Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance

Mylène C. Q. Farias
Sanjit K. Mitra

Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance

Mylène C. Q. Farias
University of Brasília
Department of Electrical Engineering
Campus Universitário Darcy Ribeiro
70919-970 Brasília, DF, Brazil
E-mail: mylene@ieee.org

Sanjit K. Mitra
University of California
Department of Electrical and Computer Engineering
Santa Barbara, California 93106

Abstract. *To develop a no-reference video quality model, it is important to know how the perceived strengths of artifacts are related to their physical strengths and to the perceived annoyance. When more than one artifact is present, it is important to know whether and how its corresponding perceived strength depends on the presence of other artifacts and how perceived strengths combine to determine the overall annoyance. We study the characteristics of different types of artifacts commonly found in compressed videos. We create artifact signals predominantly perceived as blocky, blurry, ringing, and noisy and combine them in various proportions. Then, we perform two psychophysical experiments to independently measure the strength and overall annoyance of these artifact signals when presented alone or in combination. We analyze the data from these experiments and propose models for the overall annoyance based on combinations of the perceptual strengths of the individual artifact signals. We also test the interactions among different types of artifact signals. The results provide interesting information that may help the development of video quality models based on artifact measurements. © 2012 SPIE and IS&T.* [DOI: 10.1117/1.JEI.21.4.043013]

1 Introduction

A video impairment is any change in a video signal that, if sufficiently strong, may reduce the perceived quality. Video impairments can be introduced during capture, transmission, storage, and/or display, as well as by any image processing algorithm (e.g., compression) that may be applied along the way. Most impairments have more than one perceptual feature, but it is possible to produce impairments that are relatively pure. We use the term artifacts to refer to the perceptual features of impairments and artifact signal to refer to the physical signal that produces the artifact. Examples of artifacts introduced by digital video systems are blurriness, noisiness, ringing, and blockiness.[1,2]

There is an ongoing effort to develop video quality models that can detect impairments and estimate their annoyance as perceived by human viewers.[3] Most successful video quality models are full reference. These models estimate the quality of a video by comparing original and impaired videos.[4–6] Requiring the reference video or even limited information about it becomes a serious impediment in many real-time applications. In such cases, it becomes essential to develop ways of blindly estimating the quality of a video by using a no-reference (NR) video quality model.

Although human observers can usually assess the quality of a video without using the reference, creating a model that can provide the same is a difficult task. One possible approach is designing algorithms for estimating the strength of individual artifacts and then combining the artifact strengths to obtain an overall annoyance model. The assumption here is that, instead of trying to detect and estimate the strength of an "unknown" impairment that consists of a combination of artifacts, it is easier to detect individual artifact signals and estimate their strength, because we know their appearance and the type of process which generates them.[7–10]

To design a NR model using this approach, it is important to find a model that describes how the individual artifact measurements (signal strengths) can be combined to provide the overall annoyance or quality. We believe an extensive study of the most relevant artifacts is still necessary, since we still do not have a good understanding of how artifacts depend on the physical properties of the video and how they combine to produce the overall annoyance. Unfortunately, little work has been done on studying and characterizing the individual artifacts,[11–14] as Moorthy and Bovik pointed out recently.[15]

In this paper, we study the characteristics of four different types of artifacts commonly found in compressed videos (blockiness, blurriness, ringing, and noisiness). We are interested in the relationship between the perceptual strength of these individual artifact signals and their overall annoyance. To this end, we created artifact signals that are predominantly perceived as blocky, blurry, ringing, and noisy and combined them in various proportions. Then, we performed two psychophysical experiments to independently measure the strength and the overall annoyance of these artifact signals

Paper 12104 received Mar. 17, 2012; revised manuscript received Jul. 2, 2012; accepted for publication Sep. 18, 2012; published online Nov. 19, 2012.

when presented alone or in combination. We analyzed the data from the experiments and propose models for the overall annoyance based on combinations of the perceptual strengths of the individual artifact signals.

This paper is divided as follows. In Sec. 2, we describe the algorithms used to generate blockiness, blurriness, ringing, and noisiness artifacts. In Sec. 3, we describe the psychophysical experiment methodology, which includes the type of equipment used in the test, the tasks, and choice of the test sequences used in the experiment. In Secs. 4 and 5, we describe the experiments performed and discuss their results. Finally, in Sec. 6, we give the conclusions.

2 Generation of Test Sequences

Our goal is to study and characterize artifacts present in digitally compressed or processed videos. More specifically, we want to find a perceptual model that describes how the physical characteristics of individual artifacts (e.g., strengths) determine the overall annoyance or quality of the video. To achieve this goal, it is necessary to perform a set of psychophysical experiments using a set of test sequences with several combinations of individual artifacts at different proportions. Unfortunately, it is not easy to find videos with a good distribution of artifacts at different proportions from real applications. A compression algorithm, for example, is known to introduce a specific set of artifacts in a nonuniform way, according to the video content. Therefore, different videos at the same bitrate may contain artifacts at different strengths and proportions. Also, for the same video, a particular artifact may be perceptually more important at a given bitrate, while another artifact may be perceptually more important at another bitrate.[16]

To obtain a good understanding of the characteristics of various individual artifacts (and attributes), their mutual interference, and their interference with the content of the image material, the ITU-T Recommendation P.930 proposes an adjustable video reference system that generates synthetic artifacts that look like "real" artifacts yet are simpler, purer, and easier to describe.[2,13] Synthetic artifacts offer advantages for experimental research on video quality, because they make it possible to control the amplitude, distribution, and mixture of different types of artifacts, making it possible to study the different types of artifacts. Recommendation P.930 gives definitions of different types of artifacts and descriptions of algorithms for generating them synthetically. According to it, the created synthetic artifacts must be relatively pure and easily adjusted, and they must be combined to match the appearance of the full range of compression impairments. Also, the algorithms for generating them must be well defined in a way that the artifacts can be easily reproduced. In this work, we add the condition that the synthetic artifacts must produce psychometric and annoyance functions similar to those of compression artifacts.

In this paper, we use a previously developed system for generating synthetic artifacts based on the algorithms described in Recommendation P.930.[2,17] With this system, we can create artifacts synthetically and use linear scaling to control their signal strength. The method allows us to combine as many or as few artifacts as needed at several strengths. This way, we can control both the appearance and the strength of the artifacts in order to measure the psychophysical characteristics of each type of artifact signal separately or in combination.[13,18] The artifact signals are created so that their combination roughly matches "real" artifacts in terms of appearance, mean, and variance of their luminance distributions.[19] The implemented set of artifacts is composed of four artifacts (blurriness, noisiness, blockiness, and ringing) considered the most salient in digital videos.[20] The set is not extensive, and in practice, further variations of each type of artifact may occur. Nevertheless, restricting the number of artifacts to four is necessary, because experiments that estimate annoyance and visibility of artifacts require a large amount of data: a reasonable number of originals and about six strength levels for each artifact.

In the following sections, we briefly describe the algorithms used for the creation of these four type of artifacts (blockiness, blurriness, ringing, and noisiness) and the techniques used to combine them in order to create realistic degradations. More details about these algorithms can be found in our previous works.[13,17,18]

2.1 *Blockiness*

Blockiness is the appearance of the underlying block encoding structure of typical compression schemes often caused by coarse quantization of the spatial frequency components during the encoding process.[1,2] The algorithm for generating blockiness takes into account not only the average of the block of pixels, but also the average of the surrounding blocks. The first step of the algorithm is to calculate the average of each 8×8 block of the frame and the average of the 24×24 surrounding block (with the current 8×8 block as its center). Next, the difference D between these two averages is calculated. Then, to each block of the original frame, we add the corresponding element of the difference matrix D:

$$Y(i,j) = X_0(i,j) + n \cdot D(i,j), \tag{1}$$

where X_0 is the original frame, Y is the frame with blockiness, n is a constant number, and i and j denote the spatial position of the pixel in the frame. The values of $D(i,j)$ are limited to keep the pixels from becoming saturated. Before adding the blockiness, the average of the frame is adjusted to keep the artifacts from becoming more visible than intended. To adjust the average, we calculate the average of the frame before and after introducing the artifacts. Then, we subtract both averages and add the difference to all pixels in the frame. The algorithm for generating synthetic blockiness can be easily modified to use different block sizes and to include spatial shifts.

2.2 *Blurriness*

Blurriness is characterized by the loss of spatial details and a reduction in sharpness.[1,2] Recommendation P.930 suggests the generation of blurriness with the use of a simple lowpass filter.[2] To control the amount of blurriness, we can use different sizes of filters with different cutoff frequencies. Using a big range of filters dramatically increases the "types" of blurriness. Since we want to study four types of artifacts, it is not possible to also study different types of each artifact due to the limit on the number of videos that can be shown in a single experiment. For this reason, in this work, we used only a simple 5×5 moving average filter to generate blurriness.

2.3 *Noisiness*

Noisiness or noise is defined as random pixel intensity fluctuations that are not part of the original video image.[1] The algorithm used here for creating noisiness is similar to the one proposed in Recommendation P.930 and consists of replacing the luminance value of pixels at random locations with a constrained random value (Gaussian distribution, zero mean, and variance equal to one).[2] We change the range of luminance values used by the recommendation to the range [10,120] to avoid making the artifact more visible than intended. The ratio of impaired to nonimpaired number of pixels in the frame is set to 0.1. The bigger this ratio is, the higher the level of noisiness presented in the video frame.

2.4 *Ringing*

Ringing occurs when the quantization of individual coefficients of a given transform (discrete cosine transform, wavelet, Fourier, etc.) results in high-frequency irregularities of the reconstructed block.[1] Recommendation P.930 suggests generating the ringing artifact by using a filter with ripples in the passband amplitude response.[2] The problem with this approach is that, besides generating ringing, it also introduces blurriness and possibly noisiness. Our algorithm for generating ringing consists of a pair of delay-complementary filters related through

$$H(z) + G(z) = \rho \cdot z^{-n_0}, \tag{2}$$

where $H(z)$ and $G(z)$ are the transfer functions of N-tap highpass and lowpass FIR filters, respectively.[11] For $\rho = 1$ and $n_0 = 0$, the output of our system in the z-domain is given by

$$Y(z) = [H(z) + G(z)] \cdot X_0(z). \tag{3}$$

Thus, except for a shift, Y is equal to $X_0(z)$, 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. Since ringing is visible only around edges, the algorithm is applied only to the pixels of the video corresponding to edges in both the horizontal and vertical directions. The resulting effect is very similar to the ringing artifact found in compressed images, but without any blurriness or noisiness.

2.5 *Combination of Individual Artifacts*

Although the majority of psychophysical studies vary the defect strength by changing the bitrate and/or the codec implementation,[21–23] some studies have controlled the defect strength by changing its amplitude linearly.[18,24] Libert, Fenimore, and Roitman[24] compared the two methods and concluded that linear scaling can validly approximate the changes produced by varying the MPEG-2 bitrate goal. We used linear scaling to control the signal strength of our artifacts, because it allowed us to combine as many or as few artifacts as needed at several strengths.

Besides the scaling method, we use spatial and temporal binary masks to restrict artifacts to an isolated region (defect zone) of the video clip for a short time interval.[18] The degraded videos with a combination of artifact signals can be generated separately and added later to the defect zones. In this work, the defect zones correspond to central rectangular strips (horizontal or vertical) taking approximately one third of the frame. They are one second long and do not occur during the first and last seconds of the video. The use of defect zones allows us to test the interaction between the content and the artifacts.

The algorithm for generating test sequences consists of the following steps. First, we generate videos with one type of artifact signal at a relatively high level of annoyance. This distorted video can be expressed mathematically as a sum of the original sequence X_0 and the artifact signal E_l:

$$X_l(i, j, k) = X_0(i, j, k) + E_l(i, j, k), \tag{4}$$

where the index l refers to the type of artifact signal being introduced, k is the frame number, and i and j are the spatial coordinates. Therefore, the artifact signal is the difference of the video degraded with that particular artifact from the original:

$$E_l(i, j, k) = X_l(i, j, k) - X_0(i, j, k). \tag{5}$$

The settings for obtaining the maximum level of annoyance for each artifact were obtained at a previous experiment, where we matched the strength of artifact signals to perceptual strengths of real digital video artifacts.[25]

The test sequences (Y) are generated by combining the original video linearly with the artifact signals in different proportions. To create a test sequence Y with up to L artifacts, we used the following expression:

$$Y(i, j, k) = X_0(i, j, k) + \sum_{l=1}^{L} r_l \cdot E_l(i, j, k), \tag{6}$$

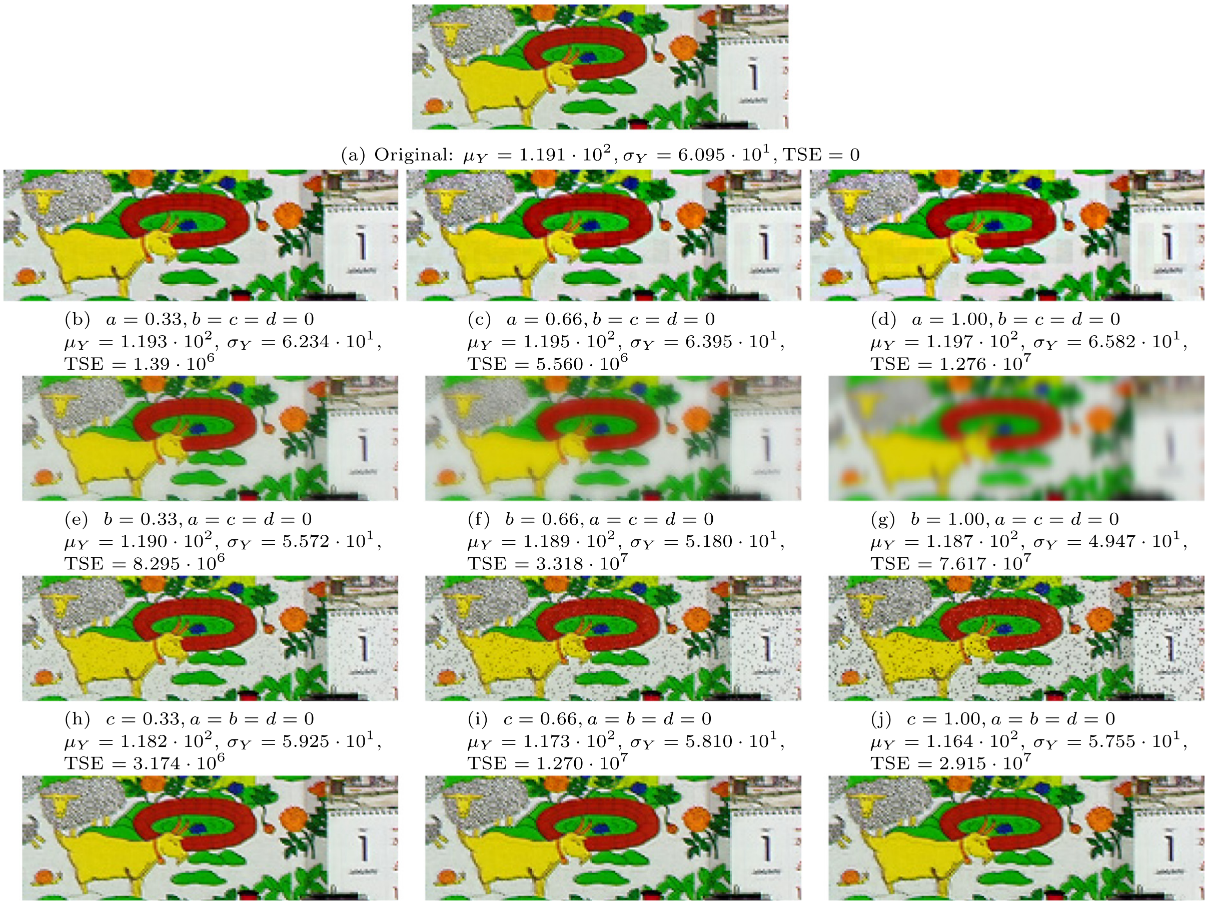
where $X_0(i, j, k)$ is the original video, and r_l $(0 \le r_l \le 1)$ is the relative strength parameter corresponding to the l-th artifact signal. In general, $\sum_l r_l = 1$, but, in some cases, we allowed $\sum_l r_l \ge 1$, making the artifact signal strength stronger.

