Some Properties of Synthetic Blocky and Blurry Artifacts

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

This work addresses the problem of studying and characterizing individual artifacts found in digital video applications (e.g., blockiness, blurriness). In particular, the goal of this paper was to examine the properties of synthetic blocky and blurry artifacts, designed to be relatively pure, and a combination of these two artifacts. We performed a psychophysical experiment in which human subjects were asked to detect these artifacts, identify their types, and rate their annoyance. In most cases, the blocky and blurry artifacts were identified as blocky and blurry, respectively. In combined blocky-blurry artifacts the salience of blockiness increased and blurriness decreased as artifact strength increased. The blocky artifacts produced higher annoyance values than the blurry ones when the total squared error was the same.

Keywords: artifacts, perceptual image quality, video.

1. INTRODUCTION

An impairment or a defect 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) that may be applied along the way. Impairments can be very complex in their physical description and also in their perceptual description. Most of them have more than one perceptual feature. Nevertheless, it is possible to produce impairments that are relatively pure. To differentiate impairments from their features, we will use the term 'perceived artifacts' to refer to the features. Examples of perceived artifacts introduced by digital systems are blurriness, noisiness 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. A study of the individual perceived artifacts is necessary since the quality of a video depends on the type of artifacts as well as their visibility.[2] Psychophysical scaling experiments have shown that the overall annoyance of impairments increases when different artifacts are combined simultaneously.[3] However, 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.

One approach for studying impairments is to work with synthetic artifacts that look like "real" artifacts, yet are simpler, purer, and easier to describe. 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 study the importance of each type of artifact for human observers. Such artifacts are necessary components of the kind of reference impairment system recommended by the ITU-T for the measurement of image quality[4] and offer advantages for experimental research on video quality.

There are several properties that are desirable in synthetic artifacts, if they are to be useful for these purposes. The synthetic artifacts should:

- ❑ be generated by a precisely defined and easily replicated algorithm,

* Further author information: (Send correspondence to M.C.Q.F.)
M.C.Q.F.: E-mail: mylene@ece.ucsb.edu, Telephone: 1 805 893 8312
J.M.F.: E-mail: foley@psych.ucsb.edu, Telephone: 1 805 893 2030
S.K.M.: E-mail: mitra@ece.ucsb.edu, Telephone: 1 805 893 8312

Human Vision and Electronic Imaging VIII, Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Editors, Proceedings of SPIE–IS&T Electronic Imaging, SPIE Vol. 5007 (2003)

- be relatively pure and easily adjusted and combined to match the appearance of the full range of compression impairments, and
- produce psychometric functions and annoyance functions that are similar to those for compression artifacts.

We created two synthetic artifacts; one is perceived as predominantly blurry and the other as predominantly blocky. The goal of the study was to examine properties of these two synthetic artifacts and their combinations. We performed a psychophysical experiment in which human subjects detected these artifacts, identified their features, and rated their annoyance.

2. GENERATION OF SYNTHETIC ARTIFACTS

Blockiness is a distortion of the image characterized by the visibility of an underlying block encoding structure[1,4] and is often caused by coarse quantization of the spatial frequency components during the encoding process. We produced blocky artifacts by using the difference between the average of each block and the average of the surrounding area to make each block stand out.

The algorithm for producing the blocky artifacts is as follows: The first step is to calculate the average of each 8×8 block of the frame and of the 24×24-surrounding block, which has the current 8×8 block as a center. The next step is to calculate the difference, $D(i,j)$, between these two averages for each block. This difference is added to each block of the original frame $X(k,l)$:

$$X_{blocky}(k,l) \;=\; X(k,l) + D(i,j), \tag{1}$$

where $i = round(k/8)$, with $1 \leq i \leq Rows/8$, and $j = round(l/8)$, with $1 \leq i \leq Cols/8$. X_{blocky} is the frame with blocky artifacts, $Rows$ is the total number of rows of the frame, and $Cols$ is the number of columns of the frame. The frame average was matched to the original by adding or subtracting a constant to each pixel. This procedure guarantees that the average of the frame does not change much. A saturation control was also implemented by limiting the value of $D(i,j)$ added to each pixel. The same algorithm was applied to the luminance and the two color components of the video.

Blurriness is the reduction in sharpness of edges and spatial detail.[1,4] In compressed images blurriness is often caused by trading off bits to code resolution and motion. Blurry artifacts were generated by applying a symmetric, two-dimensional FIR (finite duration impulse response) low-pass filter to the digital image array. We used a 5×5 mean filter[7] in this experiment. Different filters with varying cut-off frequencies could be used to control the amount of blurriness introduced. A detailed description of how blocky and blurry artifacts were created can be found in Ref. 5.

