DETECTABILITY AND ANNOYANCE OF SYNTHETIC BLOCKINESS, BLURRINESS, NOISINESS, AND RINGING IN VIDEO SEQUENCES*

Mylène C. Q. Farias [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

Synthetic artifacts offer many advantages for experimental research on video quality because of the degree of control that the researchers have with respect to the amplitude, distribution, and mixture of artifacts. We have developed algorithms for synthetically generating four types of artifacts commonly found in digital videos: blockiness, blurriness, noisiness, and ringing. In this paper, we inserted these artifacts in short video sequences and performed a psychophysical experiment where we measured the probability of detection and the mean annoyance values of these artifacts as a function of their total squared error (TSE). The results show that, although the different artifacts looked different and affected the videos differently, there is no consistent difference between either their visibility thresholds or mid-annoyance TSE. Mid-annoyance values are positively correlated with the visibility threshold, and the relation can be described by a linear function.

1. INTRODUCTION

An impairment is defined as a perceived flaw introduced into an image or video during capture, transmission, storage, and/or display, as well as by any image processing algorithm (e.g. enhancement, compression). Most impairments are very complex and can be decomposed into one or more perceptual features or artifacts. Examples of artifacts introduced by digital systems 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 [2,3]. Psychophysical scaling experiments have shown that the overall annoyance of impairments increases when different artifacts are combined simultaneously. However, we do not yet have a good understanding of how annoyance and visibility depend on the type of artifact and video content.

One approach for studying impairments is to work with synthetic artifacts that look like "real" artifacts, yet are simpler, purer, and easier to describe [4]. This approach is promising because of the degree of control it offers with respect to the amplitude, distribution, and mixture of different types of artifacts. This control makes it possible, for example, to isolate individual artifacts for study. We have developed algorithms for synthetically generating four types of artifacts commonly found in digital videos: blockiness, blurriness, noisiness, and ringing [5-7]. In this paper, our goal was to estimate the visibility and annoyance of these four synthetic artifacts. We also wanted to compare the visibility threshold and mid-annoyance values of these artifacts. To this end, we inserted these artifacts in short video sequences and performed a psychophysical experiment where we measured the probability of detection and the mean annoyance values of the individual artifacts as a function of TSE.

2. SYNTHETIC ARTIFACTS

Blockiness is a distortion of the images characterized by the visibility of an underlying block encoding structure and is often caused by a coarse quantization of the spatial frequency components during the encoding process [1,4]. We produced blockiness by adding to each pixel a constant proportional to the difference between the average of each block and the average of the surrounding area. This causes each block to stand out creating blockiness [5].

Blurriness is a reduction in sharpness of edges and spatial detail. In compressed videos blurriness is often caused by trading off bits to code resolution and motion [1,4]. We produced blurriness by applying a symmetric, two-dimensional Finite duration Impulse Response (FIR) low-pass filter through the digital frame array [5].

Noise is defined as an undesired, uncontrolled or unpredicted pattern of intensity fluctuations. In every part of the video chain, from the source to display, the video may be impaired by noisiness. Sources of noisiness include electronic noise, photon noise, film-grain noise, and quantization noise [1,4]. We created noisiness by replacing the luminance value of pixels at random locations with a constrained random value. The color components were left untouched. The random location of the pixels to change was determined by drawing two random numbers, corresponding to the coordinates of the pixel. After a pixel location was determined, the pixel value was replaced by a random value. Additional pixel locations were selected until the desired ratio of impaired/non-impaired number of pixels is obtained [6].

Ringing is fundamentally related to the Gibb's phenomenon. It occurs when the quantization of individual DCT coefficients results in high frequency irregularities of the reconstructed block. Ringing manifests itself in the form of spurious

*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.

oscillations of the reconstructed pixel values [1,4]. The algorithm for synthetically generating ringing consisted of applying a pair of delay-complementary highpass and lowpass filters to the pixels of the video frame forming the edges [7]. Except for a shift, the output of this system is equal to the input, given that the initial conditions of both filters are exactly the same. If we make the initial conditions different, a decaying "noise" is introduced in the first $N/2$ samples. The resulting effect is very similar to the ringing artifact found in compressed images, but without any blurriness or noisiness [7].