For notation simplification, we use the coefficients a, b, c, and d (instead of r_1, r_2, r_3, and r_4) in our experiments to refer to the relative strengths of blockiness, blurriness, noisiness, and ringing artifact signals, respectively. To generate a sequence with combinations of blockiness, blurriness, noisiness, and ringing artifact signals, we use

$$\begin{aligned} Y(i, j, k) = X_0(i, j, k) + a \cdot E_{bk}(i, j, k) + \\ b \cdot E_{br}(i, j, k) + c \cdot E_{ns}(i, j, k) + d \cdot E_{rg}(i, j, k), \end{aligned} \tag{7}$$

where E_{bk}, E_{br}, E_{ns}, and E_{rg} are the blockiness, blurriness, noisiness, and ringing artifact signals, respectively.

Next, we examine the appearance and some characteristics of the artifacts generated using this technique. In Figs. 1 and 2, we show examples of the synthetic artifacts (by themselves) in the 50th frame of the videos "Calendar" and "Hockey," respectively. The images correspond with a zoomed area of the frames. The image in the first row corresponds with the original frame, while the second, third, and fourth rows correspond with frames with blockiness, blurriness, noisiness, and ringing artifacts with strength values of 0.33, 0.66, and 0.66, respectively. Below each image, we indicate the average (μ_Y) and standard deviation (σ_Y) of the intensity values of the frames, along with the total squared



(a) Original: $\mu_Y = 1.191 \cdot 10^2, \sigma_Y = 6.095 \cdot 10^1$, TSE $= 0$

(b) $a = 0.33, b = c = d = 0$ $\mu_Y = 1.193 \cdot 10^2$, $\sigma_Y = 6.234 \cdot 10^1$, TSE $= 1.39 \cdot 10^6$

(c) $a = 0.66, b = c = d = 0$ $\mu_Y = 1.195 \cdot 10^2$, $\sigma_Y = 6.395 \cdot 10^1$, TSE $= 5.560 \cdot 10^6$

(d) $a = 1.00, b = c = d = 0$ $\mu_Y = 1.197 \cdot 10^2$, $\sigma_Y = 6.582 \cdot 10^1$, TSE $= 1.276 \cdot 10^7$

(e) $b = 0.33, a = c = d = 0$ $\mu_Y = 1.190 \cdot 10^2$, $\sigma_Y = 5.572 \cdot 10^1$, TSE $= 8.295 \cdot 10^6$

(f) $b = 0.66, a = c = d = 0$ $\mu_Y = 1.189 \cdot 10^2$, $\sigma_Y = 5.180 \cdot 10^1$, TSE $= 3.318 \cdot 10^7$

(g) $b = 1.00, a = c = d = 0$ $\mu_Y = 1.187 \cdot 10^2$, $\sigma_Y = 4.947 \cdot 10^1$, TSE $= 7.617 \cdot 10^7$

(h) $c = 0.33, a = b = d = 0$ $\mu_Y = 1.182 \cdot 10^2$, $\sigma_Y = 5.925 \cdot 10^1$, TSE $= 3.174 \cdot 10^6$

(i) $c = 0.66, a = b = d = 0$ $\mu_Y = 1.173 \cdot 10^2$, $\sigma_Y = 5.810 \cdot 10^1$, TSE $= 1.270 \cdot 10^7$

(j) $c = 1.00, a = b = d = 0$ $\mu_Y = 1.164 \cdot 10^2$, $\sigma_Y = 5.755 \cdot 10^1$, TSE $= 2.915 \cdot 10^7$

(k) $d = 0.33, a = b = c = 0$ $\mu_Y = 1.180 \cdot 10^2$, $\sigma_Y = 6.053 \cdot 10^1$, TSE $= 4.128 \cdot 10^5$

(l) $d = 0.66, a = b = c = 0$ $\mu_Y = 1.169 \cdot 10^2$, $\sigma_Y = 6.016 \cdot 10^1$, TSE $= 1.651 \cdot 10^6$

(m) $d = 1.00, a = b = c = 0$ $\mu_Y = 1.157 \cdot 10^2$, $\sigma_Y = 5.984 \cdot 10^1$, TSE $= 3.791 \cdot 10^6$

Fig. 1 Zoom of the 50th frame of the video "Calendar" with blockiness, blurriness, noisiness, and ringing artifacts at different strengths (a, b, c, d). For each sample, the average (μ_Y) and standard deviation (σ_Y) of the luminance component of the frame is given, along with the total squared error (TSE) in relation to the original frame.

error (TSE) in relation to the original frame. As expected, the TSE values increase with the strength value but are dependent on the type of artifact. Blurriness is the artifact with the highest TSE values, followed by noisiness, blockiness, and ringing. Concerning the average intensity, when we compare the μ_Y values with the μ_Y values obtained for the original frame, we notice that the differences are bigger for noisiness and ringing.

It is important to point out that, in the experiments, the artifacts are added only to a defect zone. This is implemented using a binary mask $M(i,j,k)$ that has values equal to 1 for pixels inside the defect zone and 0 for pixels outside it. The final test sequences are given as

$$\tilde{Y}(i,j,k) = \begin{cases} Y(i,j,k), & \text{if } M(i,j,k) = 1 \\ X_0(i,j,k), & \text{if } M(i,j,k) = 0 \end{cases}. \tag{8}$$

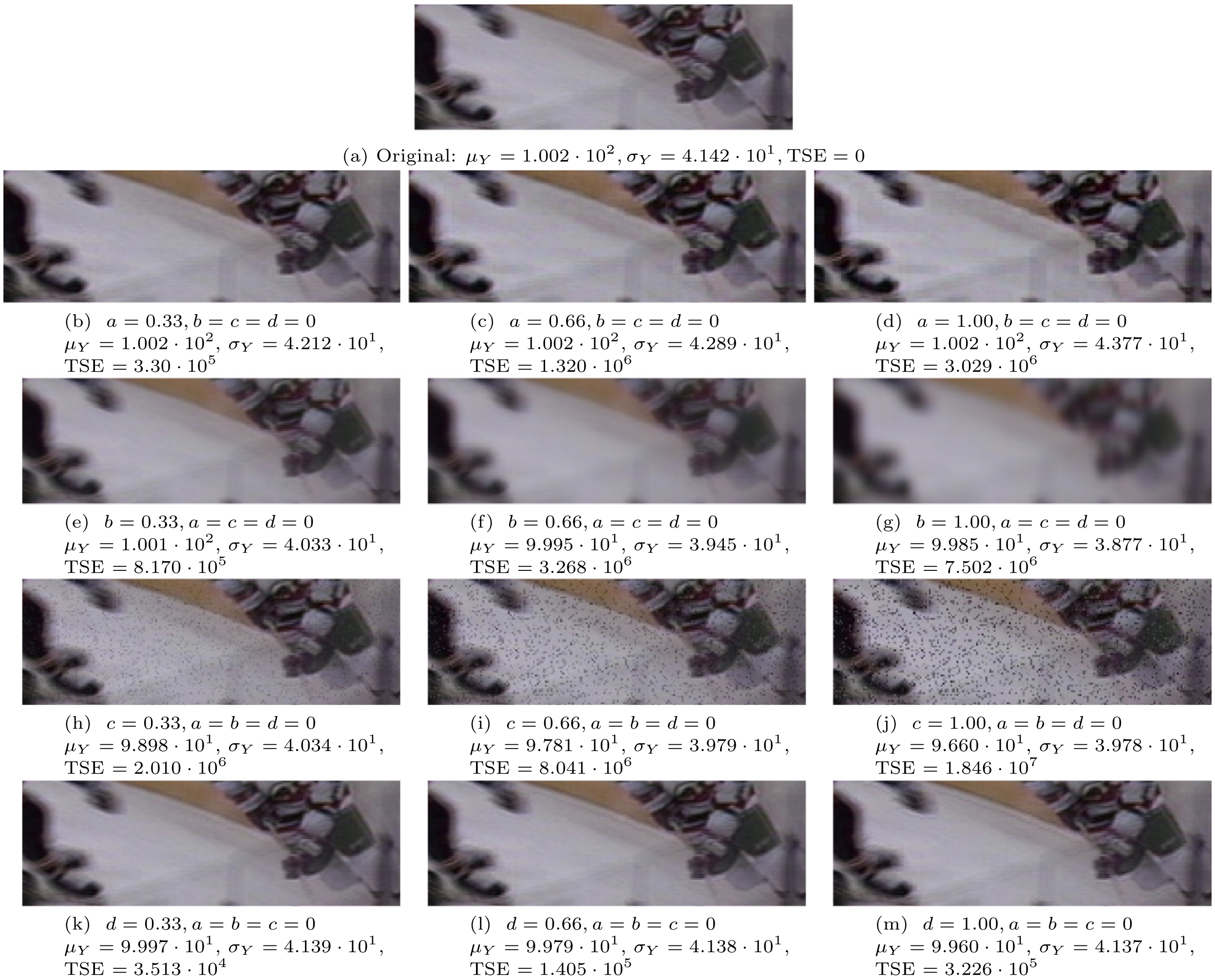
After the artifact signals are added, the borders of the defect zones are faded to avoid increasing their visibility. More details on the methodology used for generating the artifacts and combining them can be found in a previous work.[17]

3 Psychophysical Experiment

In this section, we describe the physical conditions of the psychophysical experiment and the experimental methodology used.

3.1 *Apparatus and Physical Conditions*

The apparatus used in the experiments consists of a computer, a broadcast video monitor, a computer monitor, a keyboard, and a mouse. The test video sequences are stored in the hard disk of an NEC server (PC computer) and are displayed using a subset of the PC cards normally provided with the Tektronix PQA-200 picture quality analyzer. A generator card is used to store the video locally and stream it out in a serial digital (SDI) component format. The analog output is displayed on a 14-inch Sony PVM-1343 monitor. A special-purpose program in Visual C++ is used to run the experiment and record the subject's data. After each test sequence is shown to the subject, this program displays a series of questions on the computer monitor and records the subject's responses in a data file. This configuration ensures that no further impairments are added to the videos while they are being displayed.



(a) Original: $\mu_Y = 1.002 \cdot 10^2, \sigma_Y = 4.142 \cdot 10^1$, TSE $= 0$

(b) $a = 0.33, b = c = d = 0$ $\mu_Y = 1.002 \cdot 10^2$, $\sigma_Y = 4.212 \cdot 10^1$, TSE $= 3.30 \cdot 10^5$

(c) $a = 0.66, b = c = d = 0$ $\mu_Y = 1.002 \cdot 10^2$, $\sigma_Y = 4.289 \cdot 10^1$, TSE $= 1.320 \cdot 10^6$

(d) $a = 1.00, b = c = d = 0$ $\mu_Y = 1.002 \cdot 10^2$, $\sigma_Y = 4.377 \cdot 10^1$, TSE $= 3.029 \cdot 10^6$

(e) $b = 0.33, a = c = d = 0$ $\mu_Y = 1.001 \cdot 10^2$, $\sigma_Y = 4.033 \cdot 10^1$, TSE $= 8.170 \cdot 10^5$

(f) $b = 0.66, a = c = d = 0$ $\mu_Y = 9.995 \cdot 10^1$, $\sigma_Y = 3.945 \cdot 10^1$, TSE $= 3.268 \cdot 10^6$

(g) $b = 1.00, a = c = d = 0$ $\mu_Y = 9.985 \cdot 10^1$, $\sigma_Y = 3.877 \cdot 10^1$, TSE $= 7.502 \cdot 10^6$

(h) $c = 0.33, a = b = d = 0$ $\mu_Y = 9.898 \cdot 10^1$, $\sigma_Y = 4.034 \cdot 10^1$, TSE $= 2.010 \cdot 10^6$

(i) $c = 0.66, a = b = d = 0$ $\mu_Y = 9.781 \cdot 10^1$, $\sigma_Y = 3.979 \cdot 10^1$, TSE $= 8.041 \cdot 10^6$

(j) $c = 1.00, a = b = d = 0$ $\mu_Y = 9.660 \cdot 10^1$, $\sigma_Y = 3.978 \cdot 10^1$, TSE $= 1.846 \cdot 10^7$

(k) $d = 0.33, a = b = c = 0$ $\mu_Y = 9.997 \cdot 10^1$, $\sigma_Y = 4.139 \cdot 10^1$, TSE $= 3.513 \cdot 10^4$

(l) $d = 0.66, a = b = c = 0$ $\mu_Y = 9.979 \cdot 10^1$, $\sigma_Y = 4.138 \cdot 10^1$, TSE $= 1.405 \cdot 10^5$

(m) $d = 1.00, a = b = c = 0$ $\mu_Y = 9.960 \cdot 10^1$, $\sigma_Y = 4.137 \cdot 10^1$, TSE $= 3.226 \cdot 10^5$

Fig. 2 Zoom of the 50th frame of the video “Hockey” with blockiness, blurriness, noisiness, and ringing artifacts at different strengths (a, b, c, d). For each sample, the average (μ_Y) and standard deviation (σ_Y) of the luminance component of the frame is given, along with the total squared error (TSE) in relation to the original frame.