The test videos used in this experiment contained synthetic blocky and blurry artifacts and a combination of these two types of artifacts. To generate the test video sequences, we started by choosing a set of five original video sequences of assumed high quality: 'Bus', 'Cheerleader', 'Flower-garden', 'Football' and 'Hockey'. These videos are commonly used for video experiments and are publicly available.[6] The video clips are all 5 seconds long and contain scenes that we think are typical of normal television. The second step was to generate videos in which one type of artifact dominated by using the algorithms described above. For each original, we created a sequence with blocky artifacts, X_{blocky}, and a sequence with blurry artifacts, X_{blurry}. Before creating the artifacts, the videos were transformed to the linear light domain using a gamma approximation.

Then for each original, we created a video , X_{comb}, with the blocky and blurry artifacts combined by using a fixed mixture of these two artifacts:

$$X_{comb} \;=\; 0.5 \cdot X_{blocky} + 0.5 \cdot X_{blurry}\,, \tag{2}$$

The artifacts were created using the same parameters as in Ref. 5, where we tried to create a combination of blocky and blurry that looked as similar as possible to a MPEG-2 sequence compressed at 7.5 Mbps. The same strength of blockiness and blurriness was used for this experiment, but this time the artifacts were presented alone as well as combined. Finally, the test sequences were generated by linearly combining the original video with the video with artifacts in different proportions.

Table 1 Scaling factors and fitting parameters for annoyance functions..

Videos	*r*	$\overline{x}$	β	Residuals
Blocky Bus	0.48,0.68,0.80,0.92,1.15,1.20	3.48	0.5	3.39
Blocky Cheer	0.44,0.60,0.72, 0.92,1.25,1.70	3.92	0.51	8.29
Blocky Flower	0.44,0.60,0.72, 0.88,1.25,1.70	3.38	0.47	4.66
Blocky Football	0.48,0.68,0.80,1.10,1.50,1.70	3.17	0.35	4.38
Blocky Hockey	0.48,0.68,0.80,0.92,1.25,1.50	3.39	0.63	6.59
Blurry Bus	0.36,0.48,0.68,0.80,0.92,1.05	4.15	0.36	6.1
Blurry Cheer	0.32,0.44,0.60,0.72,0.92,1.05	3.83	0.5	5.68
Blurry Flower	0.32,0.44,0.60,0.72,0.88,1.05	4.15	0.28	6.02
Blurry Football	0.36,0.48,0.68,0.80,0.92,1.05	3.5	0.37	3.63
Blurry Hockey	0.48,0.68,0.80,0.92,1.05,1.25	3.45	0.45	3.98
Comb Bus	0.27,0.41,0.54,0.77,0.90,1.00	4.06	0.35	7.05
Comb Cheer	0.23,0.36,0.50,0.68,0.79,1.00	3.83	0.37	7.58
Comb Flower	0.23,0.36,0.50,0.68,0.81,0.99	3.91	0.32	2.95
Comb Football	0.27,0.41,0.54,0.77,0.90,1.00	3.44	0.34	3.61
Comb Hockey	0.27,0.41,0.54,0.77,0.90,1.00	3.3	0.43	7.04

The basic formula is:

$$Y = X + r \cdot (X_{\mathrm{I}} - X), \qquad (3)$$

where Y is the result, X is the original, X_{I} is the impaired video with artifacts (X_{comb}, X_{blocky} or X_{blurry}), and r is the scaling factor. By varying the scaling factor, we could vary the magnitude of the artifacts relative to the magnitude produced by our algorithms. The values of r used in this experiment are shown in Table 1.

3. METHOD

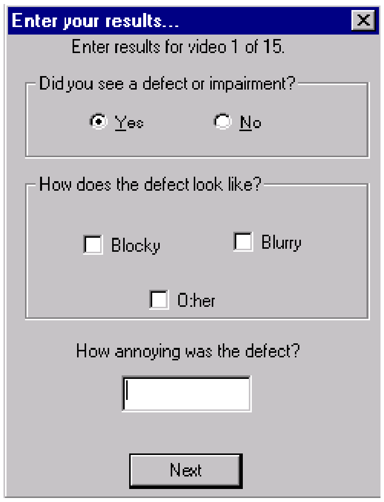
The normal approach to subjective quality testing is to degrade a video by a variable amount and ask the test subjects for a quality/impairment rating.[8] The degradation is usually applied to the entire video. In this research we have been using an experiment paradigm that measures the annoyance value of brief, spatially limited artifacts in video.[2] We degrade one specific region of the video for a short time interval. The rest of the video clip is left in its original state. Different regions were used for each original to prevent the test subjects from learning the locations where the defects appear. The regions used in this experiment were centered strips (horizontal or vertical) taking 1/3 of the frame. They were 1 second long and did not occur during the first and last seconds of the video.

The Image Processing Laboratory (IPLAB) at UCSB, in conjunction with the Visual Perception Laboratory, has been performing experiments on video quality for the last three years. 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.

For our experiments, the test sequences are stored on the hard disk of an NEC server. Each video is displayed using a subset of the PC cards normally provided with the Tektronix PQA-200 picture quality analyzer. Each test sequence can be loaded and displayed in six to eight seconds. A generator card is used to locally store the video and stream it out in a serial digital (SDI) component format. 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. The result is five seconds of broadcast quality (except for the impairment), full-resolution, NTSC video.

