to four five-second videos: Bus, Cheerleader, Flower-garden, and Hockey. These videos are commonly used for video experiments and are publicly available at the Video Quality Experts Group website (http://www.vqeg.org).

The normal approach to subjective video quality testing is to degrade the entire video by a variable amount and to ask test subjects for a quality rating.[7] In this research we have been using an experimental paradigm that measures the annoyance value of brief, spatially limited defects inserted in the video.[6] A *defect zone* is defined to be the spatial and temporal region of the video where a defect is inserted. The rest of the video clip is left in its original state. The degradations are generated separately and added to specific spatial and temporal regions of the original videos. The test subjects are then asked to search each video clip for defective regions and to indicate the annoyance value or impairment strengths of the features in the defects seen.

The regions used in this experiment were created by dividing the frame into three equal strips, either horizontally or vertically. They were 1 second long and did not occur during the first and last seconds of the video. Different regions were used for each original to prevent the test subjects from learning the locations where the defects appear. Due to time limitations, only two defect zones for each original were used. Columns 1 and 2 of Table 1 present the defect zones and original videos used for this experiment.

The impairments used in this experiment were the same type of MPEG-2 impairments of two previous experiments.[6] The test sequences were generated in the following way. First, we took an original sequence and compressed it using an MPEG-2 codec with two specific bitrate goals: 1.0 Mbps and 7.5 Mbps. The types of defects and their appearance in the resulting reconstructed videos differed with the different bitrate goals. The extreme bitrate goal resulted in videos that were very blurry and had visible blocks. We called these defects **Type I** impairments or ‘blocky-blurry’ impairments. The relaxed bitrate goal resulted in videos degraded mostly due to quantization noise. To use the word provided by our test subjects, the sequences looked ‘fuzzy’. We will call these defects **Type II** impairments or ‘fuzzy’ impairments. In general, the raw Type II impairments were significantly weaker than the raw Type I impairments.

To create our stimuli we started with three sets of videos. One set contains the original videos. The second set contains the reconstructed videos with mostly Type I impairments. The third set contains the reconstructed videos with mostly Type II impairments. Most impairment subjective tests vary the strength of the impairments in the test signals by varying the bitrate goal and/or the codec used to compress the original video.[7] This approach changes both the strength and the type of impairments. We were interested in varying only the strength of the impairments – not their type. To do that, we linearly combined the original video with a video with impairments. The basic formula is

$$Y = X_0 + r\left(X_1 - X_0\right), \tag{1}$$

where X_0 is the original, X_1 is the video with impairments, *r* is the weighting factor and *Y* is the resulting sequence. We also created test sequences that combined Type I and Type II impairments. This was done by combining three sequences instead of two. Specifically,

$$Y = X_0 + a\left(X_1 - X_0\right) + b\left(X_2 - X_0\right), \tag{2}$$

where X_0 is the original video, X_1 and X_2 are videos with Type I and Type II impairments, respectively, and *a* and *b* were the two weighting factors. The total squared error (TSE) of Y (Eq.(1)) is given by:

$$TSE = a^2 \sum \left(X_1 - X_0\right)^2 + b^2 \sum \left(X_2 - X_0\right)^2 + 2ab \sum \left(X_1 - X_0\right)\left(X_2 - X_0\right) = a^2 TSE_1 + b^2 TSE_2 + 2ab TSE_{12}\,, \tag{3}$$

where the value of TSE_1 and TSE_2 are simply the TSE values calculated for (a = 1, b = 0) and (a = 0, b = 1), respectively. The value of TSE_{12} can be calculated by creating a combined version of the video (non-zero *a* and *b*) and

Table 1. Summary of original sequences, defect zones, strengths and weighting factors (*a*, *b*) used in the experiment.

Original	Defect zone	Strength	*a*				*b*			
			I	II	66%I	66%II	I	II	66%I	66%II
Bus	Top.	1	1.00	0.00	0.64	0.45	0.00	1.30	0.59	0.83
		2	0.79	0.00	0.50	0.36	0.00	1.03	0.46	0.66
		3	0.51	0.00	0.32	0.23	0.00	0.67	0.30	0.42
	Middle	1	1.00	0.00	0.66	0.46	0.00	1.48	0.69	0.97
		2	0.79	0.00	0.52	0.37	0.00	1.17	0.54	0.77
		3	0.49	0.00	0.32	0.23	0.00	0.73	0.34	0.48
Flower	Houses	1	1.00	0.00	0.67	0.48	0.00	1.82	0.86	1.22
		2	0.80	0.00	0.54	0.38	0.00	1.45	0.69	0.97
		3	0.52	0.00	0.35	0.25	0.00	0.95	0.45	0.64
	Garden.	1	1.00	0.00	0.68	0.48	0.00	1.66	0.80	1.12
		2	0.79	0.00	0.54	0.38	0.00	1.32	0.63	0.89
		3	0.51	0.00	0.34	0.24	0.00	0.84	0.40	0.57
Football	Left	1	1.00	0.00	0.69	0.49	0.00	1.82	0.89	1.25
		2	0.78	0.00	0.54	0.38	0.00	1.42	0.69	0.98
		3.	0.48	0.00	0.33	0.23	0.00	0.86	0.42	0.60
	Middle	1	1.00	0.00	0.69	0.49	0.00	1.77	0.86	1.22
		2	0.80	0.00	0.55	0.39	0.00	1.41	0.69	0.97
		3	0.52	0.00	0.35	0.25	0.00	0.91	0.44	0.63
Hockey	Middle	1	1.00	0.00	0.71	0.50	0.00	2.15	1.08	1.52
		2	0.79	0.00	0.56	0.39	0.00	1.69	0.85	1.20
		3	0.49	0.00	0.34	0.24	0.00	1.04	0.52	0.74
	Right	1	1.00	0.00	0.64	0.45	0.00	1.11	0.50	0.71
		2	0.85	0.00	0.55	0.39	0.00	0.95	0.43	0.61
		3	0.68	0.00	0.43	0.31	0.00	0.75	0.34	0.48

measuring its TSE:

$$TSE_{12} = \frac{TSE_{combined} - a^2 TSE_1 - b^2 TSE_2}{2ab}. \tag{4}$$

Using these three values, we can calculate values for *a* and *b* for any arbitrary TSE and 'mixture' of Type I and Type II impairments. If the value of TSE$_{12}$ is very small (impairments are relatively independent), then a proportion variable *p* can be defined:

$$p = \frac{b^2 TSE_2}{a^2 TSE_1} = \frac{\text{TSE from Type II}}{\text{TSE from Type I}}. \tag{5}$$