The experiments were performed in a reserved room in the Department of Psychology of University of California Santa Barbara (UCSB). The lights of the room were dimmed, so that no light was reflected on the monitor. The measured luminance of the monitor when no signal was presented was approximately 25 $\mathrm{cd/m^2}$, while the measured luminance for the maximum output of the monitor was 225 $\mathrm{cd/m^2}$. The estimated γ of the Sony monitor was approximately 1.4 for the luminance and 1.6 for each of the independent colors (R, G, B).

Our subjects were drawn from a pool of students in an introductory psychology course. They were considered naïve of most kinds of digital video defects and the associated terminology. No vision test was performed on the subjects, but they were asked to wear glasses or contact lenses if they need them to watch TV. To guarantee robust results, at least 22 subjects were used in each experiment.[26] The experiments were run with one subject at a time. The subject was seated straight ahead of the monitor, centered at or slightly below eye height for most subjects, with the keyboard and mouse in easy reach. The computer monitor was located to one side of the subject. The distance between the subject's eyes and the video monitor was of four video monitor screen heights. The video monitor is 20 cm tall, resulting in a viewing distance of 80 cm. Four screen heights is a conservative estimate of the viewing distance, according to Recommendation BT.500.[23]

The duration of an experiment trial was limited to no more than 40 min to reduce fatigue effects on the human subjects.[23] Since only 100 to 125 sequences can be shown during a 40-min test session, and the experiments required that several defect strengths were tested, the total number of originals used for each experiment was kept low. A total of five videos of assumed high quality were used in this work: “Bus,” “Calendar,” “Cheerleader,” “Football,” and “Hockey.” These videos were all five seconds long and were in ITU-R BT.601 format (formerly CCIR-601), i.e., the videos are 60 Hz (NTSC), 4∶2∶0 YCrCb format, 486 lines by 720 columns. Representative frames of these videos used are shown in Fig. 3.

This set of videos consisted of common broadcasting scenes, with varied content that included scenes with slow and fast motion; textures, edges, and uniform areas; people and objects; and high and low contrast. Since the visibility

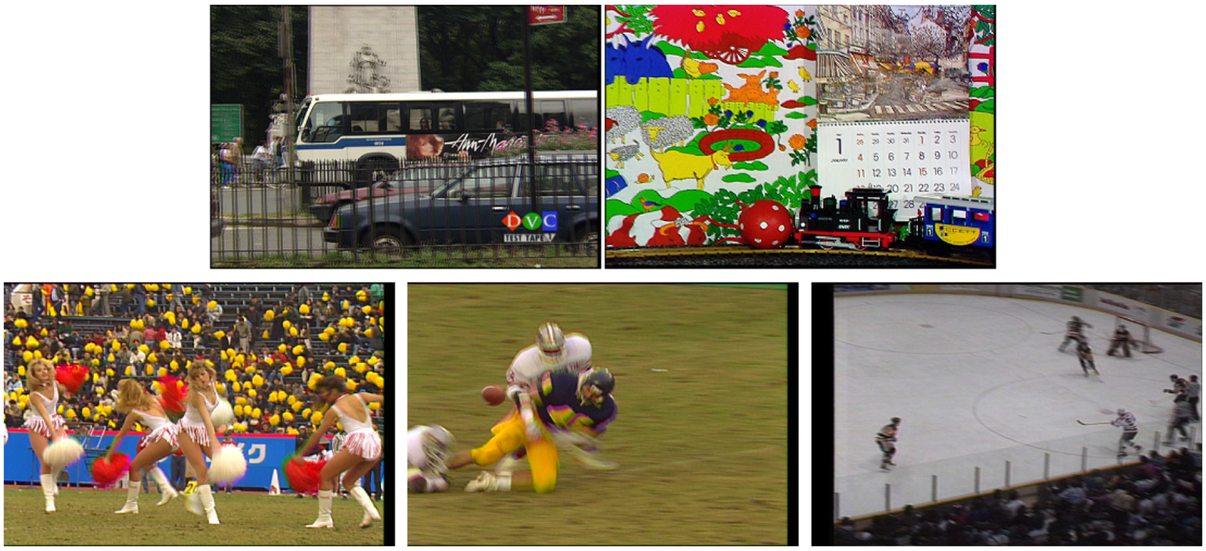


Fig. 3 Sample frame of original videos "Bus," "Calendar," "Cheerleaders," "Football," and "Hockey."

and annoyance of each type of artifact are content dependent, the choice of the scenes ensured that each of the different types of artifacts is visible in at least two videos. For example, ringing is only visible when the scene contains edges, borders, or text in a uniform background. To test ringing, we included the sequences "Calendar" and "Bus." Similarly, blockiness and noisiness are more visible in uniform areas, which are present in "Hockey" and "Calendar." Blurriness, on the other hand, is more visible in sequences with lots of details, like "Cheerleader," "Football," and "Bus."

3.2 *Methodology*

A experimental (test) session was broken into the following five stages:

1. oral instructions
2. training
3. practice trials
4. experimental trials, and
5. interview.

Prior to each experiment, a script was elaborated to help the experimenter perform the experiment. The script contained details of what the experimenter should do at each step of the experiment. More importantly, the script contained oral instructions to be given to the subject to make sure she/he understood the task to be performed. Before giving the oral instructions, the experimenter needed to make sure the subject was properly seated at the adequate distance. The tasks to be performed in the experimental trials were then explained to the subject, who was told to disregard the content of the videos and judge only the impairments.

In every experiment, the subject was asked to perform a task which consists of entering a judgment about an impairment seen in the video. To complete this task, the subject needed to have an idea of how the different types of artifacts looked and how videos with no impairments (originals) compared with videos with strong impairments. With this goal, we included a training session in the experimental session that consisted of two parts. In the first part, we showed a set of groups of videos, each with a different artifact type. Before each group was displayed, the subject was told the name of the type of artifact and given a brief description of its appearance. The second part of the training consisted of displaying the original videos, followed by examples of videos with the strongest impairments found in the experiment. The subjects were instructed to watch these videos carefully and assign a maximum value of 100 to the worst or strongest impairments in this subset.

The initial judgments of a test subject are generally erratic. It takes time for a subject to get used to the task of judging/detecting impairments. The ITU recommendation suggests throwing away the first 5 to 10 trials.[23] In our methodology, instead of discarding the first trials, we included practice trials. Before beginning this stage, subjects were told that this was a practice stage, and that no data would be recorded. Besides eliminating erratic answers, the practice trials had other benefits. It exposed the subjects to sequences with a good range of impairments and gave the subjects a chance to try out the data entry procedure. They also allowed subjects to gain confidence in their judgments. In this work, we used 12 practice trials, because of the complexity of the tasks being performed.

The subjective data for each experiment was gathered during the experimental trials. This stage was performed with the complete set of test sequences presented in random order. For each experiment, several random-ordered lists of the test sequence were generated. The lists were used sequentially and repeated as necessary. The videos were played twice, and subjects were not allowed to go back and watch them again. Subjects were instructed to search each video for impairments and to perform a specific task. After each video was played, a pop-up window appeared on the computer monitor with a question about the video. Although all subjects watched and judged the same test sequences, half of the subjects performed detection and annoyance tasks, while the other half performed a strength task.

The detection task consisted of detecting a spatially and temporally localized impairment in a five-second video sequence. In the experimental trials, after each test sequence was played, the question “Did you see a defect or an impairment?” appeared in the computer monitor. The subject was supposed to choose a “yes” or “no” answer. The annoyance task consisted of giving a numerical judgment of how annoying/bad the detected impairment was. The most annoying videos in the training stage were to be assigned a value of 100. After each test sequence was played, the subject was instructed to enter a positive numerical value indicating how annoying the impairment was. Any defect as annoying as the worst impairments in the training stage were to be given a value of 100. Those that were half as annoying were to be given 50, those 10 percent as annoying 10, and so forth. Although the subjects were asked to enter annoyance values in the range of 0 to 100, they were told that values higher than 100 could be assigned if they thought the impairment was worse than the most annoying impairments in the training stage.

The annoyance task was always performed together with the detection task. The dialog box initially assumed that a defect had been seen. If a defect had not been seen, the subject hit “No” or used the mouse for choosing “No” for “no defect.” If a defect had been seen, the subject simply started typing in the annoyance value. After entering that data, the subject hit “return” to play the next video. The program did not move on unless either “No” or a valid annoyance value was entered. Annoyance values less than zero were not accepted, but the program did not impose any upper limit on the annoyance values. Nonnumbers were also rejected. While the subject was entering data, the computer started to load the next video sequence. After the value had been accepted and the video had completed loading, the next video was shown.

The strength task consisted of asking the subjects for an estimate of how strong or visible a set of artifacts were in the detected impairment. As mentioned earlier, this type of task required teaching the subjects how each artifact looked. In the training stage, subjects were shown a set of sequences illustrating the set of artifacts being measured. In the trials, after the video was played, subjects were asked to enter a number from 0 to 10 corresponding to the strength of that artifact or feature. If no impairments were seen, subjects were instructed not to enter any number and just click “Next” to go on to the next trial.

After the trials were complete, the test subjects were asked a few questions before they left. These questions gathered interesting information that could not be gathered during the experiment. They represented the subject’s general impression of the set of test sequences and could not be associated with specific sequences. However, they were useful in guiding the design of future experiments.

We used standard methods to analyze the annoyance judgments provided by the test subjects.[27] These methods were designed to ensure that a single annoyance scale was applied to all artifact signal combinations. The mean observer score (MOS) was our subjective measure and was calculated by averaging the scores over all observers for each video. The data gathered from subjects in the first group provided one MOS value for each test sequence—the mean annoyance values (MAVs). The data gathered from subjects in the second group provided one MOS value for each artifact and each test sequence—the mean strength values (MSVs). For simplification, the MSVs for blockiness, blurriness, noisiness, and ringing are represented by S_{bk}, S_{br}, S_{ns}, and S_{rg}, respectively.

4 Experiment I: Combinations of Blocky, Blurry, and Noisy Artifacts

In Experiment I, 23 subjects performed detection and annoyance tasks, while 30 subjects performed only strength tasks. During the instructions stage, subjects of both groups were told that the test videos might contain up to three different types of artifacts: blockiness, blurriness, and noisiness. The 30 combinations of a, b, and c values used to generate the test sequences are shown in column 2 of Table 1 [Eq. (7) with $d = 0$] . We did not use all possible combinations of the three artifact signals, because that would have made the experiment too long. Four original video sequences were used in this experiment: “Bus,” “Cheerleader,” “Football,” and “Hockey.” A total of 120 test videos were used (4 originals times 30 combinations times 1 defect region).

To give an idea of what a, b, and c values correspond with in terms of error, in Fig. 4, we show graphs of the TSE values of the videos with these combinations of artifacts. In Fig. 4(a), we show the TSE values for combinations 2 to 4, which correspond to videos with only blockiness ($b = c = 0$) at three different strengths. In Fig. 4(b), we show the TSE values for combinations 5 to 7, which correspond to videos with only blurriness ($a = c = 0$) at three different strengths. In Fig. 4(c), we show the TSE values for combinations 8 to 10, which correspond to videos with only noisiness ($a = b = 0$) at three different strengths. Finally, in Fig. 4(d), the TSE values for combinations 11 to 30 are shown. Notice that for the videos containing only one type of artifact, blurriness is the artifact with highest TSE, followed by noisiness and blockiness. Because of this, the videos with more than one type of artifact that have the highest TSE are the ones with the highest values of b [combinations 25 and 26 in Fig. 4(d)].

4.1 *Results*

The average gathered values for the mean strength values ($\overline{S}_{bk}$, $\overline{S}_{br}$, and $\overline{S}_{ns}$) and mean annoyance values ($\overline{\text{MAV}}$) over all videos are shown in columns 3 to 6 of Table 1. In Fig. 5(a) to 5(d), we show the graphs of TSE versus S_{bk}, S_{br}, S_{ns}, and MAV, respectively. From the graphs in Fig. 5(a) to 5(c), we can notice that videos with high TSEs can have a small MSV (S_{bk}, S_{br} or S_{ns}) if the corresponding artifact is not present in the video. For example, S_{bk} is small for combinations that do not contain blockiness. As a result, the graphs have a high concentration of points at the bottom region, forming a horizontal line of points. For the MAV graph, we can notice that there is a big spread in points, indicating that TSE is a poor annoyance predictor. This is in agreement with findings that TSE and other similar fidelity metrics are not good quality estimators.[28]

To have a closer look at the MSVs given by subjects, bar plots of S_{bk}, S_{br}, and S_{ns} corresponding to some samples of the combinations in Table 1 are depicted in Figs. 6 to 8, along with the confidence intervals of the measurements. Combination 1, shown in Fig. 6(a), corresponds with the original videos ($a = 0$, $b = 0$, and $c = 0$). It is interesting to notice

Table 1 Experiment I: artifact signal strength combinations and average of mean strength values ($\overline{S}_{bk}$, $\overline{S}_{br}$, and $\overline{S}_{ns}$) and mean annoyance values ($\overline{\text{MAV}}$) over all test sequences.