In addition to storing the video sequences, the server is also used to run the experiment and collect data. A special-purpose program records each subject's name, displays the video clips, and runs the experiment. After each test sequence is shown, the experiment program displays a series of questions on a computer monitor and records the subject's responses in a subject-specific data file.

The experiments are run with one test subject at a time. The subjects are asked to wear any vision correction devices (glasses or contacts) that they would normally wear to watch television. Each subject is seated in front of the



Enter your results...

Enter results for video 1 of 15.

Did you see a defect or impairment?

● Yes ○ No

How does the defect look like?

□ Blocky □ Blurry

□ Other

How annoying was the defect?

Figure 1 Dialog box used to collect data from the experiment subjects.

computer keyboard at one end of a table. Directly ahead of the subject is the Sony video monitor, located at or slightly below eye height for most subjects. The subjects are positioned at a distance of four screen heights (80 cm) from the video monitor. The subjects are instructed to keep their heads at this distance during the experiment, and their position is monitored by the experimenter and corrected when needed.

The course of each experimental session goes through five stages: instructions, examples, practice, experimental trials, and interview. In the first stage, the subject is verbally given instructions. In the second stage, sample sequences are shown to the subject. The sample sequences represent the impairment extremes for the experiment and are used to establish the annoyance value range. The practice trials are identical to the experimental trials, except that no data is recorded. The practice trials are also used to familiarize the subject with the experiment. Twelve practice trials are included in this session to allow the subjects' responses to stabilize before the experimental trials begin.

After the practice trials, the main experiment is performed with the complete set of test sequences. In the main experiment, subjects enter data using the dialog box shown in Figure 1. The first question in the dialog box is whether an impairment or defect was seen. If the answer is 'no', the subject can hit next and the following test sequence is presented. If the answer is 'yes', two more questions are asked. The second question is 'How would you describe the defect?'. To answer this question the subject was presented with three options – 'blocky', 'blurry' and 'other'. 'Other' meant that the video contained other artifacts besides blockiness and blurriness. The subjects were instructed to pick one, two or all three of these options, as necessary, to describe the impairment. The third question was 'How annoying was the defect?'. To answer this the subject entered a numerical value, where '0' means that the defect is not annoying at all and '100' means that it is as annoying as the worst sample video. Any half as annoying should be given 50, and any twice as annoying 200 and so forth. Although we try to include the worst test sequences in the sample set, we acknowledge the fact that the subjects might find some of the other tests clips to be more annoying and specifically instruct them to go above 100 in that case.

At the end of the experimental trials, we ask 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 defect feature analysis experiments.

The total number of test sequences in this experiment was 95, which included 90 test sequences (5 originals x types of artifacts times 6 scaling factors) plus the five original sequences. The sequences were shown in different random orders for different groups of observers during the main experiment.

4. DATA ANALYSIS

We used the standard methods[8] for analyzing the annoyance judgments provided by the test subjects. We first computed two measures: the Total Squared Error (TSE) and the Mean Annoyance Value (MAV) for each test sequence.

The TSE is our objective error measure and is defined as:

$$\mathrm{TSE} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - X_i\right)^2 , \tag{4}$$

where Y_i is i-th pixel value of the test sequence, X_i is the corresponding pixel of the original sequence, and *N* is the total number of pixels in the video. The *MAV* is our subjective error measure and is calculated by averaging the annoyance levels over all observers for each video:

$$MAV = \frac{1}{M}\sum_{i=1}^{M} A(i), \tag{5}$$

where *A (i)* is the annoyance level reported by the *i*-th observer. *M* is the total number of observers.

The mean annoyance values for each test sequence were fitted with the standard logistic function: [8]

$$PMAV = y_{\min} + \frac{\left(y_{\max} - y_{\min}\right)}{\left(1 + \exp\left(-\frac{\left(x - \bar{x}\right)}{|\beta|}\right)\right)}, \tag{6}$$

where $PMAV$ is the predicted mean annoyance value and x is the log10 (TSE). The parameters $y_{\max}$ and $y_{\min}$ establish the limits of the annoyance value range. The parameter $\bar{x}$ translates the curve in the *x*-direction and the parameter β is inversely related to the steepness of the curve.

Figures 2-6 depict the mean annoyance values versus the log total squared error (annoyance functions) for the five videos with the three types of impairments. The graphs show three curves for each video, one for each type of impairment – blocky, blurry and combined blocky and blurry artifacts.

These graphs show that at the functions for the blocky artifacts are higher than the functions for the blurry artifacts for three of the videos – Bus, Flower and Football. For these three videos, the combined blocky and blurry artifacts produced intermediate annoyance functions. For the two other videos - Cheerleader and Hockey - the annoyance functions are all very similar. The fitting parameters for these functions are presented in Table 1. When we compare across artifacts for the same video, the values of $\bar{x}$ are generally least for blocky and greatest and they are correlated. The values of the parameter β for blocky artifacts are greater than for blurry or combined artifacts; this means the annoyance curve is less steep for the blocky artifacts. Across videos, the values of the parameters $\bar{x}$ and βdo not vary greatly, but they do show that annoyance depends on the video as well as the artifact.