From this relationship, expressions for *a* and *b* are derived:

$$a = \sqrt{\frac{TSE_{goal}}{(1+p)TSE_1 + 2TSE_{12}\sqrt{p\,TSE_1/TSE_2}}},\ b = a\sqrt{\frac{pTSE_1}{TSE_2}}. \tag{6}$$

The twelve pairs of (*a*,*b*) used in this experiment are shown in columns 4 and 5 of Table 1. Each pair corresponds to a choice of original, defect zone, strength, and impairment proportion. The constants *a* and *b* were selected in order to create three different strength levels – weak (3), medium (2), and strong (1) - of 'constant' TSE. Each line of values for *a* and *b* correspond to a strength level. Each column of *a* and *b* correspond to a different proportion of impairments ('type I', 'type 66%I', 'type 66%II', and 'type II'). Therefore, from these twelve pairs, three pairs corresponded to sequences with only Type I impairment (*a* varies, *b* = 0), three to sequences with only Type II impairment (*a* = 0, *b* varies), and six pairs to sequences with combined impairments ('type 66%I', 'type 66%II').

3. METHOD

The entire apparatus consisted of the following components: a computer, a broadcast video monitor, a computer monitor, a keyboard, and a mouse. The test video sequences are stored on the hard disk of an NEC server (PC computer). The videos are displayed using a subset of the PC cards normally provided with the Tektronix PQA-200 picture quality analyzer. The test sequence length is limited to five seconds by the generator card. The analog output is then displayed on a Sony PVM-1343 monitor (14 inches). Each test sequence can be loaded and displayed in six to eight seconds.

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 concerning video artifacts and the associated terminology. The experiments were run with one subject at a time. The subject was seated straight ahead of the monitor, located at or slightly below eye height for most subjects, with the keyboard and mouse in easy reach. The distance between the subject's eyes and the monitor was of four video monitor screen heights from the video monitor. The video monitor was 20 cm tall resulting in a distance view of 80 cm.

The test subjects were divided into two groups – ‘Annoyance’ and ‘Feature’. Each group performed one of the two experimental tasks. The task performed by ‘Annoyance’ group consisted of reporting annoyance values. The task performed by the ‘Feature’ group consisted of indicating impairment strengths. However, the general flow of the experiment was the same for both groups. The course of each experimental session goes through five stages: instructions, training, practice trials, experimental trials, and interview. In the first stage, the subject is verbally given instructions.

In the training stage, sample sequences are shown to the subject. The sample sequences had two important functions. First, the sequences represented the impairment extremes for the experiment, so that annoyance value range could be established prior to the start of the experiment. Second, the sequences taught the test subject to recognize each of the three impairment features: blurriness, blockiness, and fuzziness. The training stage varied according to the group of subjects.

Subjects in the ‘Annoyance’ group watched two sets of sample sequences. The first set consisted of original sequences. The second set consisted of sequences with the worst defects, i.e., the sequences with the highest TSE. Since the annoyance of a defect may depend also on the original video, the defect zone or the type of defect, we instructed the subjects to assign a value of ‘100’ to the defects they think are the worst in this set. Subjects in the ‘Feature’ group watched four sets of sample sequences. The first set consisted of the original. The second, third and fourth sets consisted of the worst defects with mostly one impairment feature. Before each set, the experimenter explained the type of impairment feature to be displayed.

To create a set of sequences with mostly fuzziness, we simply used a Type II impaired video. Some of the type II impairments look fuzzy without any apparent blurring or blocking. However, the Type I videos can not easily be categorized as only blurry or only blocky. Therefore, the videos with mostly blockiness and mostly blurriness had to be generated in another way. Blurring is a reduction in the sharpness of edges and a loss of spatial detail. Sequences with blurriness were created artificially by running a lowpass filter across each field in a video. Blockiness is a little harder to simulate. It is defined as a ‘distortion of the image characterized by the appearance of an underlying block encoding structure.’ Basically, the boundaries between blocks become visible while details within the blocks are lost. Because details are lost as both blurriness and blockiness increase, it is hard to simulate blockiness without also including blurriness. So, we showed a set consisting of those videos in which blockiness was most apparent. The test subjects were then informed that one type of impairment dominates each training set, but that the other impairments may also be present.

The practice trials are identical to the experimental trials, except that no data is recorded. The practice trials are used to familiarize the subject with the experiment. Twelve practice trials were included in this experiment to allow the subjects’ responses to stabilize before the experimental trials begin. In the experimental trial the set of all test sequences are shown in random order. After each video has finished, a dialog box appeared on an adjacent computer monitor. The content of the dialog box depended on the data being requested (annoyance values or impairment strengths).

A subject in the ‘Annoyance’ group was first asked if a defect or impairment was seen. If the answer was yes, the subject was asked to enter a numerical value for the annoyance using the keyboard. The subject was instructed to enter a positive numerical value indicating how annoying the defect was compared to the worst defects in the training stage. Any defect half as annoying as should be given ‘100’, half as annoying ‘50’, twice as annoying ‘200’ and so forth. If the answer was no, no number needed to be entered. After all of the data was entered, the subject was able to proceed to the next sequence by clicking on Next or simply hitting the Enter key.

A subject in the 'Feature' group was asked to rate the strength of each kind of impairment using three scale bars. Each bar was labeled with an eleven-point scale (0 – 10). However, each bar contained far more than eleven points and intermediate values were allowed. The subject entered the scores by using the mouse to click on each scale. The scale bars were updated to show the entered strength. Until the dialog box was closed, the entered values could be adjusted by re-clicking on the scale bars. After the impairment strengths were entered, the user would proceed to the next sequence by clicking on Next or hitting the Enter key. For the 'Feature' group, the subject was never explicitly asked if a defect was seen. Instead, all three of the scale bars were initialized to zero. If the subject clicked on Next without changing any of the impairment strengths, we assumed that no defect was seen

At the end of the experimental trials, we asked the test subjects for qualitative descriptions of the defects that were seen. The qualitative descriptions are useful for categorizing the defect features seen in each experiment and help in the design of future defect feature analysis experiments.