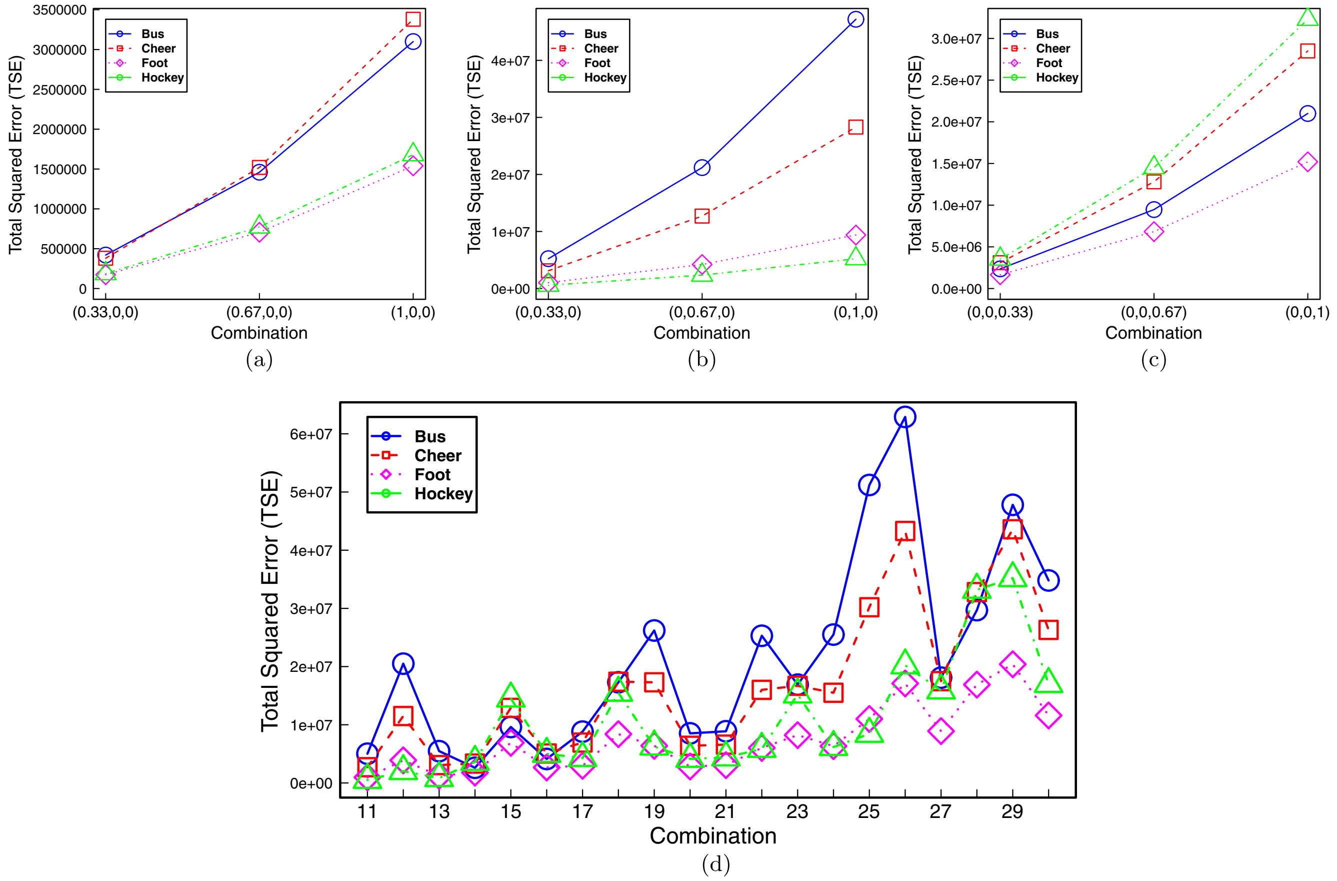
Comb.	(a,b,c)	$\overline{S}_{bk}$	$\overline{S}_{br}$	$\overline{S}_{ns}$	$\overline{\text{MAV}}$
1	(0, 0, 0)	0.05	0.22	0.05	8.16
2	(0.33, 0, 0)	1.02	0.21	0.05	5.36
3	(0.67, 0, 0)	4.06	0.40	0.08	16.61
4	(1, 0, 0)	6.51	0.33	0.09	32.93
5	(0, 0.33, 0)	0.04	1.95	0.07	9.26
6	(0, 0.67, 0)	0.09	6.59	0.02	37.53
7	(0, 1, 0)	0.15	8.6	0.12	74.76
8	(0, 0, 0.33)	0.04	0.20	3.05	15.85
9	(0, 0, 0.67)	0.04	0.26	5.37	26.51
10	(0, 0, 1)	0.07	0.16	7.26	45.12
11	(0.33, 0.33, 0)	2.10	1.65	0.02	12.69
12	(0.33, 0.67, 0)	3.88	5.51	0.11	43.87
13	(0.67, 0.33, 0)	5.07	0.84	0.06	28.03
14	(0.33, 0, 0.33)	0.25	0.19	2.92	14.82
15	(0.33, 0, 0.67)	0.02	0.10	5.18	32.7
16	(0.67, 0, 0.33)	3.44	0.30	3.75	26.41
17	(0, 0.33, 0.33)	0.10	1.94	3.76	24.46
18	(0, 0.33, 0.67)	0.07	1.33	5.86	37.71
19	(0, 0.67, 0.33)	0.24	6.22	4.25	51.38
20	(0.33, 0.33, 0.33)	0.6	0.98	3.82	21.595
21	(0.67, 0.33, 0.33)	4.78	1.02	3.53	35.94
22	(0.33, 0.67, 0.33)	1.78	4.75	4.16	53.88
23	(0.33, 0.33, 0.67)	0.14	0.91	5.60	38.37
24	(1, 0.67, 0.33)	7.71	2.49	3.08	72.11
25	(0.67, 1, 0.33)	7.02	4.49	4.03	85.17
26	(0.33, 1, 0.67)	1.97	6.70	6.05	81.64
27	(1, 0.33, 0.67)	5.76	0.92	5.43	58.52
28	(0.67, 0.33, 1)	1.63	0.55	7.30	55.01
29	(0.33, 0.67, 1)	0.17	4.31	7.10	72.63
30	(0.67, 0.67, 0.67)	5.12	2.23	5.66	68.47

that, for some videos, the values of S_{bk}, S_{br}, and S_{ns} corresponding to the originals are not zero, indicating that subjects reported that these videos contained some type of artifact signals and annoyance levels different from zero. In the video "Hockey," for example, subjects reported some blurriness. This is not very surprising, considering that this original does not look as sharp as the other originals.

The test combinations 2 to 4, 5 to 7, and 8 to 10 (see Table 1 and Fig. 6) correspond to videos with only blocky, blurry, or noisy artifact signals, respectively. For these combinations, the highest MSV was obtained for the corresponding pure artifact signals, while the other two types of artifacts received zero or much smaller values, which indicates that subjects correctly identified the artifact introduced in the video. The MSVs were highest for videos that contained blurriness, followed by noisiness and blockiness, which is the same order obtained for TSE values [see Figs. 4(a) to 4(c)]. It is interesting to notice that, for the video "Cheerleader," subjects gave smaller MSVs for blockiness and noisiness. This is probably due to the high spatial activity of this video, which makes it harder to detect these artifacts in the backround (busy audience) or in the foreground (moving cheerleaders). Although other scenes like "Bus," "Football," and "Calendar" are also very busy, they do have uniform areas that facilitate their detection.

The test combinations 11 to 13, 14 to 16, and 17 to 19 correspond to videos with two types of artifact signals: blocky-blurry, blocky-noisy, or blurry-noisy (see Fig. 7). For these combinations, the artifact signal corresponding with the highest weight in the combination received the highest MSV. Nevertheless, an increase in the artifact signal strength did not always result in a proportional increase of the corresponding MSV (S_{bk}, S_{br}, or S_{ns}). For example, for the blocky-blurry artifact signals, an increase in the weight of the blurry artifact signal caused an increase in the perceived strength of not only blurriness, but also blockiness (as can be seen when comparing combinations 11 and 12 in Fig. 7). This was especially true for the sequences "Bus" and "Football," which are sequences with the highest visibility of blockiness [see Fig. 4(b)]. Therefore, if blockiness is visible, its strength seems to be accentuated by an increase in blurriness. On the other hand, increasing the weight of blockiness did not have the same effect on the perceived strength of blurriness (see combinations 11 and 13 in Fig. 7). In the case of blocky-noisy and blurry-noisy artifact signals, the presence of noisiness decreased the perceived strength of the other two artifacts, especially for higher weights of noisiness (see combinations 15, 16, and 18 in Fig. 7). The reason for this is probably that noisiness is easily detectable in higher strengths and uniform areas. Therefore, when present, it causes the subject to disregard less prominent artifacts.

The test combinations 20 to 30 (see samples in Fig. 8) correspond to videos with the three types of artifact signals. Also, for these combinations, the artifact signals corresponding to the greatest weights received higher MSVs. Again, the noisy artifact signals seemed to decrease the perceived strength of the two other artifacts (compare combinations 20 and 23), while blurry artifact signals seemed to increase them (compare combinations 20 and 22). Thus, there seem to be interactions between the three artifact signals when determining the perceived strengths of the artifacts.



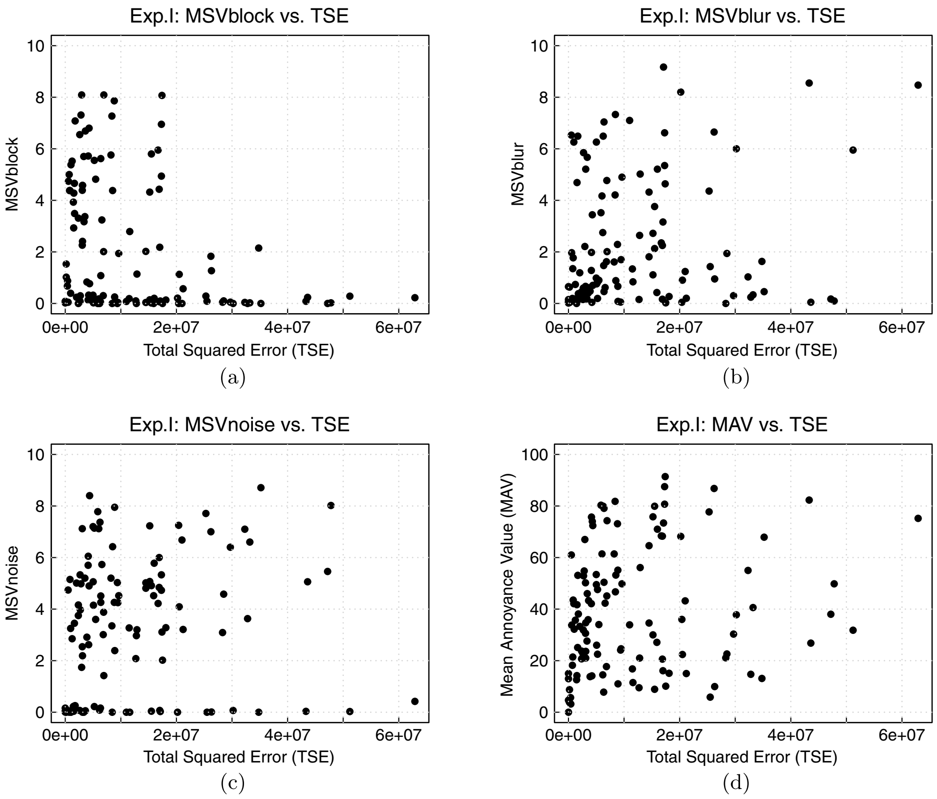
(a)

(b)

(c)

(d)

Fig. 4 Experiment I: total squared errors (TSE) for combinations: (a) 2 to 4 (only blockiness), (b) 5 to 7 (only blurriness), (c) 8 to 10 (only noisiness), and (d) 11 to 30.



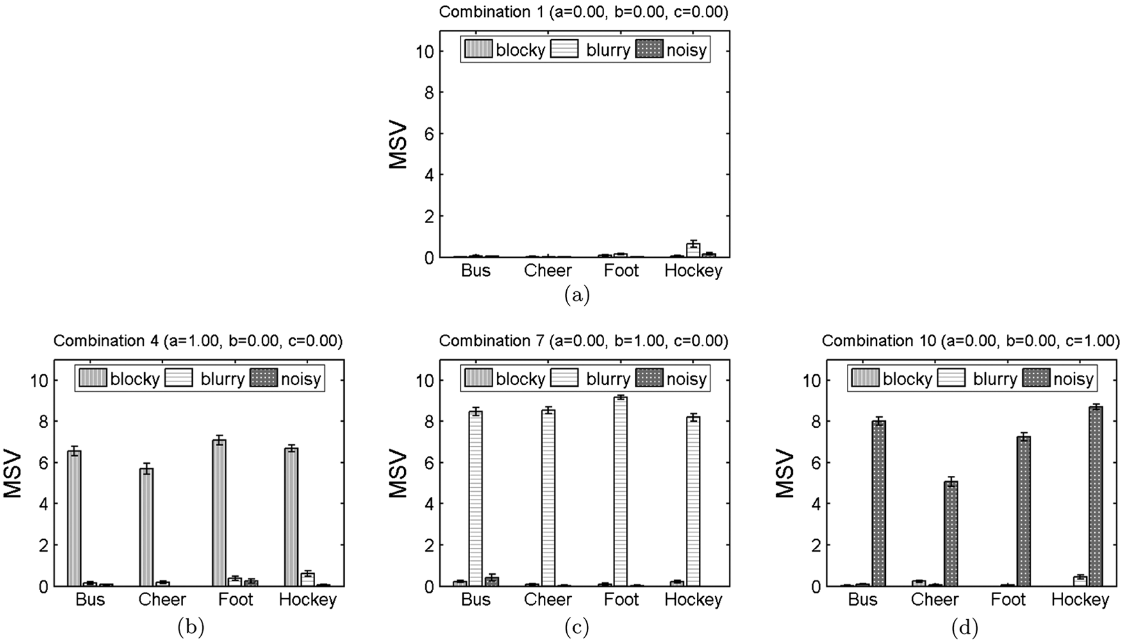
(a)

(b)

(c)

(d)

Fig. 5 Experiment I: total squared errors (TSE) versus (a) S_{bk}, (b) S_{br}, (c) S_{ns}, and (d) MAV.



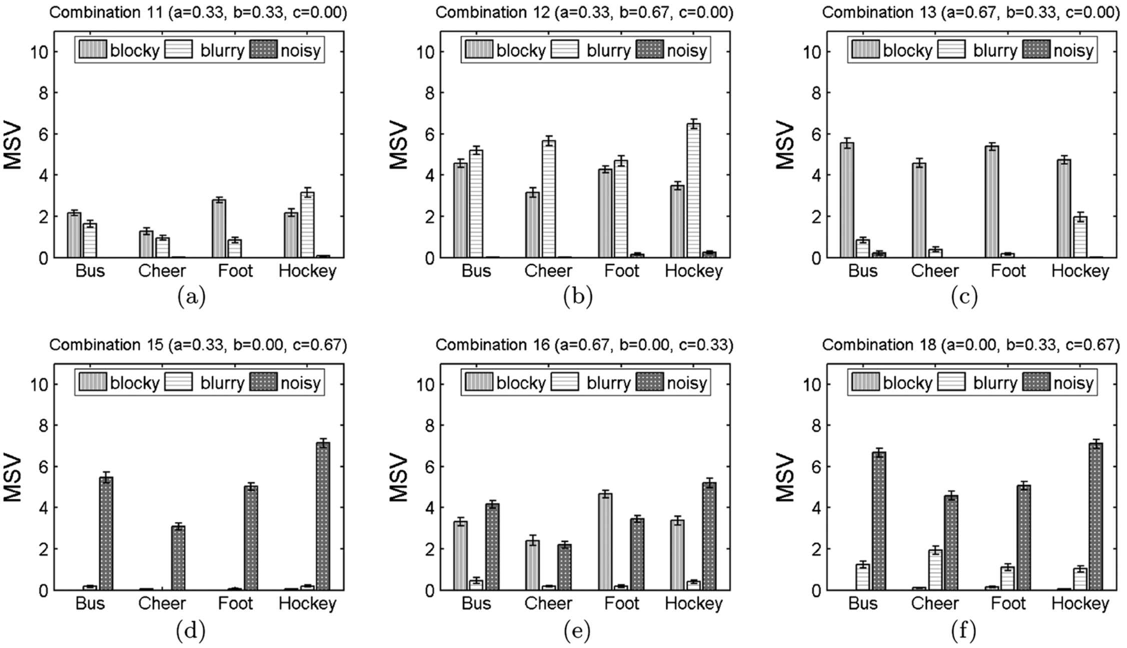
(a)

(b)

(c)

(d)

Fig. 6 Experiment I: MSV bar plots for combinations (a) 1, (b) 4, (c) 7, and (d) 10.