To analyze how the subjects described the impairments that they saw, we first determined the percentage of subjects who described the impairments as 'blocky', blurry' or 'other'. Figures 7-21 depict the percentage of subjects, who judged the impairments as 'blocky', 'blurry' and 'other' versus the log total squared error, for all the test videos. Figures 7-11 correspond to sequences with blocky artifacts, Figures 12-16 to the blurry artifacts and, Figures 17-21 to sequences with combined blocky and blurry artifacts. Since subjects could mark more than one feature for each sequence, these percentages often add to more than 100.

From Figures 7-16 it is clear that almost all subjects in most conditions judged the artifacts designed to be blocky and blurry to be 'blocky' and 'blurry', respectively. These figures also show that there is a percentage of subjects that judged the sequences designed to have one artifact to have some of the other types of artifacts as well. The percentage of classifications as 'blocky' and 'other' were small (under 15%) for all blurry test sequences (Figs. 13-16). On the other hand, for the blocky test sequences these percentages were higher, especially for the lowest TSEs. These percentages go over 20% for most videos and hockey is judged to be 'blurry' by more than 65% and flower is judged to be 'other' by more than 50%. That the Hockey video was reported to appear blurry at low TSE may be due to a blurry appearance of the ice even in the original video.

Figures 17-21 show that for all videos the percentage of subjects that judged these combined artifacts as 'blocky' increased with the TSE of the artifact, although the proportions of the two impaired videos in the combined video were constant for all TSEs. This effect was already observed (weakly) in the case of the blocky artifact. For higher log TSE levels the blocky artifact seems to become more salient than the blurring and some of the subjects classify it only as 'blocky'. Thus the appearance of the blocky and combined artifacts depends in a complex way on the physical signals. Blurry artifacts, on the other hand, are almost always seen as only blurry.

5. CONCLUSIONS

This paper presented a study of the way two relatively pure synthetic artifacts – blockiness and blurriness and their combinations are perceived. We performed a psychophysical experiment in which human subjects detected these artifacts, identified their features, and rated their annoyance. The blocky and blurry artifacts were usually identified as blocky and blurry, respectively. In combined blocky-blurry artifacts the salience of blockiness increased and blurriness decreased as artifact strength increased. The blocky artifacts produced higher annoyance values than the blurry ones when TSE was the same.

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 matched support from Philips Research Laboratories.

REFERENCES

1. M. Yuen, and H.R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, **70**, pp. 247-278, 1998.
2. M.S. Moore, "Psychophysical measurement and prediction of digital video quality," Ph.D. thesis, University of California Santa Barbara, June 2002.
3. H. de Ridder, and M.C. Willemsen, "Percentage scaling: A new Method for evaluating multiply impaired images," *Proc. of the SPIE*, Human Vision and Electronic Imaging V, San Jose, CA, vol. 3016, pp. 68-77, January 2000.
4. Recommendation ITU-R BT.500-930, "Principals of a reference impairment system for video," ITU-T 1996.
5. M.C.Q. Farias, M.S. Moore, J.M. Foley, and S.K. Mitra, "Detectability and annoyance of synthetic blocky and blurry artifacts," *Proc. of SID International Symposium*, Boston, MA, May, 2002.
6. Video Quality Experts Group, "VQEG subjective test plan," 1999.
7. R.C. Gonzalez, and R.E. Woods, *Digital Image Processing*, Addison Wesley, 1992.
8. ITU Recommendation BT.500-8, "Methodology for subjective assessment of the quality of television pictures," 1998.
9. 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, pp. 16-26, January 1992.

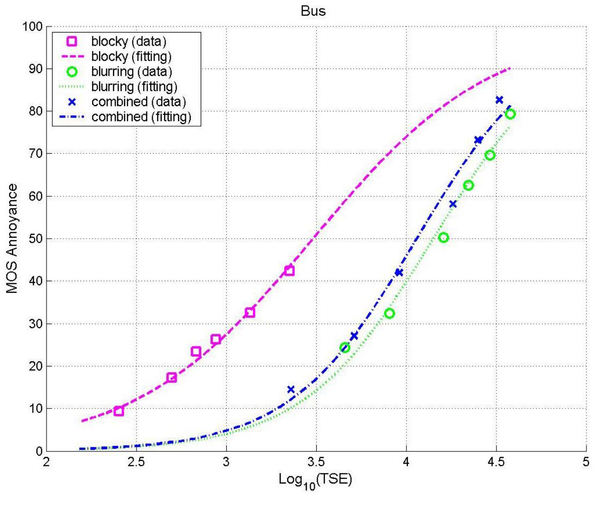


Figure 2. Annoyance functions for the Bus video.