3. TEST SEQUENCE GENERATION

To generate the test video sequences, we started by choosing a set of five original video sequences of assumed high quality: 'Bus', 'Calendar', 'Cheerleader', 'Flower', and 'Hockey'. These videos are 5 seconds long and are commonly used for video experiments and publicly available. The second step was to generate videos in which one type of artifact dominated and produced a relatively high level of annoyance. For each original, 4 new videos were created: X_{blurry}, with only blurriness, X_{blocky}, with only blockiness, X_{ringy}, with only ringing, and or X_{noisy}, with only noisiness. The artifacts were roughly matched in annoyance. However, we found that the highest possible TSE's for blockiness and ringing did not produce annoyance values anywhere near as high as the highest possible TSE's for blurriness and noisiness. Since we wanted to examine the full range of each artifact, the initial TSE's were higher for these artifacts.

Then, the test sequences (Y) were generated by linearly combining the original video with the video containing the individual artifact (X_{blurry}, X_{blocky}, X_{ringy}, or X_{noisy}) in different proportions, as given by the following equation:

$$Y = X + r \cdot (X_I - X), \quad (1)$$

where X is the original video, X_I is the sequence with the artifact, and r is the strength parameter of the test sequence ($r \geq 0$). Before adding them, the videos were transformed to the linear light domain using a gamma approximation. We used 6 strength values for each artifact in order to determine the annoyance and psychometric functions. Using results of a pilot study, we chose values of r that covered the ranges of both the psychometric and the annoyance functions in so far as possible.

The usual approach to subjective quality testing is to degrade the whole video by a variable amount and ask the test subjects for a quality rating [8]. Since both the type and the strength of artifacts vary from frame to frame and region to region, this method cannot be used to measure the visibility and annoyance produced by specific artifacts at specific strengths. In this work the artifacts were added only to isolated regions (defect zones) of the video clip for a short time interval [3]. The rest of the video was left in its original state. We used three rectangular defect zones, each of size equivalent to approximately 1/3 of the frame. This procedure has the advantages of making it possible to analyze the effect of different content on the visibility/annoyance of the impairments and of simplifying the subject's task.

A total of 465 test sequences were used in this experiment (5 originals x 4 artifacts x 3 defect zones x 6 strengths + 5 originals). Since it would be impossible to perform all these test conditions using only one set of subjects, we divided our subjects in three independent groups. Each group performed 1/3 of the set conditions (125 sequences), consisting of the all originals, all types of artifacts, all strength values, but different groups of defect zones.

4. METHOD

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. They were asked to wear any vision correcting devices (glasses or contacts) that they normally wear to watch television. There were five stages to the experimental session: instructions, training, practice, experiment, and interview [3, 5-7].

In the first stage, the subject was verbally given instructions. In the training stage, we showed sample sequences to the subject to establish the range for the strength and annoyance scales. In the practice stage, the subject carried out 8 practice trials to allow the responses to stabilize. At the interview stage, we asked the subject for qualitative descriptions of the defects that were seen. The main experiment was performed with the set of test sequences presented in random order. The test subject was instructed to search each video for defective regions. After each video was played the subject was asked two questions. The first question was "Did you see a defect or impairment?" If the answer was 'no', no further questions were asked. If the answer was 'yes', the subject was instructed to enter a positive numerical value indicating how annoying the defect was. Any defect half as annoying as the most annoying defect in the training stage should be given 50, as annoying 100, twice as annoying 200 and so forth.

5. DATA ANALYSIS

We used the standard methods [8] for analyzing the visibility and annoyance judgments provided by the test subjects. We first computed the Probability of Detection (*PD*) and the Mean Annoyance Value (*MAV*). *PD* was estimated by dividing the number of subjects who detected the artifact by the total number of subjects. *MAV* is calculated by averaging the annoyance scores over all observers for each video. The test sequences were divided in *test groups* corresponding to same original, same defect zone, same artifact type, and varying strength values.

Using the probability of detection data, we estimated the visibility detection threshold of the artifacts. The probability as a function of the log(TSE) (*psychometric function*) is fitted using the Weibull function [8], which has an S-shape similar to the experimental data and is defined as:

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

where $P(x)$ is the probability of detection, x is the logarithm of the TSE, S is the sensitivity, and k is a constant that determines the slope of the transition. The 50% detection threshold is simply $x_T = 1/S$. Figures 1-2 show the psychometric functions for the videos Calendar (defect zone Right) and Flower (defect zone Houses). Each graph contains four different curves, one for each type of artifact. The MAV, as a function of the log(TSE) (*annoyance function*), is fitted with the standard logistic function [8]:

$$y = (100)/\left(1 + \exp\left(-(x - \bar{x})/|\beta|\right)\right) \quad (3)$$

where y is the predicted annoyance and x is the logarithm of the TSE. The parameter $\bar{x}$ (mid-annoyance TSE) translates the curve in the x-direction and β controls the steepness.