3. DATA ANALYSIS

We used standard methods[7] for analyzing the annoyance and impairment strength judgments provided by the test subjects. We first computed the Total Squared Error (TSE) of the test sequences. Then we calculated the mean observer score (MOS) of each of the responses given by the subjects by averaging the scores over all subjects, using the following equation:

$$MOS_j = 1/N \sum_{i=1}^{N} S(i,j) \quad , \qquad (7)$$

where $S(i,j)$ is the score value reported by the i-th subject for the j-th test sequence, and N is the total number of subjects.

The data gathered from subjects in the 'Annoyance' group provided two measures for each test sequence: the detection probability and the Mean Annoyance Value (*MAV*). We estimated the detection probability by counting the number of subjects who detected the impairment and dividing it by the total number of subjects. Then, the relation between detection probability and total squared error (TSE) was fitted using the Weibull function,[7] which has an *S*-shape similar to our data and is defined as

$$P(x) = 1 - 2^{-(S \cdot x)^k} \quad , \qquad (8)$$

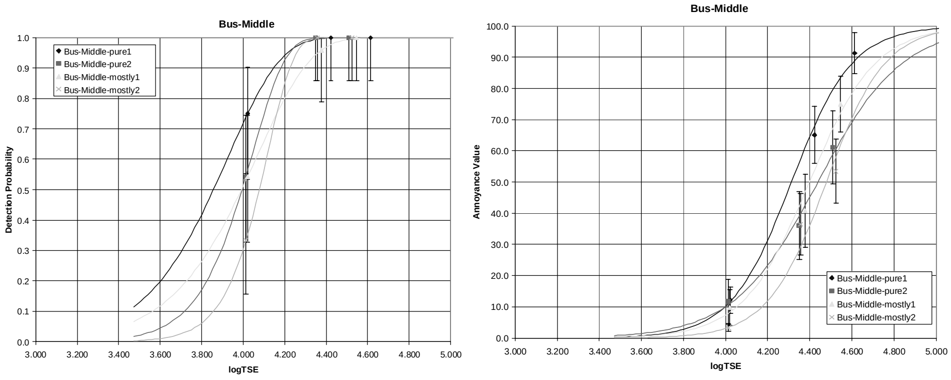
where $P(x)$ is the probability of detection, x is the logarithm of the TSE of the corresponding test sequence, $X_T = 1/S$ is the 50% detection threshold in logarithmic error energy, and k is a constant that determines the steepness of the function. The resulting curves are called psychometric functions.

The *MAV* is calculated by estimating the *MOS* of the annoyance values, i.e., by averaging the annoyance scores over all observers. These *MAV* values were then fitted with a standard logistic function:[7]

$$PMAV = y_{\min} + \left(y_{\max} - y_{\min}\right) \Big/ \left(1 + \exp\left(-\frac{(x - X_{mean})}{|\beta|}\right)\right), \qquad (9)$$

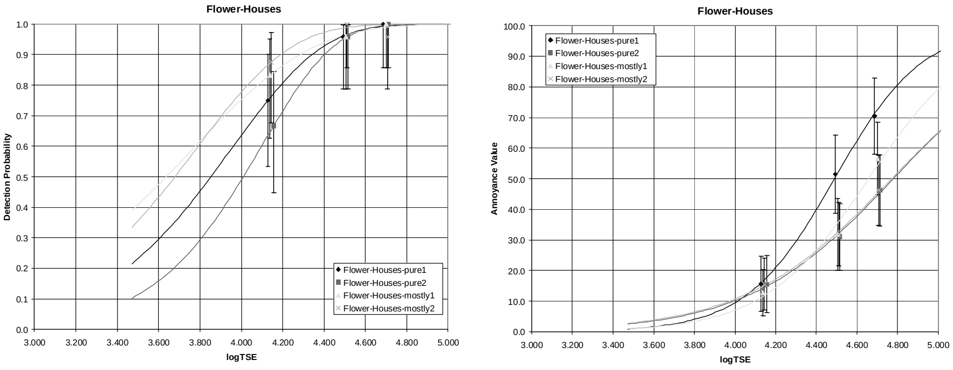
where $PMAV$ is the predicted mean annoyance value and x is the $\log_{10}$ (TSE) of the corresponding test sequence. The parameters $y_{\max}$ and $y_{\min}$ establish the limits of the annoyance value range. The parameter X_{mean} translates the curve in the x-direction and the parameter β is inversely related to the steepness of the curve. The resulting curves are called annoyance functions. Figures 1-3 depict the psychometric and annoyance functions for a subset of the test sequences. Table 2 summarizes the fitting parameters of the psychometric and annoyance functions for all the groups of test sequences (same original, same defect zone). The last column of Table 2 shows the detection threshold (X_T) values for each group.

To compare the fitting parameters Xmean, β, *S*, and *k* for both the impairments, we plotted type I parameters against the corresponding type II parameters. Figure 4 depicts the graphs for the parameters X_{mean} and β, and Figure 5 for parameters *S* and *k*. The values of X_{mean} for the impairments type I and II (Figure 4(a)) are highly correlated. The



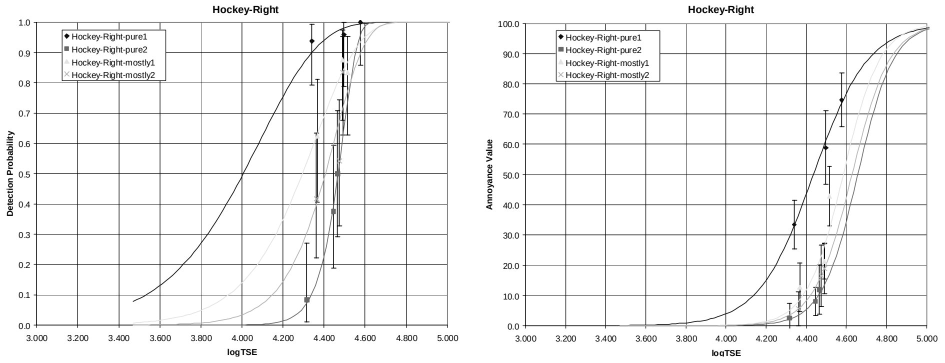
(a) (b)

Figure 1. (a) Annoyance and (b) Psychometric functions for video 'Bus', defect zone 'Middle'.