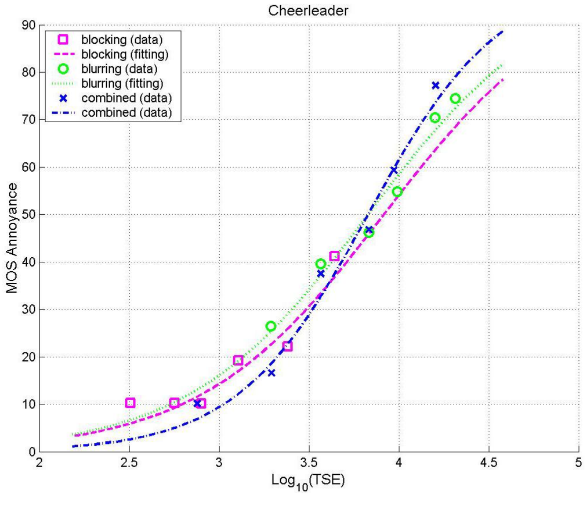


Figure 3. Annoyance curves for the Cheerleader video.

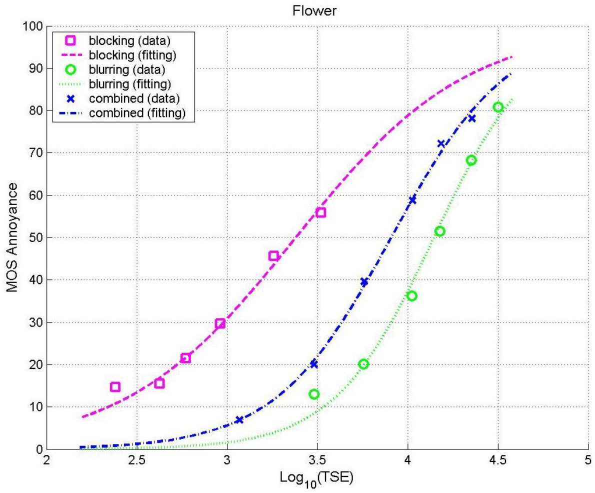


Figure 4. Annoyance curves for the Flower video.

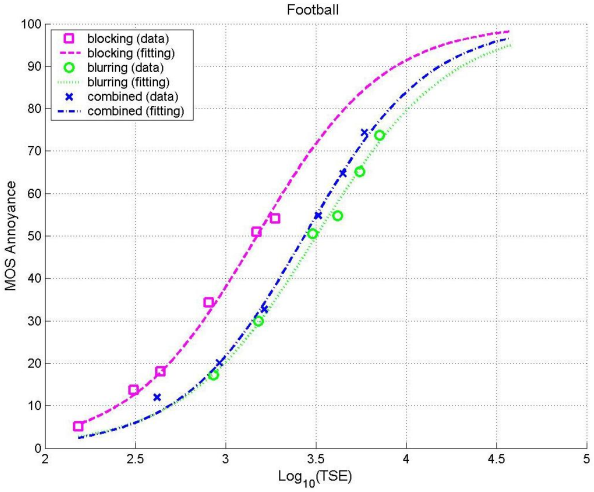


Figure 5. Annoyance curves for the Football video.

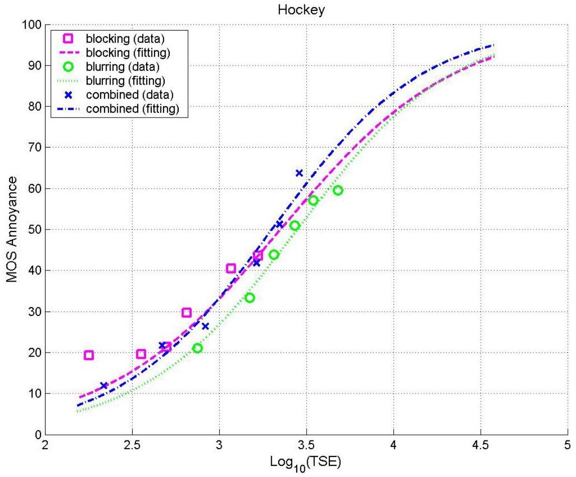


Figure 6. Annoyance curves for the Hockey video.

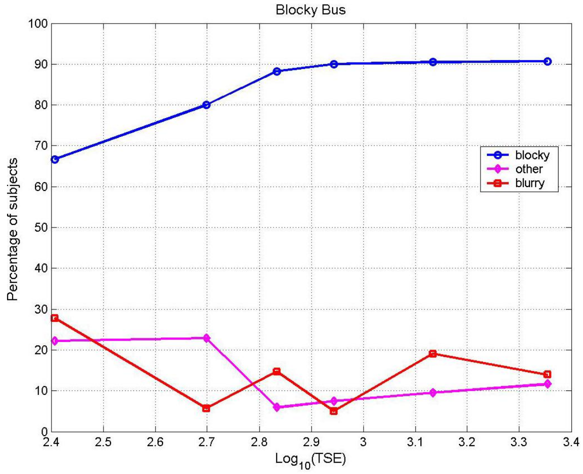


Figure 7. Percentage of subjects that judged the blocky video Bus as 'blocky', 'blurry', and 'other'.