(a)

(b)

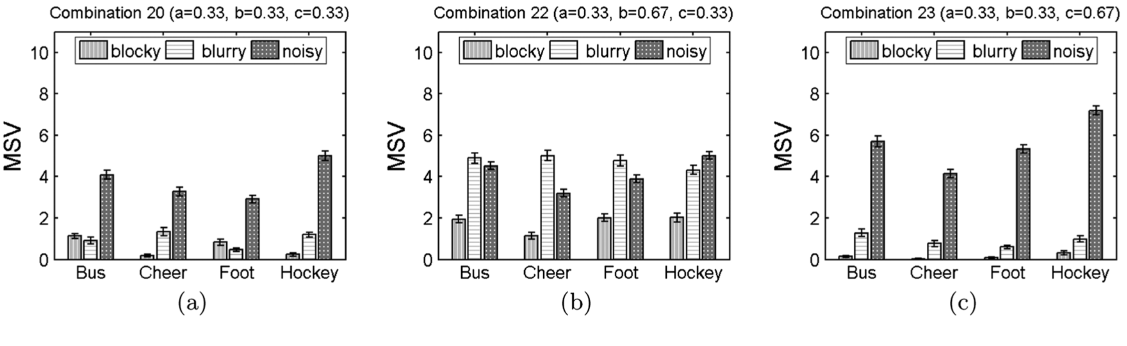
(c)

(d)

(e)

(f)

Fig. 7 Experiment I: MSV bar plots for combinations (a) 11, (b) 12, (c) 13, (d) 15, (e) 16, and (f) 18.

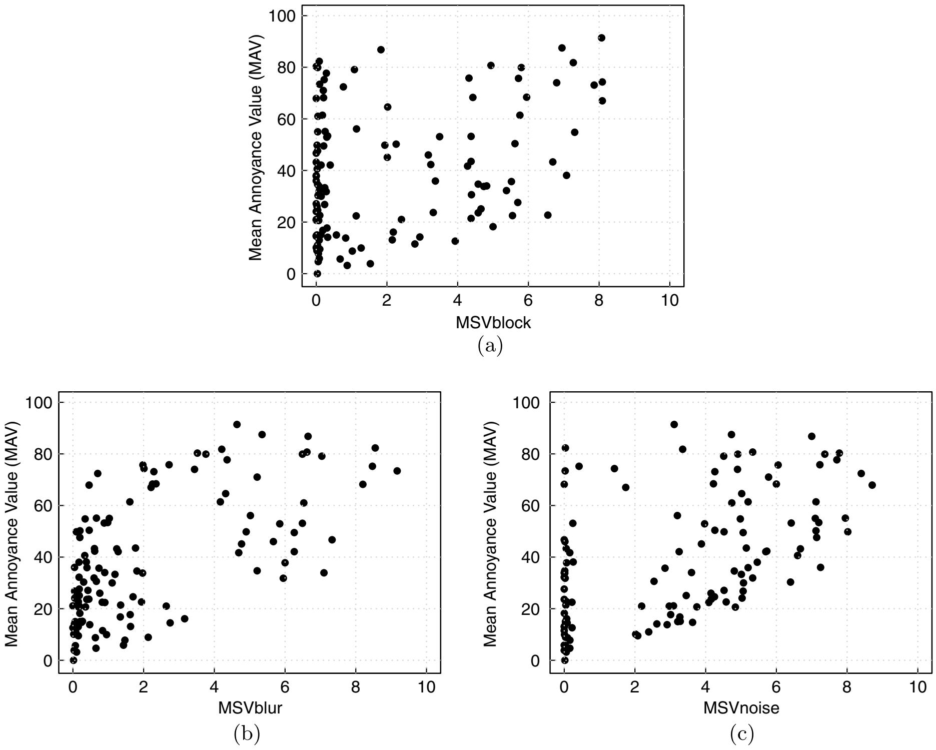


(a)

(b)

(c)

Fig. 8 Experiment I: MSV bar plots for combinations (a) 20, (b) 22, and (c) 23.



(a)

(b)

(c)

Fig. 9 Experiment I: MSV versus MAV: (a) S_{bk}, (b) S_{br}, (c) S_{ns}.

4.2 *Model Predictions*

Our principal interest in measuring the artifacts' strength is to investigate the relationship between the artifact perceptual strengths (S_{bk}, S_{br}, and S_{ns}) and the overall annoyance (MAV). The first question that needs to be answered is whether MAV can be predicted by a single artifact MSV. In Fig. 9, we show graphs of each individual MSV versus MAV. From these plots, it is hard to believe that MAV can be predicted from a single MSV corresponding to perceptual strength measurements of only one type of artifact. The points in the top part of these graphs correspond to combination indices greater than 18 that correspond to combinations of at least two artifacts and, therefore, have higher MAVs. For combinations that did not contain the artifact being estimated, the corresponding MSV value was small, which occurred in a concentration of points in the graphs within values between 0 and 1. Since annoyance cannot be modeled as a function of a single artifact, there is a chance that the annoyance model is a multidimensional function that depends on the strengths of the most "important" or relevant artifacts.

Given that a single measurement cannot predict annoyance, we tested a set of combination or multidimensional models by performing a nonlinear least-squares data fitting to the MAV and the MSVs (S_{bk}, S_{br}, and S_{ns}) using the model equation. The first model we tested was the Minkowski metric, which is given by

$$\mathrm{PA}_M = (S_{bk}^p + S_{br}^p + S_{ns}^p)^{1/p}, \tag{9}$$

where PA_M is the predicted annoyance, and p is the Minkowski power. For this fit, we found a Minkowski power equal to $p = 0.376255$ and Pearson and Spearman correlation coefficients equal to 0.7867718 and 0.8015848, respectively. This is the same combination rule used by Huib de

Table 2 Experiment I: fitting parameters for weighted Minkowski metric.

Group	p	α	β	γ	R_σ	PCC	SCC
Bus	1.0617	3.6812	7.1707	6.2468	8.997	0.939083	0.955718
Cheer	1.2087	8.9370	11.9469	6.2184	7.439	0.965283	0.964957
Football	1.0699	5.3116	7.3802	5.2427	7.854	0.943737	0.968294
Hockey	1.3025	8.7545	12.3897	12.6920	7.407	0.960720	0.965624
All	1.1726	6.3868	9.6265	7.9377	8.462	0.943426	0.958771

Ridder in a previous work that predicted annoyance caused by blockiness, blurriness, and ringing.[12] However, our results are different from the values of $p = 1.67$ and 2.37 reported by De Ridder for his annoyance model. Nevertheless, De Ridder's model is for still images, and it is tested on a smaller data set (two images: "Child" and "Girls").

To try to find a better model, we modified the Minkowski metric to include weights for each individual artifact's strength. This way, different contributions of each artifact can be estimated, as given by

$$\mathrm{PA}_{\mathrm{WM}} = (\alpha \cdot S_{bk}^{p} + \beta \cdot S_{br}^{p} + \gamma \cdot S_{ns}^{p})^{1/p}, \tag{10}$$

where $\mathrm{PA}_{\mathrm{WM}}$ is the predicted annoyance, and α, β, and γ are the weights for S_{bk}, S_{br}, and S_{ns}, respectively. In Table 2, we present the coefficients (p, α, β, and γ), residual standard deviation (R_σ), Pearson correlation coefficients (PCC), and Spearman correlation coefficients (SCC) obtained for the fits of this model to the data group corresponding to each original ("Bus," "Cheer," "Football," and "Hockey") and to the data group containing all videos ("All").

The P-values (t-test, two-tailed, $P < 0.05$) corresponding to coefficients α, β, and γ are consistent for all tested models. For the set containing all video sequences ("All"), the fit using the weighted Minkowski metric returned a Minkowski exponent $p = 1.1726$ and weight coefficients $\alpha = 6.3868$, $\beta = 9.6265$, and $\gamma = 7.9377$, corresponding to blockiness, blurriness, and noisiness, respectively. This fit is better than for the original Minkowski metric, with PCC = 0.943426 and SCC = 0.958771. Also, there is little systematic error in the predictions (see R_σ in column 6 of Table 2).

From Table 2, it can be noticed that the values found for p are all between 1.06 and 1.3. To investigate the effect of having a fixed value of p, we fit the data using Eq. (10) with p constant and within the interval [0.25, 1.50]. In Table 3, we present the results obtained for the linear case ($p = 1$). For the set containing all video sequences, the fit returned coefficients $\alpha = 4.0746$, $\beta = 6.3879$, and $\gamma = 5.3164$. The PCC and SCC of this fit are equal to 0.935995 and 0.948798, respectively.

A model comparison test indicates that there is no significant statistical difference between the weighted Minkowski and the models with $p = 1.0$, 1.13, or 1.25. This means that the results using a nonlinear model or a simpler linear model are (statistically) the same. In Figs. 10 and 11, we plotted the observed MAV versus $\mathrm{PA}_{\mathrm{WM}}$ (computed across all videos) for the Minkowski metric ($p = 1.1726$) and the linear model ($p = 1$). To differentiate both models, from now on we will refer to the linear model as $\mathrm{PA}_{\mathrm{LIN}}$.

From Tables 2 and 3, we can see that both models give more weight to blurriness, followed by noisiness and blockiness. As our results show, blurriness is the artifact with the highest individual values of MSV (and TSE), and it is the artifact with the most weight in the models tested. Although blurriness has a high impact on MAV, the weights of noisiness and blockiness are also high, indicating that they are also very significant in determining MAV.

Since we are also interested in understanding if the perceptual strengths interact with one another, we also tested a linear model with interactions, as given by

$$\begin{aligned}\mathrm{PA}_{\mathrm{LINT}} = {} & \alpha \cdot S_{bk} + \beta \cdot S_{br} + \gamma \cdot S_{ns} + \rho_1 \cdot S_{bk}S_{br} \\ & + \rho_2 \cdot S_{bk}S_{ns} + \rho_3 \cdot S_{br}S_{ns} + \tau \cdot S_{bk}S_{br}S_{ns}. \end{aligned} \tag{11}$$

The results of this fitting can be found in Table 4. Column 2 of the table shows the estimated coefficients for the model, and column 5 shows the corresponding P-values (t-test, two-tailed, $P < 0.05$). The last line of the table shows fitting correlation coefficients for the complete model. Notice that the correlation coefficients of this model are slightly higher than for the linear model with no interactions, but a model test comparison showed that the differences are not statistically significant. It can be observed that the coefficients for the main factors (α, β, and γ) are statistically significant, but the first- and second-order interactions are not statistically significant. It is important to point out that this model uses the perceived strength of the artifact and not the actual artifact signal strength. Therefore, it cannot be used to explain the results obtained in the bar plots in Figs. 6 to 8, which were based on the artifact signal strength. In the next section, we present Experiment II, which uses a factorial design to allow for a better interaction test of artifact signals.

5 Experiment II: Combinations of Blocky, Blurry, Noisy, and Ringy Artifacts

In Experiment II, the subjects were also divided into two independent groups. The first group was composed of 23 subjects who performed detection and annoyance tasks. The second group was composed of 30 subjects who performed strength tasks. As in the previous experiment, the main goal was to study the importance of the strengths of individual artifacts and to determine overall annoyance. The first difference between Experiments I and II is that the sequences in Experiment II contained combinations of

Table 3 Experiment I: fitting parameters for linear metric (weighted Minkowski with $p = 1$).

Group	α	β	γ	R_σ	PCC	SCC
Bus	3.1483	6.1581	5.4292	8.859	0.937577	0.951713
Cheer	5.1980	7.2208	3.7750	7.609	0.959483	0.949160
Football	4.4809	6.2704	4.4885	7.754	0.94183	0.965624
Hockey	3.6646	6.2264	6.2284	8.079	0.941885	0.945155
All	4.0746	6.3879	5.3164	8.655	0.935991	0.948798

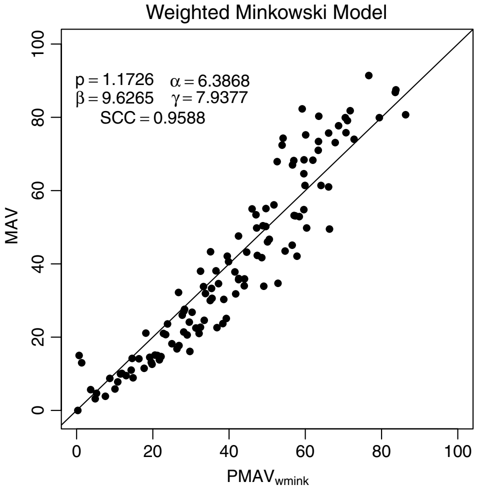


Fig. 10 Experiment I: observed MAV versus predicted MAV (PA_{WM}) using the weighted Minkowski model for the data set containing all test videos.

Table 4 Experiment I: fitting parameters for the linear model with interactions.

Coef.	Estimate	Std. error	*t* value	Pr(>|*t*|)
α	4.43538	0.49607	8.941	8.57e–15[a]
β	6.78535	0.38986	17.404	<2e–16[a]
γ	5.67310	0.30372	18.679	<2e–16[a]
$\alpha{:}\beta$	−0.29465	0.24545	−1.200	0.2325
$\alpha{:}\gamma$	−0.25081	0.17376	−1.443	0.1517
$\beta{:}\gamma$	−0.24920	0.12924	−1.928	0.0564
$\alpha{:}\beta{:}\gamma$	0.13339	0.07021	1.900	0.0600
Corr. coef.	PRCC = 0.940565		SRCC = 0.955316	

[a]Statistically significant at ($P < 0.05$).

four artifacts (blocky, blurry, noisy, and ringing), instead of the three (blocky, blurry, and noisy) used in Experiment I.

The second difference is that the set of combinations used in Experiment II includes a full factorial design[29] (combinations 1 to 16) of the four artifact signals, which enables us to identify major factors and interaction terms affecting the annoyance scores. In such a design, the levels (or strengths) of the variables are chosen in such a way that they span the complete factor space. Often, only a lower level and an upper level are chosen. In our case, we have four variables that correspond to the strengths of blocky, blurry, ringy, and noisy artifact signals (a, b, c, and d).