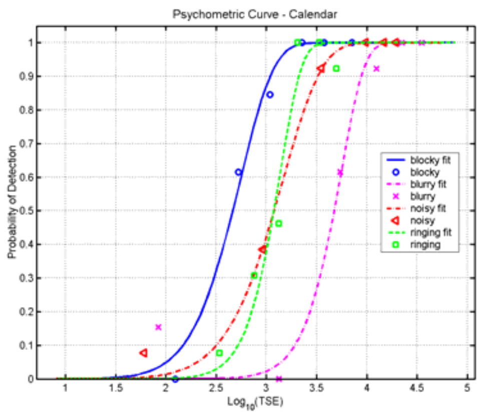

Figure 1. Psychometric function for 'Calendar'-Right.

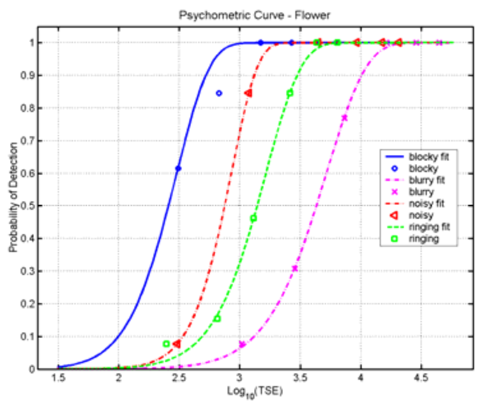

Figure 2. Psychometric function for 'Flower' - Houses.

Figures 3-4 show the annoyance functions for the videos 'Calendar' (defect zone Right) and 'Flower' (defect zone Houses). Each graph contains four different curves, one for eachtype of artifact. The corresponding curves have the same form as those for MPEG artifacts [5]. Table 1 shows the fitting parameters for both annoyance and psychometric functions. The results are divided by type of artifact (blockiness, blurriness, noisiness or ringing) and *test group*. Columns 2, 3, 6 and 7 show the psychometric function parameters (x_T and k). The empty spaces correspond to cases where a fit was not possible because more than 50% of subjects detected the weakest impairment. Columns 4, 5, 8, and 9 show the annoyance function parameters ($\bar{x}$ and β).

From Table 1 can be seen that the visibility threshold (x_T) and mid-annoyance ($\bar{x}$) parameters vary over the different *test groups*. Figures 5 and 6 show the bar plots of the threshold and mid-annoyance parameters average values for the different originals and artifact types. As can be seen from these bar plots, the original video seems to have a larger effect on the visibility threshold and mid-annoyance parameters than the type of artifact. To confirm this hypothesis, we performed an analysis of variance on the parameter values to test for the main effects of 'original', 'defect region', and 'artifact type'. Table 2 shows the *P* values obtained from this analysis (significant effects are shown in bold). The 'artifact type' did not have a significant effect on any of the fitting parameters. On the other hand, 'original video' had a significant effect on the parameters $\bar{x}$ and x_T. This analysis implies that, although the artifacts differ in their

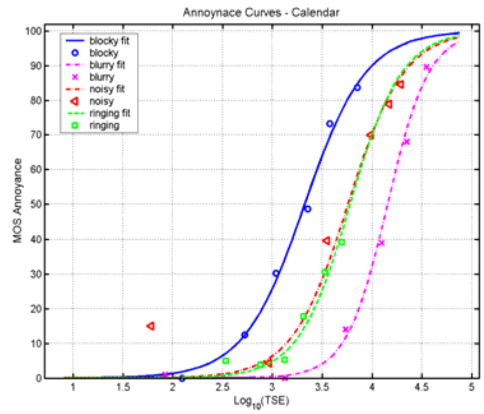

Figure 3. Annoyance function for 'Calendar'- Right.