(a) (b)

Figure 2. (a) Annoyance and (b) Psychometric functions for video 'Flower', defect zone 'Houses'.



(a) (b)

Figure 3. (a) Annoyance and (b) Psychometric functions for video 'Hockey', defect zone 'Right'.

correlation coefficient R^2 is equal to 0.99 and the *P*-value is approximately equal to 0 (the *P*-value is the probability of getting a value of the test statistics as extreme as, or more extreme than, the value observed, if the null hypothesis were true.). The fitted line obtained for these points was X_{mean_I} = (0.99 X_{mean_II} + 0.11). A similar behavior was found for the parameter *S***.** The values of *S* for the impairments type I and II (Figure 5(a)) are also highly correlated (R^2 = 0.95 and *P*-value approximately equal to 0). The fitted line obtained for these points was *S_I* = (1.17 *S_II* - 0.66). For parameters β and *k* (Figure 4(b) and 5(b)) no strong correlation was found.

We also examined the relationship between the detection threshold (X_T) and the mid-annoyance (X_{mean}) for the two types of impairments. Figures 6(a) and 6(b) depict the graphs of X_{mean} against X_T for impairment type I (Figure 6(a)) and type II (Figure 6(b)). The correlation (R^2) found was 0.88 and 0.74 and the *P*-values were 0.0005 and 0.0065,

Table 2. Fitting parameter of psychometric and annoyance functions for Annoyance group.

Sequence	X_{mean}	β	S	k	X_T
Bus-Middle-mostly1	4.043	0.36	0.282159	14.71719	3500.319
Bus-Middle-mostly2	4.343	0.25	0.275448	30.24649	4270.16
Bus-Middle-pure1	3.952	0.14	0.293798	14.87067	2533.358
Bus-Middle-pure2	4.036	0.22	0.282094	23.36706	3506.823
Bus-Top-mostly1	3.859	0.2	0.309138	18.76881	1717.14
Bus-Top-mostly2	3.994	0.16	0.314299	24.25575	1519.445
Bus-Top-pure1	3.663	0.2	0.327649	15.4432	1127.314
Bus-Top-pure2	3.675	0.14	0.326522	17.75661	1154.989
Flower-Garden-mostly1	3.675	0.31	0.254239	15.00	8576.522
Flower-Garden-mostly2	3.894	0.26	0.254433	13.24879	8517.513
Flower-Garden-pure1	4.773	0.24	0.256176	13.2937	8008.726
Flower-Garden-pure2	4.779	0.22	0.25359	15.00	8777.462
Flower-Houses-mostly1	4.673	0.28	0.311917	6.609869	1606.881
Flower-Houses-mostly2	4.836	0.25	0.30786	8.292734	1771.054
Flower-Houses-pure1	4.22	0.36	0.293159	9.037525	2577.031
Flower-Houses-pure2	4.336	0.21	0.279636	11.61388	3767.762
Football-Left-mostly1	3.51	0.16	0.303797	16.95601	1957.357
Football-Left-mostly2	3.707	0.19	0.311217	14.25315	1633.789
Football-Left-pure1	3.839	0.29	0.326086	15.00	1165.937
Football-Left-pure2	3.883	0.13	0.32268	15.00	1256.165
Football-Middle-mostly1	3.392	0.23	0.317691	12.44539	1405.109
Football-Middle-mostly2	3.5	0.25	0.316666	15.00	1438.482
Football-Middle-pure1	3.575	0.17	0.318383	15.00	1383.161
Football-Middle-pure2	3.642	0.22	0.313306	13.09499	1555.13
Hockey-Middle-mostly1	3.958	0.09	0.333399	16.5948	998.6488
Hockey-Middle-mostly2	4.18	0.09	0.353618	11.35783	672.8436
Hockey-Middle-pure1	3.472	0.27	0.359518	10.92489	604.6515
Hockey-Middle-pure2	3.528	0.14	0.362485	10.09107	573.7684
Hockey-Right-mostly1	4.375	0	0.262125	18.9976	6531.003
Hockey-Right-mostly2	-	-	0.254598	27.54701	8467.48
Hockey-Right-pure1	4.106	0.1	0.280504	15.00	3672.901
Hockey-Right-pure2	4.153	0.55	0.250575	56.0907	9790.835

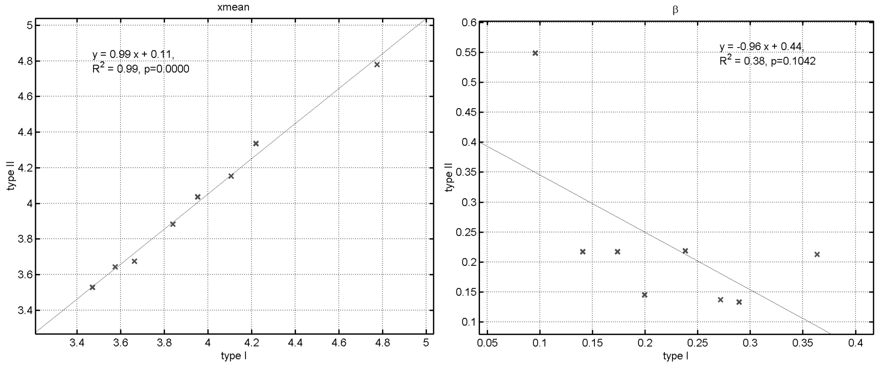
Table 3: *P* values obtained from the ANOVA analysis on the annoyance and visibility fitting parameters (Table 2).