The 24 combinations of the parameters a, b, c and d used to generate the test sequences are shown in Table 5. As can be seen in combinations 1 to 16, only two values are possible for each artifact signal strength: 0 or 1 for ringing and blockiness, and 0 or 0.67 for blurriness and noisiness. (Ringing and blockiness are given higher upper values because these two artifacts received lower annoyance values in the previous experiments.) Combinations 17 to 24 were added as samples of typical compression proportions. Five original video sequences were used in this experiment: "Bus," "Calendar," "Cheerleader," "Football," and "Hockey." A total of 120 test videos were used in this experiment (5 originals times 24 combinations times 1 defect zone).

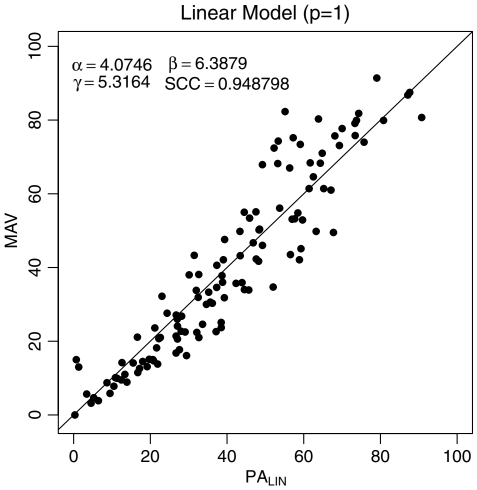


Fig. 11 Experiment I: observed MAV versus predicted MAV (PA_{LIN}) using the linear model for the data set containing all test videos.

To give an idea of what the strength values a, b, c, and d corresponded with in terms of error, we show graphs of the TSE of the videos with these combinations of artifacts in Fig. 12. The x-axis of these graphs corresponds with the combination number (according to Table 5) and the y-axis corresponds to the TSE. The graph has five curves, each corresponding to an original video. Notice that the combinations with the highest TSE values are the ones with the biggest values of a, b, and c (e.g., combination 17). The videos with only ringing ($a = b = c = 0$, combinations 22 to 24) have very small TSE values.

5.1 *Results*

The values for the average MAV and MSV for all videos are also shown in Table 5. We show the graphs of TSE versus S_{bk}, S_{br}, S_{ns}, S_{rg}, and MAV in Figs. 13(a) to 13(d) and 14. From these graphs, we can notice again that videos with high values of TSE can have a small S_{bk}, S_{br}, S_{ns} or S_{rg} if the corresponding artifact is not present in the video. Although the MAV plot still shows a large spread of points, the relationship between TSE and MAV seems to have a higher correlation than what was obtained in Experiment I. This might be due to the fact that fewer variations in artifact strengths exist in Experiment II (Table 2).

To have a closer look at the MSVs given by the subjects, in Figs. 15 to 17, we show the bar plots for S_{bk}, S_{br}, S_{ns}, and S_{rg} corresponding to some sample combinations. Combination number 1 (see Table 5) corresponds to the original videos. Again, the values for the average of MAVs and MSVs corresponding to the originals are not zero, indicating that some subjects reported that these videos contained some type of artifact and annoyance levels different from zero.

For the test combinations 2, 3, 5, and 9 (shown in Fig. 15), corresponding with videos with only one type of synthetic artifact signal (blockiness, blurriness, noisiness, or ringing),

Table 5 Experiment II: Artifact signal strength combinations and average of mean strength values ($\overline{S}_{bk}$, $\overline{S}_{br}$, $\overline{S}_{ns}$, and $\overline{S}_{rg}$) and mean annoyance values ($\overline{\mathrm{MAV}}$) over all test sequences.

	(*a*,*b*,*c*,*d*)	$\overline{S}_{bk}$	$\overline{S}_{br}$	$\overline{S}_{ns}$	$\overline{S}_{rg}$	$\overline{\mathrm{MAV}}$
1	(0, 0, 0, 0)	0.042	0.32	0.03	0.74	0.38
2	(0, 0, 0.67, 0)	0.20	0.23	5.9	0.17	35.84
3	(0, 0.67, 0, 0)	0.33	5.93	0.05	0.31	29.05
4	(0, 0.67, 0.67, 0)	0.252	4.99	6.41	0.31	62.53
5	(1, 0, 0, 0)	4.29	0.47	0.06	0.28	17.86
6	(1, 0, 0.67, 0)	1.49	0.58	5.97	0.23	43.76
7	(1, 0.67, 0, 0)	6.663	2.491	0.04	0.19	48.61
8	(1, 0.67, 0.67, 0)	4.52	2.82	6.29	0.26	67.84
9	(0, 0, 0, 1)	0.13	0.47	0.09	2.57	3.42
10	(0, 0, 0.67, 1)	0.19	0.39	6.25	0.36	38.98
11	(0, 0.67, 0, 1)	0.21	6.11	0.50	3.18	36.43
12	(0, 0.67, 0.67, 1)	0.171	4.63	6.55	0.67	64.97
13	(1, 0, 0, 1)	4.77	0.62	0.12	1.18	18.11
14	(1, 0, 0.67, 1)	1.51	0.64	6.23	0.23	45.90
15	(1, 0.67, 0, 1)	6.51	2.82	0.27	0.76	56.78
16	(1, 0.67, 0.67, 1)	4.23	2.79	6.24	0.40	70.4
17	(1, 1, 0.33, 1)	6.25	4.29	5.44	0.57	81.83
18	(1, 0.67, 0, 1)	6.61	2.92	0.17	0.97	54.14
19	(0.67, 0.67, 0, 0.67)	5.12	4.05	0.08	0.45	42.93
20	(0, 0, 0, 0.33)	0.1	0.43	0.05	0.79	0.85
21	(0, 0, 0, 0.67)	0.08	0.37	0.16	1.46	1.38
22	(0, 0, 0.1, 0)	0.03	0.33	0.25	0.59	1.66
23	(0, 0, 0.25, 0)	0.18	0.39	3.02	0.16	12.02
24	(0, 0, 0.8, 0)	0.17	0.26	6.36	0.21	41.28

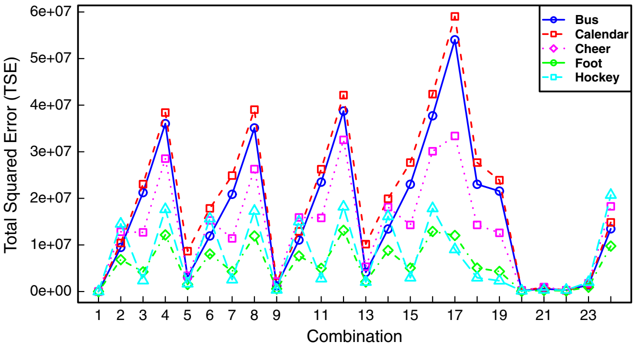


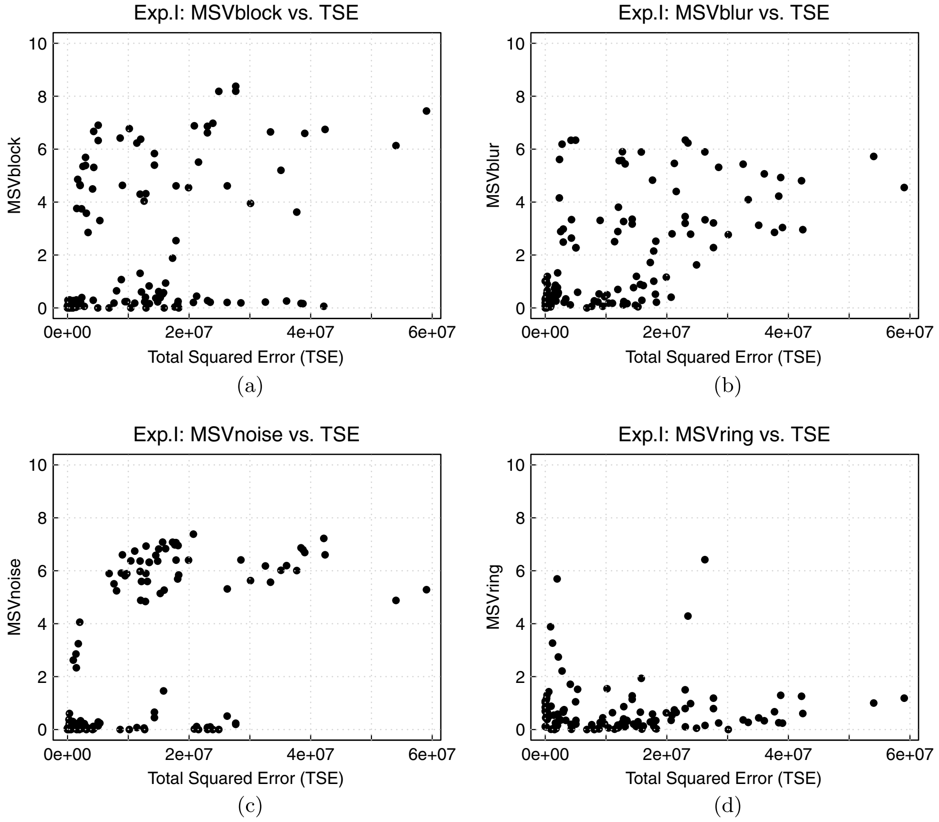
Fig. 12 Experiment II: total squared error (TSE) of all test videos.

the highest MSVs were obtained for the corresponding pure artifact, while the other three types of artifact signals received small values. MAVs are the highest for videos that contain noisy artifact signals (see the last column of Table 5). Again, given its high spatial and temporal activity, “Cheerleader” was the video for which subjects reported the lowest level of ringing and blockiness. Surprisingly, subjects were able to detect a considerable amount of ringing in this video, thanks to the relatively large uniform white areas (see cheerleaders in the foreground). Notice also that “Calendar” was the video for which subjects reported the highest levels of ringing and blockiness. This is due to the fact that this video has a good amount of borders and lines in a uniform background.

In Fig. 16, we show MSV bar plots corresponding with sample combinations containing two artifacts. Notice that for combinations 10 and 13, which contain ringing plus noisiness and blockiness, respectively, the strength corresponding to ringing is very small, although the physical signal strength of ringing is as high as that of the other artifact signal. On the other hand, for combination 11, which contains ringing plus blurriness, the strength corresponding to ringing is larger, especially for sequences “Bus” and “Calendar.” This is in agreement with the common knowledge that, for images with weak levels of degradation, ringing and blurring are two of the most relevant artifacts.[30] For combinations containing blockiness plus one more artifact, the strength for blockiness was higher when blurriness was present (combination 7) and much lower for noisiness (combination 6), while ringing (combination 13) did not seem to have a significant impact (compare it with combination 5 in Fig. 15). Overall, blurriness increased the perceived strength of both blockiness (see combinations 5 and 7) and ringing (see combinations 9 and 11). Noisiness, on the other hand, decreased the perceived strength of blockiness and ringing.

In Fig. 17, we show MSV bar plots corresponding to sample combinations containing three artifacts. Notice that, for the combinations that contain ringing plus two other artifacts (combinations 12, 14, and 15), the strength corresponding to ringing was again very low when compared with the other two artifacts. When noisiness was combined with ringing and blockiness (combination 14), the other two artifacts received very low strength scores, while noisiness received higher scores. On the other hand, when we combined noisiness, blockiness, and blurriness (combination 8), the strength score values were more comparable, with noisiness receiving the highest scores, followed by blockiness and blurriness.

Notice that, when the four artifact signals were present (combination 16), noisiness was perceived as the strongest artifact, followed by blockiness and blurriness. Given that noisiness seems to be a dominating artifact, it is worth taking a look at combination 15, which does not include noisiness. In this case, blockiness showed a significantly higher value than blurriness. These results seem to suggest that an interaction among artifacts is happening.



(a)

(b)

(c)

(d)

Fig. 13 Experiment II: total squared errors (TSE) versus (a) S_{bk}, (b) S_{br}, (c) S_{ns}, and (d) S_{rg}.

5.2 *Model Predictions*

Again, to investigate the relationship between the artifacts' perceptual strengths (S_{bk}, S_{br}, S_{ns}, and S_{rg}) and the overall MAV, we show graphs of the individual MSVs versus MAV in Fig. 18. These graphs clearly show that MAV cannot be predicted from a single MSV corresponding to perceptual strength measurements of only one type of artifact. As can be seen in all graphs in Fig. 18, for combinations that did not contain a particular artifact, the corresponding MSV value was low, which occurred in a concentration of points in the graphs with values between 0 and 1. The graph for S_{rg} [Fig. 18(d)] is interesting because it shows the low importance of the ringing artifact in predicting MAV. Most of the points in the graph are concentrated in the left part of the graph, with only a few points having an S_{rg} value greater than 2.

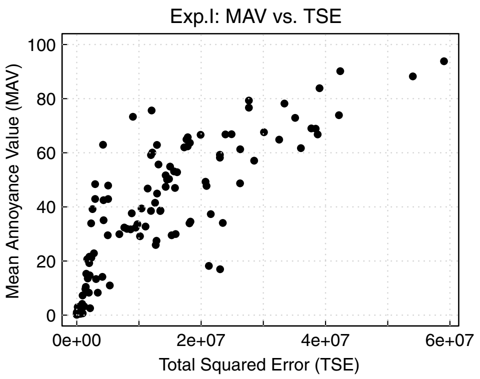


Fig. 14 Experiment II: total squared errors (TSE) versus MAV.

Given that a single measurement cannot predict annoyance, we tested a set of models by performing a nonlinear least-squares data fitting to the MAV and the MSVs of the individual artifacts (S_{bk}, S_{br}, S_{ns}, and S_{rg}) using the model equation. Again, we tested the Minkowski metric for the four artifacts, which is given by

$$\mathrm{PA}_M = (S_{bk}^p + S_{br}^p + S_{ns}^p + S_{\mathrm{ring}}^p)^{1/p}. \qquad (12)$$

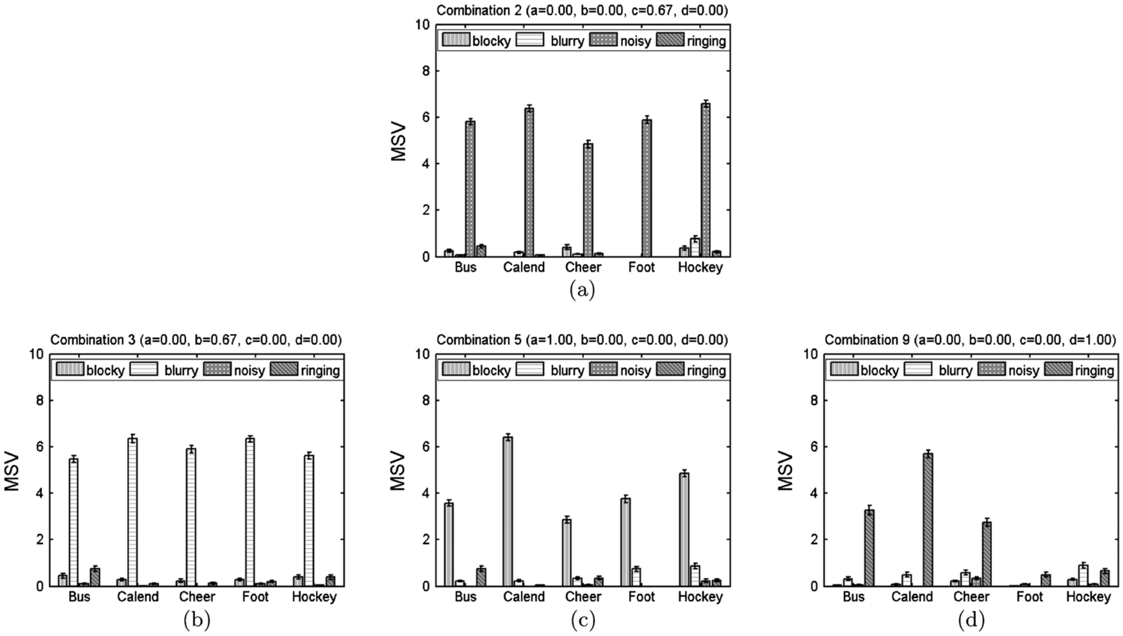
For this fit, we found $p = 0.41978$ with Pearson and Spearman correlation coefficients equal to 0.886039 and 0.9054315, respectively.