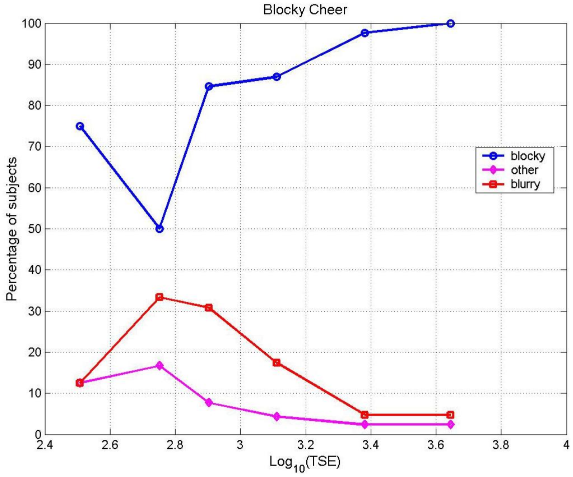


Figure 8. Percentage of subjects that judged the blocky video Cheerleader as 'blocky', 'blurry', and 'other'.

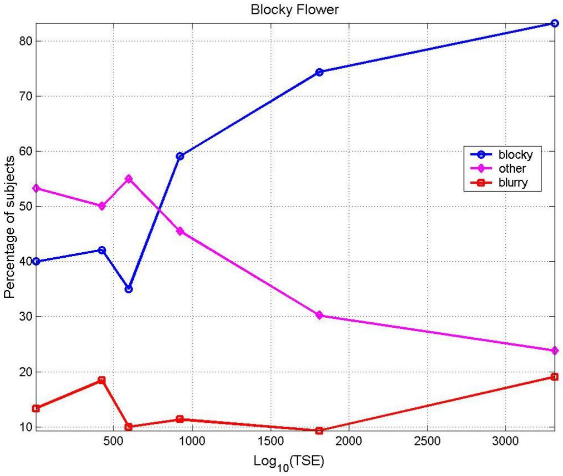


Figure 9. Percentage of subjects that judged the blocky video Flower as 'blocky', 'blurry', and 'other'.

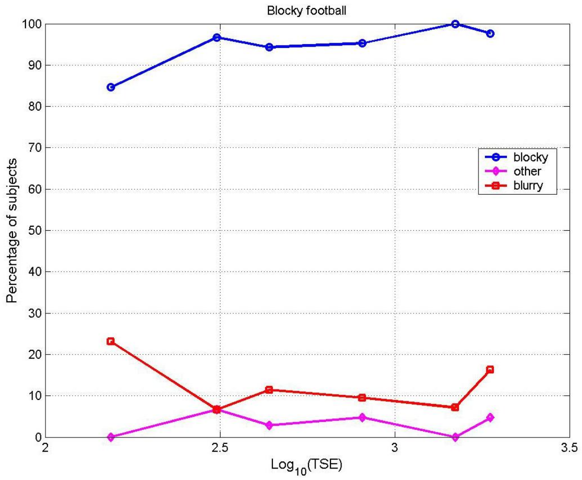


Figure 10. Percentage of subjects that judged the blocky video Flootball as 'blocky', 'blurry', and 'other'.

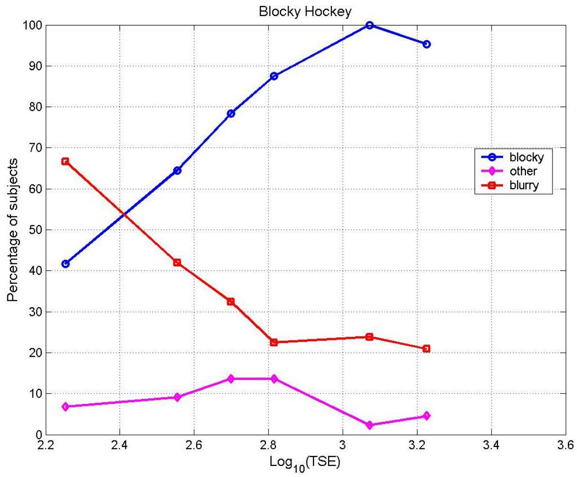


Figure 11. Percentage of subjects that judged the blocky video Hockey as 'blocky', 'blurry', and 'other'.

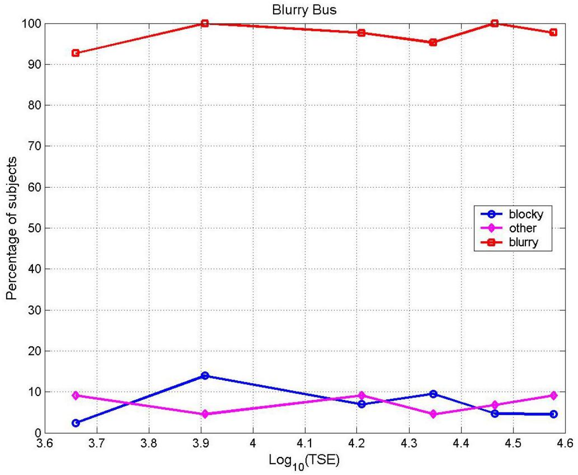


Figure 12. Percentage of subjects that judged the blurry video Bus as 'blocky', 'blurry', and 'other'.