Annoyance	$\bar{x}$		β	
	original	impairment	original	impairment
P	0.007	0.6982	0.9677	0.5623
Psychometric	*S*		*k*	
	original	impairment	original	impairment
P	0.0275	0.7857	0.1849	0.4505

respectively. The relationships found by fitting a line to the data were X_{mean_I} = (1.12 X_{T_I} + 0.26) and X_{mean_II} = (0.81 X_{T_II} + 1.25). Moore et al. found similar results for fuzzy and blocky-blurry impairments.[4]

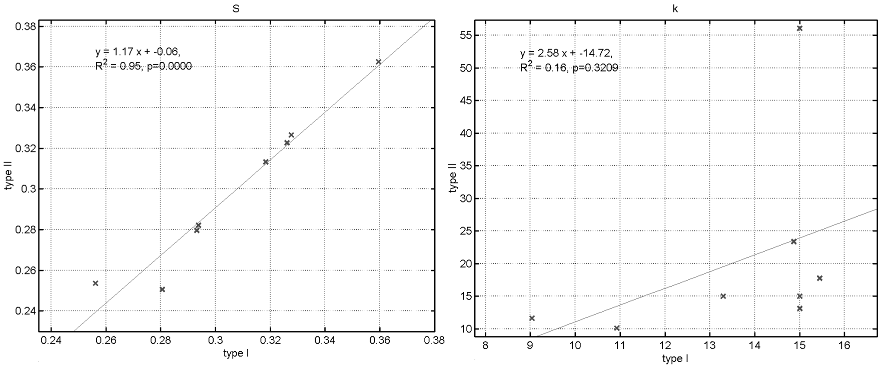
To analyze the effect of the 'impairment proportion' and the 'original' over the parameters of the annoyance and psychometric functions (X_{mean}, β, S, and k) we performed an ANOVA test. The four 'impairment proportions' considered were 'type I', 'type 66%I', 'type 66%II', and 'type II'. The *P*-values obtained are shown in Table 3. The results show that the 'impairment proportion' does not have a significant effect on any of the parameters. The original video had a significant effect on two of the parameters: X_{mean} and S. These results are in agreement with the results presented by Moore et al.[4]

The data gathered from subjects in the 'Feature' group provided 3 *MOS* values for each test sequence. These values corresponded to the Mean Strength Values (*MSV*) of blockiness, blurriness and fuzziness (impairment features). It is



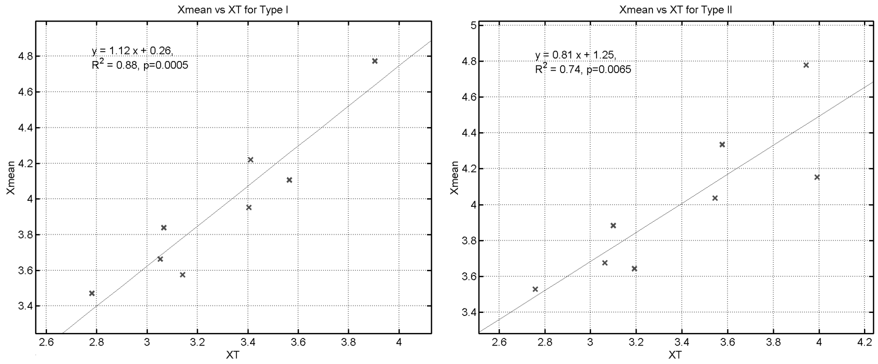
(a) (b)

Figure 4. (a) X_{mean} and (b) β parameters for impairments type I and II.



(a) (b)

Figure 5. (a) *S* and (b) *k* parameters for impairments type I and II.

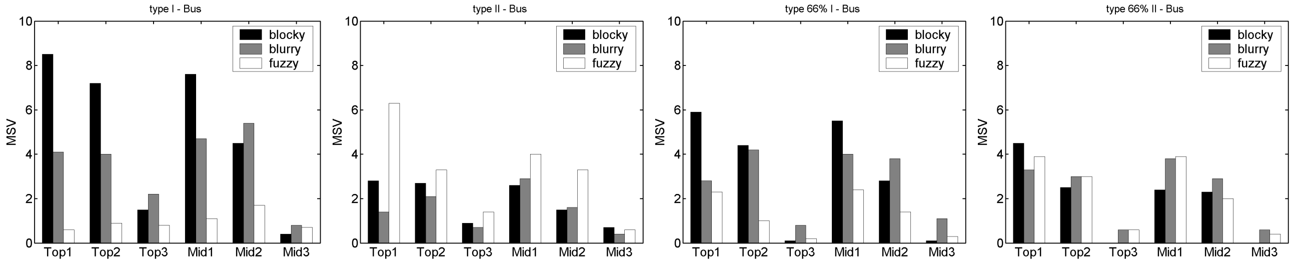


(a) (b)

Figure 6. X_{mean} versus X_T for impairments type (a) I and (b) II.

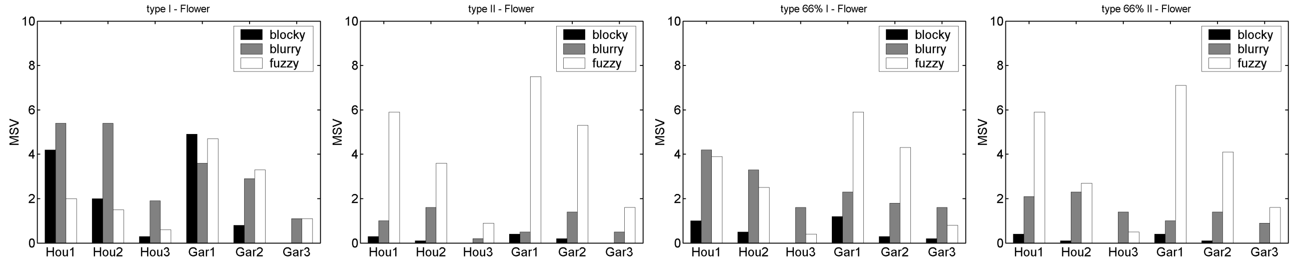
important to remember that the range of *MAV* values is from 0-100, while the range of all the *MSVs* values is from 0-10.Figures 7-9 depict the bar plots of the MSV values obtained for blockiness, blurriness and fuzziness for three of the originals. Each graph corresponds to one original and two defect zones. The labels of the x-axis correspond to the defect zone (Top, Bottom, Middle, Left, and Right) and strength (strong = 1 , medium = 2, and weak = 3) of the test sequence. For example, Mid1 corresponds to a strong defect in the Middle defect zone (Table 1).