Next, we tested the weighted Minkowski metric, which is given by

$$\mathrm{PA}_{\mathrm{WM}} = (\alpha \cdot S_{bk}^p + \beta \cdot S_{br}^p + \gamma \cdot S_{ns}^p + \zeta \cdot S_{\mathrm{ring}}^p)^{1/p}, \quad (13)$$

where $\mathrm{PA}_{\mathrm{WM}}$ is the predicted value for MAV, and α, β, γ, and ζ are the weighted coefficients for the perceptual strengths for blockiness, blurriness, noisiness, and ringing, respectively. In Table 6, we present the coefficients (p, α, β, γ, and ζ), R_σ, PCC, and SCC obtained for the fits of this model to the data corresponding to each original ("Bus," "Calendar," "Cheer," "Football," and "Hockey") and to the data containing all videos ("All").

The P-values (t-test, two-tailed, $P < 0.05$) corresponding with the coefficients α, β, and γ are consistent for all tested models. However, the values obtained for the ζ coefficients are all very low ($0 \geq \zeta \leq 1.536$), implying that ringing is the



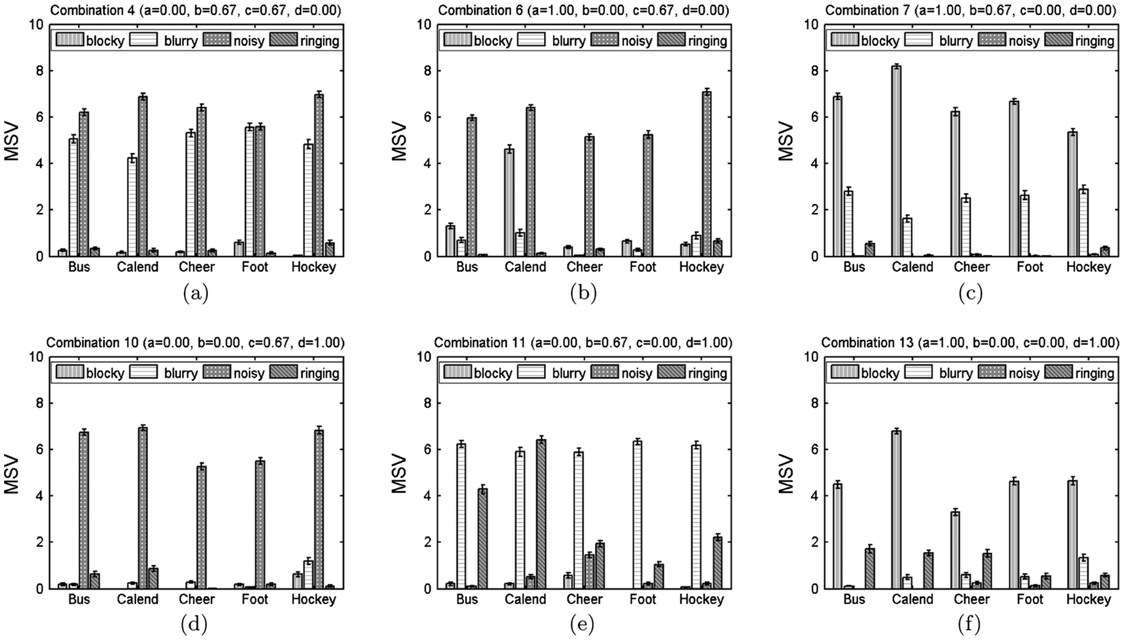
(a)

(b)

(c)

(d)

Fig. 15 Experiment II: MSV bar plots for combinations (a) 2, (b) 3, (c) 5, and (d) 9.



(a)

(b)

(c)

(d)

(e)

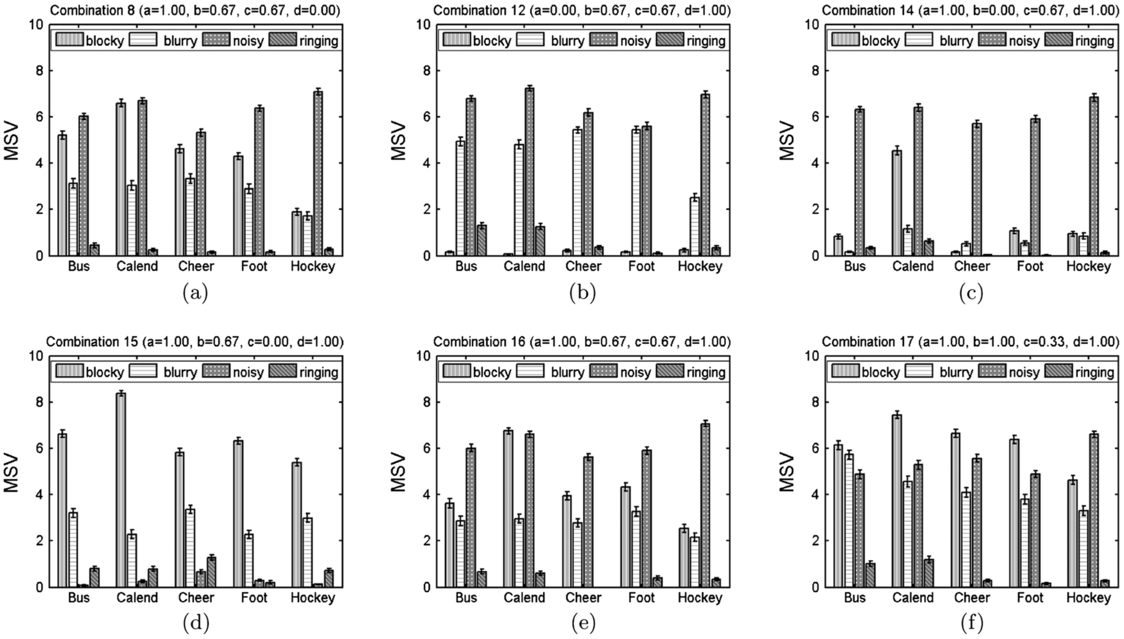
(f)

Fig. 16 Experiment II: MSV bar plots for combinations (a) 4, (b) 6, (c) 7, (d) 10, (e) 11, and (f) 13.

artifact with the lowest weight. For the set containing all video sequences, the Minkowski exponent (p) is equal to 1.0284, and α, β, γ, and ζ are equal to 5.4831, 5.0731, 6.0758, and 0.836, respectively. This fit is better than for the original Minkowski metric, with PCC $= 0.943426$ and SCC $= 0.958771$. Also, there is little systematic error in the predictions (see R_σ in column 6 of Table 6).

From Table 6, we notice that the values of the Minkowski power (p) are all between 0.96 and 1.12. Based on these results, we varied the value of p in the range from 0.8 to 1.4 and repeated the fitting procedure for each of these values. Again, we notice that, for all p values in this interval, the coefficients for ringing (ζ) are very small ($0 \leq \zeta \leq 1.68$). Table 7 summarizes the results for $p = 1.00$ (linear case). Again, the fits are reasonably good for all groups.

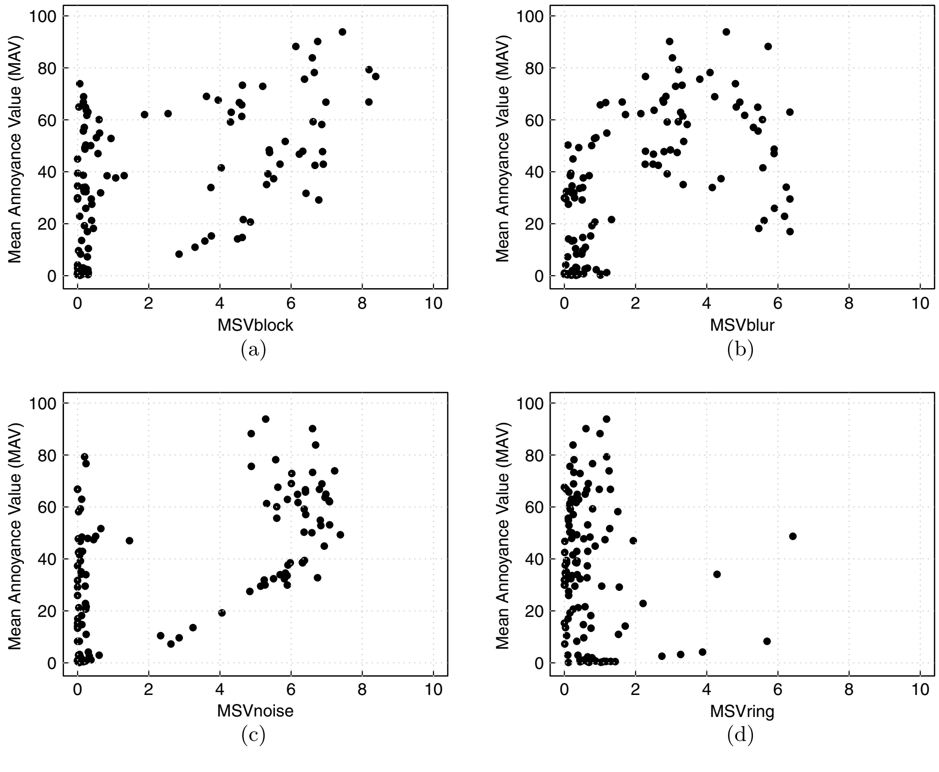
Notice that the correlation coefficients of the linear and weighted Minkowski models are very close. A model comparison test was done to compare the performance of the more generic model (Minkowski metric with p free) against the models with constant p.[29] The results indicate that there is no significant statistical difference between the Minkowski and the models with $1.00 \leq p \leq 1.25$. For models with p values outside this range, a difference was found, and the Minkowski metric performed better. These results are similar



(a) (b) (c)

(d) (e) (f)

Fig. 17 Experiment II: MSV bar plots for combinations (a) 8, (b) 12, (c) 14, (d) 15, (e) 16, and (f) 17.



(a) (b)

(c) (d)

Fig. 18 Experiment II: MSV versus MAV: (a) S_{bk}, (b) S_{br}, (c) S_{ns}, and (d) S_{rg}.

to the results found in Experiment I, where we concluded that both the linear model and the Minkowski metric had the same performance. In Figs. 19 and 20, we have plotted the MAV versus predicted annoyance using the Minkowski metric (PA_{WM}) and the linear model (PA_{LIN}), respectively. The correlation coefficients for these fits are around 0.963.

We also tested a linear model with interactions, which for a set of four artifacts is given by

$$\begin{aligned}\text{PA}_{\text{LINT}} = {} & \alpha \cdot S_{bk} + \beta \cdot S_{br} + \gamma \cdot S_{ns} + \zeta \cdot S_{rg} + \rho_1 \cdot S_{bk}S_{br} \\ & + \rho_2 \cdot S_{bk}S_{ns} + \rho_3 \cdot S_{br}S_{ns} + \rho_4 \cdot S_{bk}S_{rg} + \rho_5 \cdot S_{br}S_{rg} \\ & + \rho_6 \cdot S_{ns}S_{rg} + \tau_1 \cdot S_{bk}S_{br}S_{ns} + \tau_2 \cdot S_{bk}S_{br}S_{rg} \\ & + \tau_3 \cdot S_{bk}S_{ns}S_{rg} + \tau_4 \cdot S_{br}S_{ns}S_{rg} + \chi \cdot S_{br}S_{br}S_{ns}S_{rg}. \end{aligned} \tag{14}$$

Table 6 Experiment II: best fitting parameters for weighted Minkowski metric.

Group	p	α	β	γ	ζ	R_σ	PCC	SCC
Bus	0.8492	3.4245	3.3176	3.77	0.431	5.79	0.98112	0.98087
Calendar	1.1008	7.7850	6.2920	7.5247	1.652	8.95	0.96720	0.93565
Cheer	0.9670	4.1888	4.6616	4.9129	0.000	4.79	0.98482	0.98239
Foot	1.1212	5.0192	8.4764	6.2140	0.000	6.84	0.96284	0.92609
Hockey	1.0814	5.9317	4.1923	7.9600	0.000	5.54	0.98179	0.97652
All	1.0284	5.4831	5.0731	6.0758	0.836	7.40	0.96360	0.96340

Table 7 Experiment II: best fitting parameters for linear model (weighted Minkowski with $p = 1$).

Group	α	β	γ	ζ	R_σ	PCC	SCC
Bus	5.0110	4.9003	5.4313	0.000	5.856	0.97994	0.98000
Calendar	6.0049	4.9767	5.7768	1.536	8.876	0.96501	0.93652
Cheer	4.5872	5.0253	5.3031	0.000	4.574	0.98516	0.98239
Football	3.7382	6.3760	4.6809	0.000	6.8	0.96037	0.93739
Hockey	4.8373	3.6491	6.5309	0.000	5.312	0.98057	0.97565
All	5.0970	4.7461	5.6693	0.859	7.38	0.96307	0.96307

The results of this fitting can be found in Table 8. Column 2 of this table shows the estimated coefficients for the model, while column 5 shows the corresponding P-values (t-test, two-tailed, $P < 0.05$). It can be observed that the coefficients for all main factors are statistically significant, except for ringing (ζ). This result was already expected, since, for all previous models, the ringing coefficients were low. Almost none of the interaction coefficients are statistically significant, except for τ_2, which is a third-order interaction among the perceptual strengths of blockiness, blurriness, and ringing.