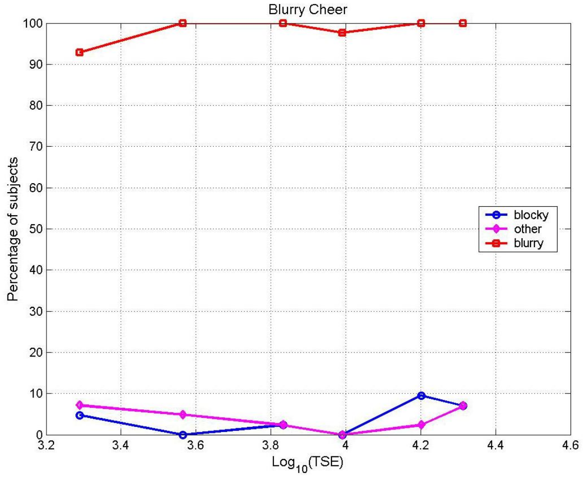


Figure 13. Percentage of subjects that judged the blurry video Cheerledar as 'blocky', 'blurry', and 'other'.

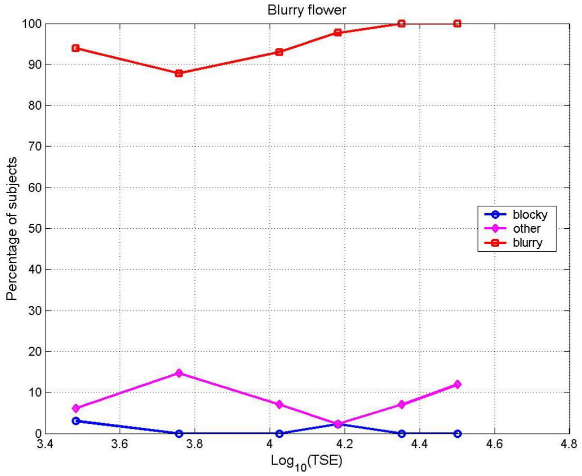


Figure 14. Percentage of subjects that judged the blurry video Flower as 'blocky', 'blurry', and 'other'.

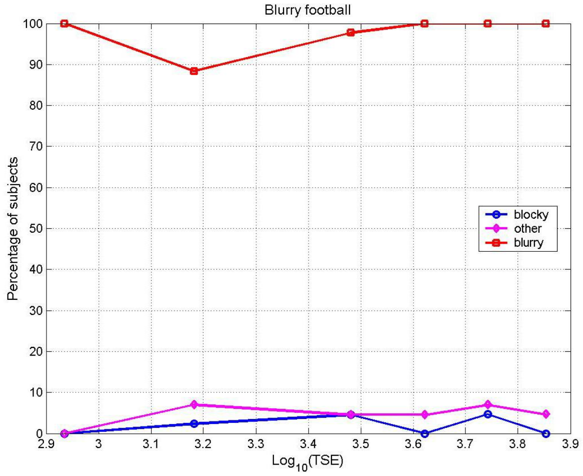


Figure 15. Percentage of subjects that judged the blurry video Football as 'blocky', 'blurry', and 'other'

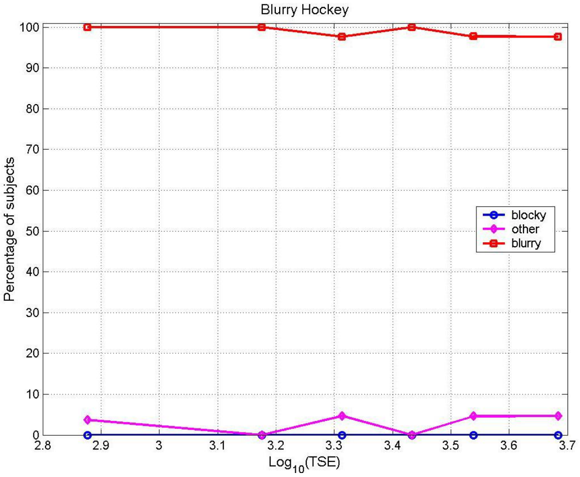


Figure 16. Percentage of subjects that judged the blurry video Hockey as ‘blocky’, ‘blurry’, and ‘other’.