As it can be noticed from these graphs, subjects thought that impairments type I (blocky-blurry) were also fuzzy. Impairments type II (fuzzy) were also judged as blocky and blurry. As expected, the combined impairments types ('type 66%I' and 'type 66%II') presented a combination of the three features. For type I test sequences the *MSV*s for blocky and blurry were generally greater than the *MSV* for fuzzy. Nevertheless, the original seemed to have an important



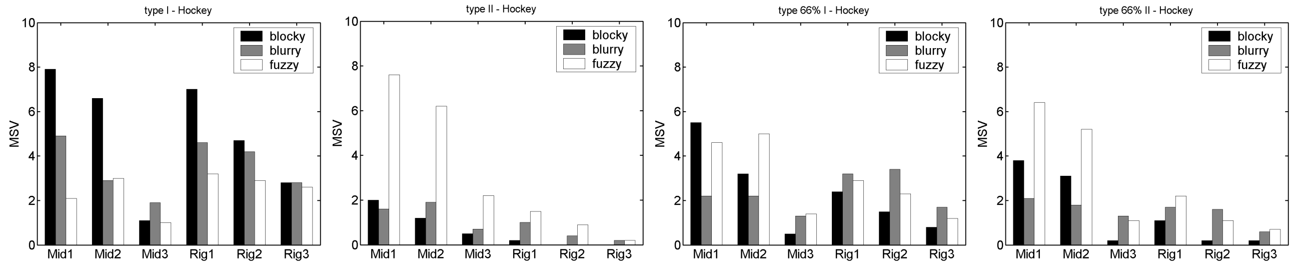
(a) (b) (c) (d)

Figure 7. MSV bar plots for impairments types (a) I , (b) II, (c) 66% I and, (d) 66%II for video Bus (Top and Middle).



(a) (b) (c) (d)

Figure 8. MSV bar plots for impairments types (a) I , (b) II, (c) 66% I and, (d) 66%II for video Flower (Houses and Garden).



(a) (b) (c) (d)

Figure 9. MSV bar plots for impairments type (a) I , (b) II, (c) 66% I and, (d) 66%II for video Hockey (Middle and Right).

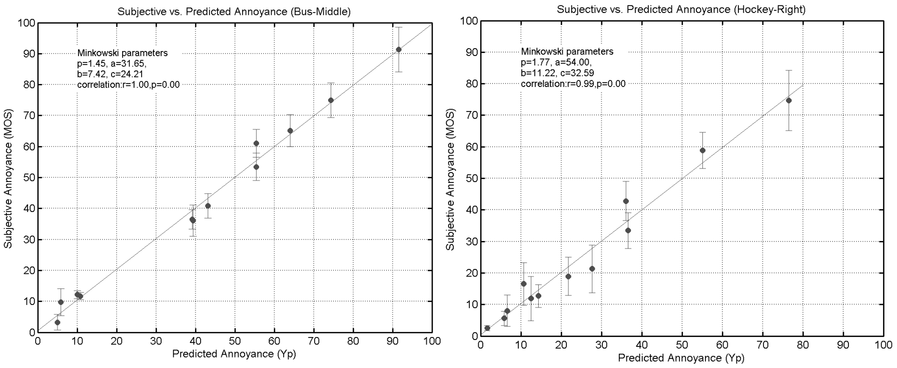
effect on these proportions. For example, the *MSV* for fuzziness for type I - Flower Garden (Figure 8) had values at least as high as the MSV for blockiness and blurriness. For type II, *MSV* for fuzziness were generally higher than the *MSV*s for blockiness and blurriness.

Our principal interest in measuring the impairments' strength is to investigate the relationship between the strength of each type of impairment and the overall annoyance. To analyze this data, we used a Minkowski metric, which is a combination rule for impairments commonly used in human perception research.[5] Using a nonlinear fitting procedure, we fitted the *MSV* of each type of impairment using the Minkowski metric:

$$PMAV_j = \left(a \cdot MSV_{blocky}{}^{p} + b \cdot MSV_{blurry}{}^{p} + c \cdot MSV_{fuzzy}{}^{p}\right)^{1/p}, \tag{10}$$

PMAV is the predicted value for *MAV*, *p* is the Minkowski exponent, and *a*, *b*, and *c* are the Minkowski coefficients. The parameters *a*, *b*, *c*, and *p* are positive real numbers. Table 4 summarizes the results obtained for this fit. Column 8 of Table 4 shows the squared sum of the fitting residuals and column 9 shows the correlation coefficient *R*. Figure 10 depicts the plots of the *MAV* versus *PMAV* for the test sequences Bus (Middle) and Hockey (Right). The graphs displays the *PMAV* values corresponding to each *MAV* and their respective confidence intervals. The line in the graphs correspond to *MAV* = *PMAV* ($y = x$). The fit is reasonably good for this type of data ($R^2 \geq$ 0.98 and the *P* values were approximately 0 for all groups).

The estimated coefficients in Table 4 show how the 'blocky', 'blurry', and 'fuzzy' *MSV*s contributed to the overall annoyance. A limitation of this model is that the values obtained for *a*, *b*, and *c* varied for some test sequence groups. Nevertheless, it is evident in Table 4 that the value of *a* (*MSV* blocky) is almost always greater than *b* (*MSV* blurry) and



(a) (b)

Figure 10. Subjective vs. Predicted Annoyance using Minkowski metric for videos 'Bus-Middle' and 'Hockey-Right'.

c (*MSV* fuzzy). Two scenes had a unique "behavior": Flower-Garden and Football-Left. The parameters *p*, *a*, *b*, and *c*for the sequence 'Flower-Garden' are very low values when compared to the other sequences. For the sequence 'Football-Left', on the other hand, the parameters obtained were extremely high.