The last line of Table 8 shows fitting correlation coefficients for the complete model. Notice that the correlation coefficients of this model are slightly bigger than those

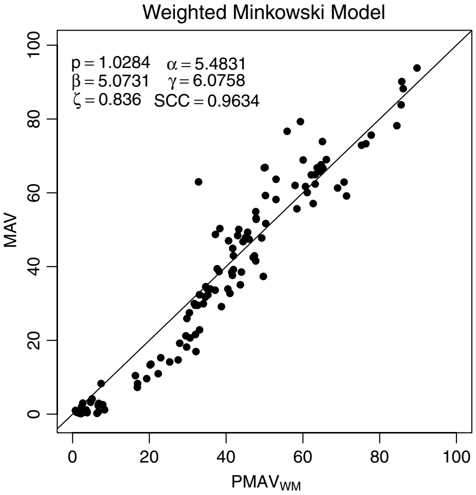


Fig. 19 Experiment II: observed MAV versus predicted MAV using the weighted Minkowski metric (PA_{WM}) for the data set containing all test videos.

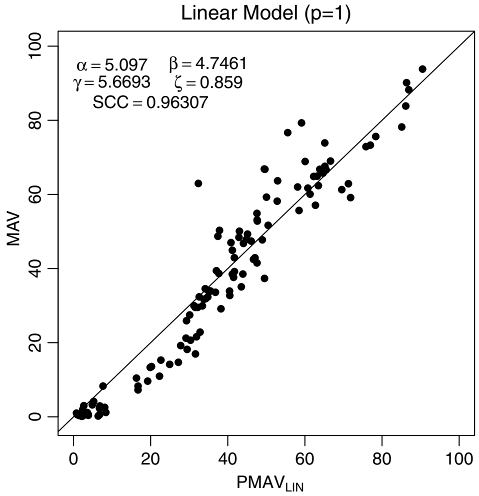


Fig. 20 Experiment II: observed MAV versus predicted MAV using the linear model (PA_{LIN}) for the data set containing all test videos.

Table 8 Experiment II: fitting parameters for the linear metric with interactions.

Coef.	Estimate	Std. error	t value	Pr(>|t|)
α	5.29296	0.72134	7.338	4.78e−11[a]
β	3.81409	0.51089	7.466	2.54e−11[a]
γ	5.60526	0.39524	14.182	<2e−16[a]
ζ	−0.06948	0.85135	−0.082	0.9351
ρ_1	−0.05560	0.32958	−0.169	0.8664
ρ_2	0.01172	0.37439	0.031	0.9751
ρ_3	0.15509	0.16878	0.919	0.3603
ρ_4	−1.33643	0.79504	−1.681	0.0957
ρ_5	0.27184	0.26506	1.026	0.3074
ρ_6	0.27565	1.07856	0.256	0.7988
τ_1	−0.06000	0.12051	−0.498	0.6196
τ_2	0.85707	0.34166	2.509	0.0137[a]
τ_3	0.40734	0.68791	0.592	0.5550
τ_4	0.09792	0.28363	0.345	0.7306
χ	−0.14757	0.17396	−0.848	0.3982
Corr. coef.	PRCC = 0.96972		SRCC = 0.96278	

[a]Statistically significant at (P < 0.05).

Table 9 Experiment II: ANOVA table for factorial design (combinations 1–16).

Source	Sum sq.	d.f.	Mean sq.	F	Prob > F
c	15066	1	15066	393.47	0.0000[a]
b	16871.6	1	16871.6	440.62	0.0000[a]
a	2978.6	1	2978.6	77.79	0.0000[a]
d	264.8	1	264.8	6.92	0.0114[a]
Group	2504.5	4	626.1	16.35	0.0000[a]
$c*b$	277.9	1	277.9	7.26	0.0096[a]
$c*a$	675.8	1	675.8	17.65	0.0001[a]
$c*d$	22.9	1	22.9	0.6	0.4429
$c*$group	722.8	4	180.7	4.72	0.0027[a]
$b*a$	4.2	1	4.2	0.11	0.7434
$b*d$	44.9	1	44.9	1.17	0.2842
$b*$group	885.9	4	221.5	5.78	0.0007[a]
$a*d$	2.6	1	2.6	0.07	0.7973
$a*$group	1109.9	4	277.5	7.25	0.0001[a]
$d*$group	271.9	4	68	1.78	0.1489
Error	1876.2	49	38.3		
Total	43580.5	79			

[a]Statistically significant at (P < 0.05).

for the linear model with no interactions. However, a comparison model test showed that the differences are not statistically significant. Again, this model uses the perceived strength of the artifact and not the actual artifact signal strength. Therefore, it is not surprising that these results contradict the results shown in the bar plots.

To investigate the effects of the actual artifact signal strength (a, b, c, and d) and the "group" (original video: "Bus," "Calendar," "Cheerleader," "Football," or "Hockey") on the MAV, we performed an analysis of variance (ANOVA) test on a subset of data consisting of combinations 1 to 16 (factorial design). Table 9 shows the ANOVA results for the main effects and interactions among terms. The results show that all artifact signals have a significant effect on MAV ($P < 0.05$). Ringing is slightly significant in a 95% test and not significant in a 99% test. Therefore, the ANOVA test shows that the addition of a ringing signal does have a (small) significant effect on the perceived MAV, but the corresponding ringing perceptual strenth (S_{rg}) does not. This might indicate that, when other artifacts are present, ringing is not being perceived as ringing (see, for example, the bar plots in Fig. 17).

The content of the video (variable "group") showed a high impact on how the artifacts were perceived. Given how each artifact signal affects the video, this is expected. For example, blockiness is more visible in uniform areas than in high-texture areas. Ringing, on the other hand, is more visible when there are edges and lines over a uniform background. The results also showed an interaction between the group and c (noisy), the group and b (blurry), a (blocky) and c (noisy), and b (blurry) and c (noisy). Noisiness was the artifact signal with the highest level of interaction with other artifacts (blockiness and blurriness). The ringing signal, on the other hand, did not seem to interfere with the perception of any of the other artifacts. This might be due to intrinsic characteristics of this artifact, which is perceived as relevant only when other artifacts are not present. In summary, the results of the ANOVA test show that all artifact signals are significant for predicting annoyance, and that these artifacts interact with one another to produce the overall annoyance.

6 Summary and Conclusions

We presented the description, statistical analysis, and conclusions of two psychophysical experiments. The goals of these experiments were to study the appearance, visibility, and annoyance of four artifacts (blockiness, blurriness, ringing, and noisiness) commonly found in digital videos and to understand how these artifacts combine and interact to produce overall annoyance. The results showed that, when the artifact signals were presented alone at a high strength, subjects were able to identify them correctly. At low strengths, on the other hand, other artifacts were reported. Annoyance increased with both the number of artifacts and their strength. The noisy artifact signals seemed to decrease the perceived strength of the other artifacts, while blurry artifact signals seemed to increase them.

Annoyance models were created by combining the artifact perceptual strengths (MSV) using a Minkowski model, a weighted Minkowski model, a linear model, and a linear model with interactions. A comparison between the Minkowski metric and the linear model showed that there is no statistical difference between these two models. Performing an ANOVA test, we found that, besides the group (content), all types of artifact signal strengths had a significant effect on MAV. The ANOVA test also indicated that there are interactions among some of the artifact signal strengths and the group.

The results presented in this paper provide information that can be used in the design of video quality models and, more specifically, on the design of NR models.[7–10] In particular, the results show that annoyance can be modeled as a multidimensional function of the individual artifact signal measurements. This implies that the NR quality model based on artifact measurements is indeed a valid approach, and it needs to include a minimal set of the most relevant artifacts. Also, although annoyance cannot be predicted using only one individual artifact signal measurement, it is not necessary to use all possible artifacts. It suffices to use the most significant (statistically) ones. For example, in Experiment II, the ringing signal was proven to have only a small effect on the prediction of MAV. Also, its importance seemed to decrease with the introduction of other artifacts. Therefore, a quality model can be designed in a way that the strength of ringing signal is estimated only when no other artifacts are present. Finally, the results show that there are interactions among artifact signals. Therefore, while designing quality models, it is important to take this into consideration to avoid underestimating or overestimating quality. To our knowledge, there are no quality models that take into account the interactions among artifact signals.

Acknowledgments

We would like to thank Prof. John M. Foley for his valuable contributions to this work. This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Fundação de Empreendimentos Científicos e Tecnológicos (Finatec), and in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

1. M. Yuen and H. R. Wu, "Reconstruction artifacts in digital video compression," *Proc. SPIE* **2419**, 455–465 (1995).
2. ITU-T Recommendation P.930: Principles of a reference impairment system for video, International Telecommunication Union (1996).
3. S. Chikkerur et al., "Objective video quality assessment methods: a classification, review, and performance comparison," *IEEE Trans. Broadcast.* **57**(2), 165–182 (2011).
4. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206, MIT Press, Cambridge, MA (1993).
5. M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," *Proc. SPIE* **5150**, 583–592 (2003).
6. Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Comm.* **19**(2), 121–132 (2004).
7. H. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.* **4**(11), 317–320 (1997).
8. Z. Wang, A. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. on Image Processing*, Vol. 3, pp. 981–984 (2000).
9. J. Caviedes et al., "Impairment metrics for digital video and their role in objective quality assessment," *Proc. SPIE* **4067**, 791–800 (2000).
10. M. Farias and S. Mitra, "No-reference video quality metric based on artifact measurements," in *Proc. IEEE Intl. Conf. on Image Processing*, pp. III141–III144 (2005).
11. V. Kayarrgadde and J. Martens, "Perceptual characterization of images degraded by blur and noise: model," *J. Opt. Soc. Am. A Opt Image Sci.* **13**(6), 1178–1188 (1996).
12. H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," *Proc. SPIE* **1666**, 16–26 (1992).
13. M. Farias, "No-reference and reduced reference video quality metrics: new contributions," Ph.D. Dissertation, University of California Santa Barbara (2004).
14. D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," *Proc. SPIE* **6057**, 60570F (2006).
15. A. Moorthy and A. Bovik, "Visual quality assessment algorithms: what does the future hold?," *Int. J. Multimedia Tools Appl.*, Special Issue on Survey Papers in Multimedia by World Experts **51**(2), 675–696 (2011).
16. T. Wolff et al., "Modeling subjectively perceived annoyance of H.264/AVC video as a function of perceived artifact strength," *Signal Process.* **90**(1), 80–92 (2010).
17. M. Farias, J. Foley, and S. Mitra, "Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts," *IEEE Trans. Signal Process.* **55**(6), 2954–2964 (2007).
18. M. Moore, J. Foley, and S. Mitra, "Defect visibility and content importance: effects on perceived impairment," *Image Commun.* **19**(2), 185–203 (2004).
19. M. Farias et al., "Detectability and annoyance of synthetic blocky and blurry artifacts," in *Proc. SID International Symposium*, Vol. XXXIII, No. II, pp. 708–712, Boston, MA (2002).
20. M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Process.* **70**(3), 247–278 (1998).
21. G. Cermak et al., "Validating objective measures of MPEG video quality," *SMPTE J. Soc. Motion Picture Television Eng.* **107**(4), 226–235 (1998).
22. VQEG subjective test plan (Phase 1), Video Quality Experts Group, ftp://ftp.crc.ca/crc/vqeg/phase1-docs (1999).
23. ITU-T Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union (1998).
24. J. Libert, C. P. Fenimore, and P. Roitman, "Simulation of graded video impairment by weighted summation: validation of the methodology," *Proc. SPIE* **3845**, 254–265 (1999).
25. M. Farias et al., "Perceptual contributions of blocky, blurry, and fuzzy impairments to overall annoyance," *Proc. SPIE* **5292**, 109–120 (2004).
26. M. Moore, "Psychophysical measurement and prediction of digital video quality," Ph.D. Dissertation, University of California Santa Barbara (2002).
27. ITU Recommendation BT.500-8: Methodology for subjective assessment of the quality of television pictures (1998).
28. Z. Wang and A. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures" *IEEE Signal Process. Magazine*, **26**(1), 98–117 (2009).
29. W. Hays, *Statistics for the Social Sciences*, 3rd ed., LLH Technology Publishing, New York, NY (1981).
30. P. Marziliano et al., "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Process. Image Commun.* **19**(2), 163–172 (2004).

Mylène C. Q. Farias received her BSc in electrical engineering from Universidade Federal de Pernambuco (UFPE), Brazil, in 1995 and her MSc in electrical engineering from the Universidade Estadual de Campinas (UNICAMP), Brazil, in 1998. She received her PhD in electrical and computer engineering from the University of California Santa Barbara, in 2004 for work in no-reference video quality metrics. She has worked as a research engineer at CPqD (Brazil) in video quality assessment and validation of video quality metrics. She has also worked as an intern for Philips Research Laboratories (the Netherlands) in video quality assessment of sharpness algorithms and for Intel Corporation (Phoenix, Arizona) developing no-reference video quality metrics. She is currently an assistant professor in the Department of Electrical Engineering at the Universidade de Brasília (UnB), Brazil. Her current interests include video quality metrics, video processing, multimedia, watermarking, and information theory. She is a member of the IEEE and IEEE Signal Processing Society.

Sanjit K. Mitra is the Stephen and Etta Varra professor in the Department of Electrical Engineering-Systems at the University of Southern California Los Angeles, and a research professor at the University of California Santa Barbara. He has served the IEEE in various capacities, including as the president of the IEEE Circuits & Systems Society in 1986. He has held visiting appointments in Australia, Austria, Brazil, Croatia, Finland, Germany, India, Japan, Singapore, Turkey, and the United Kingdom. He is the recipient of several awards, including the 1989 Education Award and 2000 Mac Van Valkenburg Society Award of the IEEE Circuits & Systems Society; the 1996 Technical Achievement Award, 2001 Society Award, and 2006 Education Award of the IEEE Signal Processing Society; the 2001 McGraw-Hill/Jacob Millman Award of the IEEE Education Society; the 2002 Technical Achievement Award of the European Association for Signal Processing (EURASIP); the 2005 SPIE Technical Achievement Award of the International Society for Optical Engineering; and the 2006 IEEE James H. Mulligan Jr. Education Medal. He is the co-recipient of the 2000 Blumlein-Browne-Willans Premium of the Institution of Electrical Engineers (London) and the 2001 IEEE Transactions on Circuits & Systems for Video Technology Best Paper Award. He has been awarded honorary doctorates from the Tampere University of Technology, Finland, the Technical University of Bucharest, Romania, and the University Medal of the Technical University of Slovakia at Bratislava. He is an academician of the Academy of Finland, a member of the U.S. National Academy of Engineering, a member of the Norwegian Academy of Technological Sciences, a foreign member of the Croatian Academy of Sciences and Arts, and a foreign member of the Academy of Engineering of Mexico. He is a Fellow of the IEEE, AAAS, and SPIE.