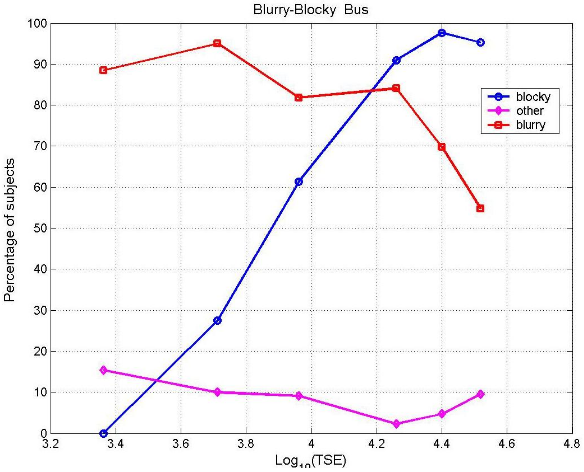


Figure 17. Percentage of subjects that judged the blocky-blurry video Bus as ‘blocky’, ‘blurry’, and ‘other’.

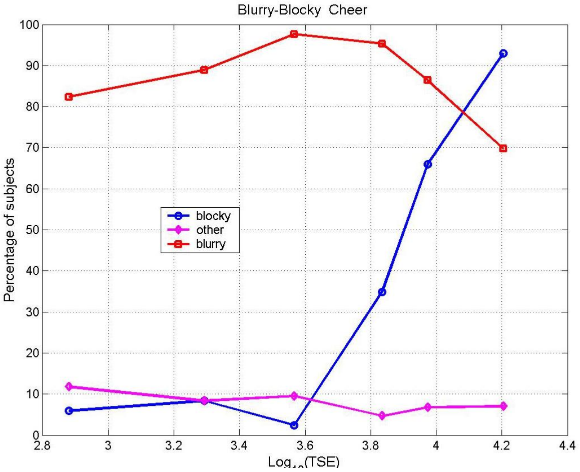


Figure 18. Percentage of subjects that judged the blocky-blurry video Cheerleader as ‘blocky’, ‘blurry’, and ‘other’.

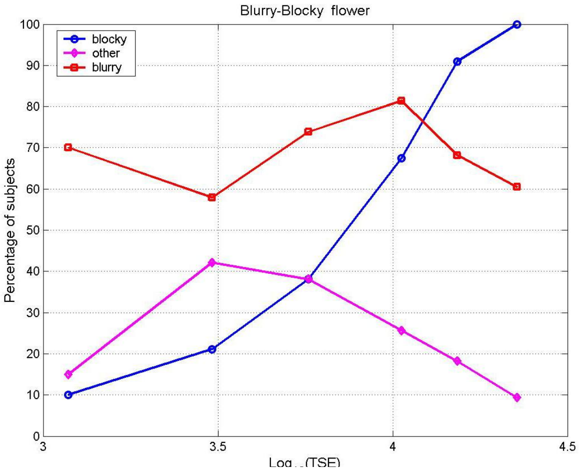


Figure 19. Percentage of subjects that judged the blocky-blurry video Flower as ‘blocky’, ‘blurry’, and ‘other’.

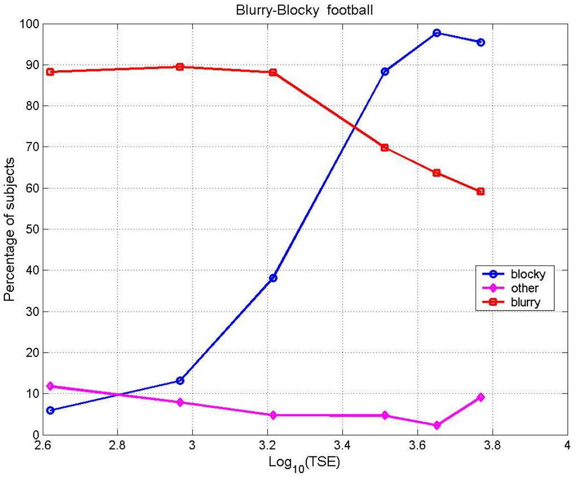


Figure 20. Percentage of subjects that judged the blocky-blurry video Football as ‘blocky’, ‘blurry’, and ‘other’.

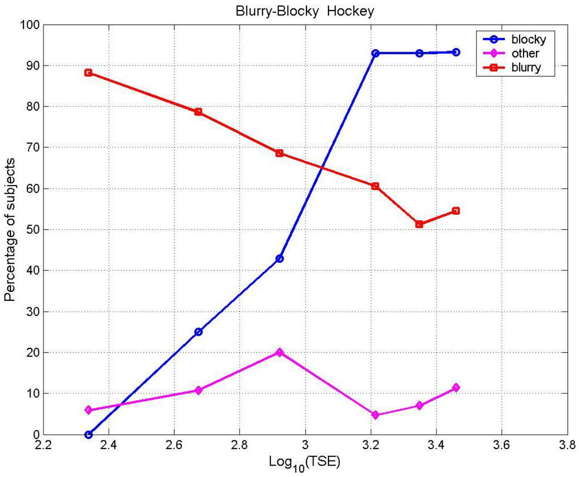


Figure 21. Percentage of subjects that judged the blocky-blurry video Hockey as ‘blocky’, ‘blurry’, and ‘other’.