We also fitted the *MSV*s values using the Minkowski metric to the data set containing all test sequences. Figure 12 (a) depicts the plot of the *MAV* versus *PMAV* obtained for this case and their respective confidence intervals. The results of this fit are shown in the last row of Table 4. The overall fit resulted in p = 1.55 and *a*, *b*, and c very close together and around 22. The correlation of this fit was 0.95 and the *P*-value was approximately 0. A Minkowski exponent value of 1.55 is in the range of values found in previous works.[5] Based on this overall fit, we decided to fix the value of p in Eq. (10) to 1.5 and make another nonlinear fit of the MSVs values using the Minkowski metric. Having a single value for p simplifies the relationship among PMAV and the MSVs given to the impairment features.

Columns 2-6 of Table 4 summarize the results obtained for this second fit. Figures 11(a) and 11(b) depict the plots of the *MAV* versus *PMAV* for the same subset of test sequences and p = 1.50. The results obtained for this fit are very similar to the previous one. The fit is also reasonably good ($R^2 \geq 0.98$ and the *P*-values were approximately 0 for all groups). The sum of the squared residuals for each group is practically the same as the ones of the previous fit, except for the sequences Flower-Garden and Football-Left that resulted in higher sums. From Table 5 it is also interesting to notice that the new parameters for the sequence 'Flower-Garden' are in the same range of the parameters for the other sequences. Figure 12 (b) depicts the plot of the *MAV* versus *PMAV* obtained using the Minkowski metric with fixed p = 1.5 for the set containing all test sequences. The plots also display the confidence intervals of the fit. The Minkowski parameters *a*, *b*, and *c* obtained for the overall case are again very close together and around 20 (last row of Table 5). The correlation was 0.95 and the *P*-value was approximately 0. A model comparison test[8] was done between the more generic model (Minkowski metric with p free) and this new model. The results indicate that there is no significant difference in performance between theses two models.

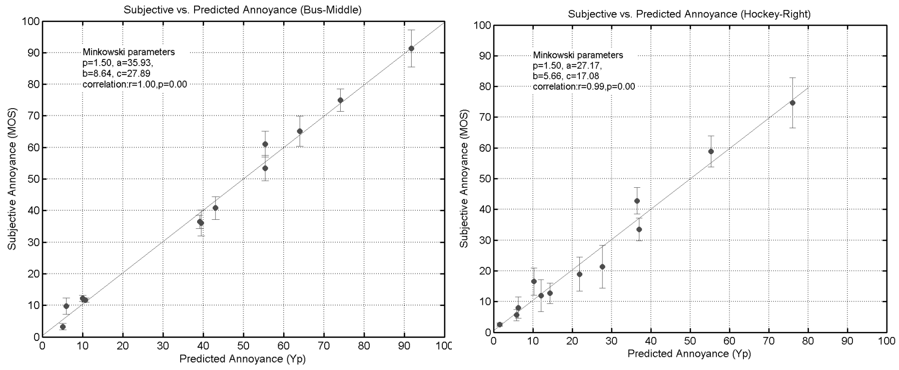
In a previous work that used synthetic artifacts[9] (blockiness, blurriness and noiseness) we found that the *PMAV* could be estimated by a simple linear combination of the impairment MSVs. For the purpose of comparison, we have also fitted our data using a simple linear combination metric, i.e., by fixing p = 1 in Eq. (10). Figure 13 depicts the plots of the MAV versus *PMAV* for the test sequences Bus and Hockey. Again, the results obtained for this fit are similar to the previous two fits. The sum of the squared residuals for is slightly higher, but this fit is also good ($R^2 \geq 0.97$ and the *P*-values were approximately 0 for all groups). Again, a model comparison test was done between the more generic model (Minkowski metric with p free) and this new model. The results indicate that the performance of the Minkowski metric is better for the data set containing all the test sequences and for the group Football Left. Since one of the artifacts was different here we would not expect the same weights as in our earlier study. Nevertheless, it is interesting to notice that the coefficients obtained for this work were in general higher than the ones obtained for the previous work.

4. CONCLUSIONS

In this work we used two types of impairments (blocky-blurry and fuzzy) in a psychophysical experiment to measure the overall annoyance and individual strength of three impairment features (blockiness, blurriness, and fuzziness). We estimated the annoyance and psychometric functions for these impairments. The results from an ANOVA test on the data showed that the 'impairment proportion' had no significant effect on any of the parameters of these two functions.

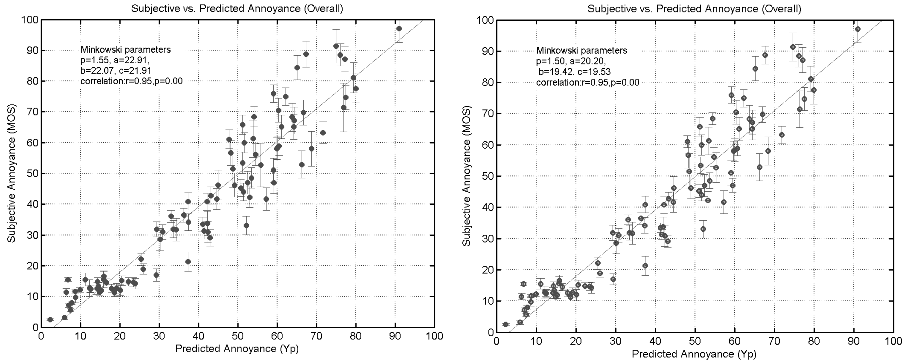
The original video, on the other hand, had a significant effect on the values of X_{mean} and S. The values of X_{mean} (Mid-annoyance) corresponding to impairment type I and the values of X_{mean} corresponding to type II are well correlated and can be related by a linear equation. X_{mean} and X_T values corresponding for each impairment type are also well correlated and related linearly. This means that, if the threshold of impairment is known, its annoyance function can be predicted.

The *MAV* of the test sequences can be estimated using the individual strengths (*MSVs*) of the impairment features. A good fit was obtained by using the Minkowski metric to combine the *MSVs*. We also fitted the data with a constant Minkowski exponent of 1.5 and 1.0 (linear case). Fixing p = 1.5 did not produce a significant worsening of the fit. For the linear case, the fit was significantly worse (P < 0.05) for the data set containing all test sequences and for one of the group of sequences. Nevertheless, in all cases the linear model provides a reasonably good description of the data.