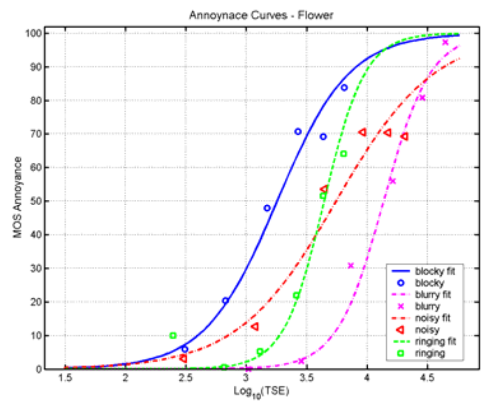

Figure 4. Annoyance function for 'Flower' - Houses.

appearance and in how they interact with the video content, the differences between either the thresholds or mid-annoyance values are not statistically significant. Moore [3] and Farias [7] also found that the effect of artifact type on threshold and mid-annoyance TSE was not statistically significant for MPEG-2 and synthetic artifacts.

We calculated the Pearson correlation [8] between xT and $\bar{x}$ for the set of all test groups. The parameters have a positive correlation of 0.783. In Figure 7 the values of $\bar{x}$ are plotted against the values of xT for the complete set of test sequences. The parameters are related by the following linear expression $\bar{x} = 0.711\ xT + 1.507$ obtained by fitting a linear equation to the data (line shown in the Figure 7). Therefore, if we know the visibility thresholds of these impairments we can estimate their annoyance using a simple linear relation. For the individual artifacts, the correlation remained almost the same, except for blurriness. The xT and $\bar{x}$ parameters obtained for blurriness have a higher correlation (0.9317) and mid-annoyance strength is proportional to threshold, $\bar{x} = 1.1325\ xT$. Moore [3] also found a positive correlation between the xT and $\bar{x}$ values and a proportional relation.

6. CONCLUDING REMARKS

The goal of this study was to estimate the visibility and annoyance of four synthetic artifacts: blockiness, blurriness, noisiness, and ringing. To this end, we inserted these artifacts in short video sequences and performed a psychophysical experiment where we measured the probability of detection and the mean annoyance values of the individual artifacts. We used standard methods to find the psychometric and annoyance functions. The corresponding curves have the same form as has been used for MPEG artifacts. We found that, although the artifacts differ in their appearance and in how they interact with the video content, there is no consistent difference between either their visibility thresholds or mid-annoyance TSEs. For all four artifacts, mid-annoyance is positively correlated with the visibility threshold, and the relation can be described by a linear function.

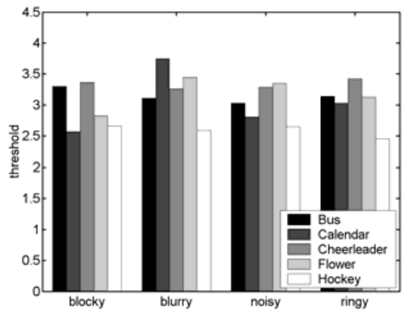

Figure 5. Bar plot of the average of threshold values (x_T).

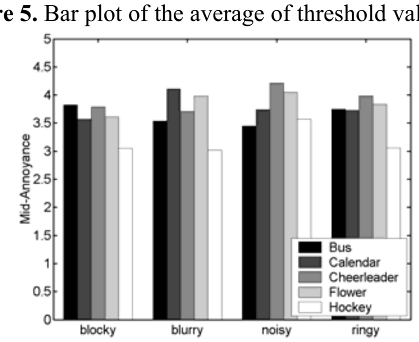

Figure 6. Bar plot of the average of mid-annoyance values ($\bar{x}$).

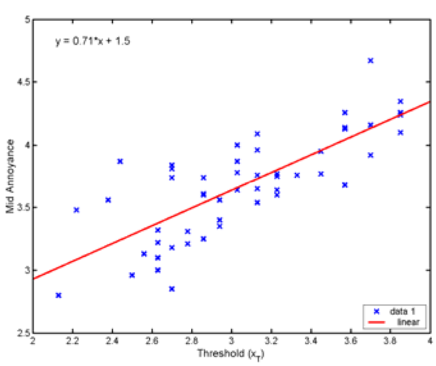

Figure 7. $\bar{x}$ versus x_T for the data set of all video sequences.

7. REFERENCES

[1] M. Yuen, and H.R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, 70 (1998), pp. 247-278.

[2] S. Winkler, "Issues in vision modeling for perceptual video quality assessment", *Signal Processing*, Vol. 78, No.2, 1999, pp. 231-252.

[3] M.S. Moore, J.M. Foley and S.K. Mitra, "Defect visibility and content inportance: Effects on perceived impairment," Image Communication, vol.19, No.2, February 2004, pp.185-203.