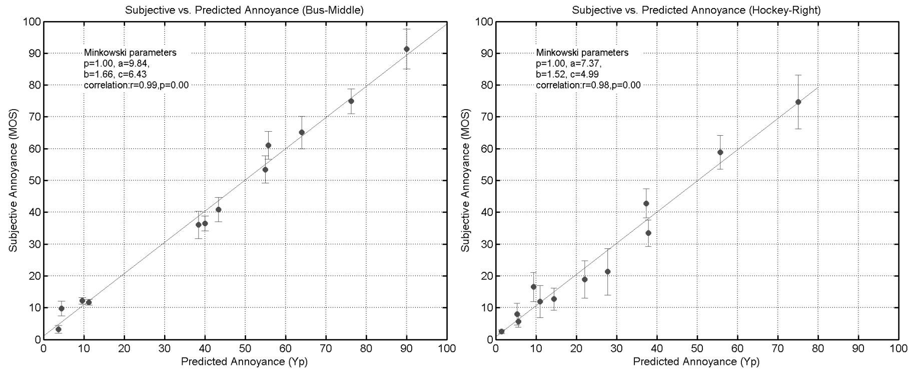
(a) (b)

Figure 11. Subjective vs. Predicted Annoyance using Minkowski metric with p=1.5 for videos ‘Bus-Middle’ and ‘Hockey-Right’.



(a) (b)

Figure 12. Subjective vs. Predicted Annoyance for data set of all videos for (a) Minkowski and (b) Minkowski with p=1.5.



(a) (b)

Figure 13. Subjective vs. Predicted Annoyance using Minkowski metric with p=1.0 for videos ‘Bus-Middle’ and ‘Hockey-Right’.

Table 4. Fitting parameters for Minkowski metric.

Group	max	min	p	*a* blocky	*b* blurry	*C* fuzzy	$\Sigma res(i)^2$	R^2
Bus Top	100	0	1.95	57.28	11.63	31.49	11.29	0.99
Bus Middle	100	0	1.45	31.65	7.42	24.21	9.28	1
Flower Houses	100	0	1.24	14.53	10.62	11.01	12.54	0.98
Flower Garden	100	0	0.28	0.49	0.26	1.41	12.28	0.99
Football Left	100	0	2.57	469.46	0.13	199.18	10.81	0.99
Football Middle	100	0	1.2	17.55	0.98	13.52	19.13	0.98
Hockey Middle	100	0	1.85	47.52	28.57	44.2	18.18	0.98
Hockey Right	100	0	1.77	54	11.22	32.59	12.68	0.99
All	100	0	1.55	22.91	22.07	21.91	77.6	0.95

Table 5. Fitting parameters for Minkowski metric with p = 1.5 and p = 1.0 (linear).

	p =1.5					*p* = 1.0				
Group	*a* blocky	*b* blurry	*c* fuzzy	$\Sigma res(i)^2$	R^2	*a* blocky	*b* blurry	*c* fuzzy	$\Sigma res(i)^2$	R^2
Bus Top	21.71	2.8	10.95	11.55	0.99	7.64	0.15	3.13	12.75	0.99
Bus Middle	35.93	8.64	27.89	9.28	1	9.84	1.66	6.43	9.82	0.99
Flower Houses	31.37	21.13	19.83	12.72	0.98	6.87	5.4	6.22	12.77	0.99
Flower Garden	24.3	19.68	19.82	14.93	0.98	5.3	3.76	6.8	13.81	0.98
Football Left	32.50	0.00	16.19	14.00	0.99	8.87	0.6	4.51	18.54	0.97
Football Middle	40.27	2.61	30.85	19.47	0.98	10.04	0.54	7.75	19.32	0.98
Hockey Middle	19.32	12.26	19.1	18.45	0.98	5.07	3.89	5.62	20.04	0.97
Hockey Right	27.17	5.66	17.08	12.74	0.99	7.37	1.52	4.99	13.24	0.98
All	20.2	19.42	19.53	77.62	0.95	5.25	5.11	5.71	80.97	0.95

ACKNOWLEDGMENTS

This work was supported in part by CAPES-Brazil, in part by a National Science Foundation Grant CCR-0105404, and in part by a University of California MICRO grant with matching support from Philips Research Laboratories.

REFERENCES

1. Michael Yuen, and H.R.Wu, "A Survey of Hybrid MC/DPCM/DCT Video Coding Distortions," *Signal Processing*, Vol. 70, 1998, pp. 247-278.
2. J. Lubin, "A human vision system model for objective picture quality measurements," *Proc. of the International Broadcasting Conference*, 1997, Amsterdam, Netherlands, pp. 498-503.
3. S. Winkler, "Issues in vision modeling for perceptual video quality assessment", *Signal Processing*, Vol. 78, No.2, 1999, pp. 231-252.
4. M.S. Moore, J.M. Foley and S.K. Mitra, "Defect visibility and content inportance: Effects on perceived impairment," Image Communication, February 2004 - to be published.
5. H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," *Proc. of the SPIE*, Human Vision and Electronic Imaging III, San Jose, CA, Vol. 1666, January 1992, pp. 16-26.
6. Michael S. Moore, John M. Foley, and Sanjit K. Mitra, "A comparison of detectability and annoyance value of MPEG-2 artifacts inserted into uncompressed video sequences," Proc. of the SPIE, Human Vision and Electronic Imaging V, San Jose, CA, vol. 4299, January 2001, pp. 90-101.
7. ITU Recommendation BT.500-8, "Methodology for subjective assessment of the quality of television pictures," 1998.
8. Khuri, A. I. & Cornell (1987) Response Surfaces: Designs and Analysis, New York: Dekker.
9. M.C. Q. Farias, S.K. Mitra, and J.M. Foley, "Perceptual contributions of blocky, blurry and noisy synthetic artifacts to overall annoyance," Proc. of the International Conference on Multimedia & Expo, Baltimore, July 2003, pp. 529 -532.