[4] Recommendation ITU-R BT.500-930, "Principals of a reference impairment system for video," ITU-T 1996.

[5] M.C.Q. Farias, J. M. Foley, and S. K. Mitra, "Some Properties of Synthetic Blocky and Blurry Artifacts," Proc. IS&T/SPIE 2003 International Symposium on Electronic Imaging Science and Technology, Santa Clara, CA, 2003.

[6] M.C.Q. Farias, J. M. Foley, and S. K. Mitra, "Perceptual Contributions of Blocky, Blurry and Noisy Artifacts to Overall Annoyance," Proc. International Conference on Multimedia & Expo, Baltimore, MD, July 2003.

[7] M.C.Q. Farias, J. M. Foley, and S. K. Mitra, "Detectability and annoyance of synthetic blurring and ringing in video sequences," Proc. International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, May 2004.

[8] ITU Recommendation BT.500-8, "Methodology for Subjective Assessment of the Quality of Television Pictures," 1998.

Table 1: Annoyance and psychometric functions parameters.

Test group	x_T	k	$\bar{x}$	β	x_T	k	$\bar{x}$	β
	blocky				noisy			
BusTop	3.125	4.94	3.96	0.5	2.941	19.37	3.35	0.27
BusMid	2.941	14.97	3.4	0.19	-	-	3.72	0.83
BusBot	3.846	6.04	4.1	0.22	3.125	15.97	3.54	0.26
CalenLeft	2.381	4.6	3.56	0.46	2.857	8.9	3.6	0.37
CalenMid	2.703	14.83	3.84	0.43	2.439	5.63	3.87	0.44
CalenRig	2.632	9.18	3.32	0.31	3.125	9.12	3.76	0.28
CalenTop	3.448	17.16	3.77	0.18	3.030	6.25	3.87	0.33
CheerMid	3.226	22.71	3.64	0.18	3.125	12.71	4.09	0.34
CheerBot	3.448	12.01	3.95	0.33	3.704	4.04	4.67	0.43
FlowTop	2.632	6.29	3.22	0.27	-	-	3.96	0.7
FlowMid	-	-	3.26	0.3	2.857	14.76	3.74	0.41
FlowBot	3.030	13.98	4	0.4	3.846	33.58	4.35	0.26
HockLeft	2.703	3.82	2.85	0.5	2.857	6.54	3.61	0.23
HockMid	2.703	10.94	3.18	0.26	2.857	7.71	3.61	0.29
HockRig	2.564	7.13	3.13	0.19	2.222	4.68	3.48	0.47
Test group	x_T	k	$\bar{x}$	β	x_T	k	$\bar{x}$	β
	blurry				ringing			
BusTop	2.857	11.32	3.25	0.15	-	-	4.03	0.5
BusMid	3.704	31.34	4.16	0.28	2.703	4.7	3.81	0.4
BusBot	2.778	11.57	3.21	0.17	3.571	11.56	3.68	0.06
CalenLeft	3.704	16.47	3.92	0.50	3.030	9.25	3.78	0.31
CalenMid	3.846	12.67	4.26	0.50	3.030	12.53	3.64	0.24
CalenRig	3.704	16.83	4.16	0.21	3.030	15.38	3.78	0.26
CalenTop	3.226	9.08	3.77	0.22	3.226	7.15	3.75	0.18
CheeMid	3.333	8.68	3.76	0.2	3.448	12.79	3.95	0.18
CheerBot	3.226	16.59	3.6	0.26	3.571	7.27	4.26	0.4
FlowTop	2.941	9.40	3.56	0.26	2.703	14.1	3.74	0.34
FlowMid	3.571	12.32	4.14	0.19	3.125	12.07	3.65	0.17
FlowBot	3.846	19.62	4.24	0.15	3.571	16.34	4.13	0.21
HockLeft	2.632	13.89	3.00	0.18	-	-	3.8	0.5
HockMid	2.632	7.73	3.10	0.18	2.778	8.49	3.31	0.2
HockRig	2.500	5.80	2.96	0.18	2.128	4.17	2.8	0.33

Table 2: P values obtained from the ANOVA analysis on the annoyance and visibility fitting parameters (Table 1).

Annoyance	$\bar{x}$			β		
	orig	artif	reg	orig	artif	Reg
P	**0**	0.120	0.367	0.552	0.022	0.184
Visibility	S			K		
	orig	artif	reg	orig	Artif	Reg
P	**0.003**	0.535	0.974	0.093	0.408	0.150