Using Perceptual Strength Estimates to Predict the Perceived Annoyance of Videos with Combinations of Spatial and Temporal Artifacts

Alexandre F. Silva[a], **Mylène C. Q. Farias**[b]
[a]Federal Institute of Triângulo Mineiro, Department of Computer Science, Paracatu, Brazil.
[b]University of Brasilia, Department of Electrical Engineering, Brasília, Brazil.

Abstract. We report the results of a set of psychophysical experiments that measure the perceptual strengths of videos with different combinations of blockiness, blurriness, and packet-loss artifacts. Participants were instructed to search each video for impairments and rate the strength of their individual features (artifacts). A repeated-measure ANOVA (RM-ANOVA) performed on the data showed that artifact physical strengths have a significant effect on annoyance judgments. We tested a weighted Minkowski model, a support vector regression model, and a linear model on the experimental data. We found that all these models give a good description of the relation between individual artifact perceptual strengths and the overall annoyance. In the other words, all models presented a very good correlation with the experimental data, showing that annoyance can be modeled as a multidimensional function of the individual artifact perceptual strengths. Additionally, results show that there are interactions among artifact signals.

Keywords: Video quality assessment, compression artifacts, subjective quality, human visual system modeling.

*Alexandre F. Silva: alexandrefieno@gmail.com; Mylene C.Q. Farias: mylene@ieee.org

1 Introduction

A video impairment is any change in a video signal that can be introduced during capture, transmission, storage, as well as by any signal processing algorithm (e.g. compression) that may be applied to the content and, if sufficiently strong, reduce its perceived quality. Impairments can be very complex in their physical and perceptual descriptions.[1] Most impairments have more than one perceptual feature, but it is possible to produce impairments that are relatively pure. The perceptual features of impairments are known as "artifacts", while the physical signals that produce the artifacts are known as artifact signals. Many types of artifacts can be introduced by digital video systems, but, in this work, we limit ourselves to three artifacts (blockiness, blurriness, and packet-loss), which are among the most relevant artifacts in digital transmission scenarios.[2]

Designing a video quality metric that can detect impairments and estimate their annoyance (as perceived by human viewers) is not an easy task.[2] In the past decade, a big effort in the scientific community has been devoted to the development of video quality metrics that correlate well with the human perception of quality.[3–6] But, although a great number of video quality metrics has been proposed in the literature, most of these metrics estimate impairment annoyance by comparing original and impaired videos, i.e. most available metrics are full-reference (FR).[7,8]

When a video is degraded by the presence of several types of artifacts, the perceived quality is affected.[5,9–13] Therefore, alternatives to regular quality metrics include artifact metrics[14,15] that measure the strength of individual artifacts. Given that the overall video quality can be estimated by combining the individual artifact *perceptual* strengths, the output of these metrics can be combined to obtain an overall annoyance score.[1] There is a considerable number of no-reference (NR) metrics that uses this 'multidimensional' approach for measuring the overall quality of a video.[16–18]

Naturally, the performance of an artifact-based metric depends on the performance of the individual artifact metrics. Therefore, the design good artifact metrics requires a good understanding of the perceptual characteristics of each artifact, as well as a knowledge of how the strength of each artifact contributes to the overall quality.[16,19] For example, Farias *et al.*[2,20] performed experiments that measured the perceived artifact strengths and the overall annoyance of combinations of artifacts. Their results showed that, when presented in combination and, at a low strength, artifacts that otherwise would be clearly recognized are mistaken by others. Also, the presence of noise in videos seems to decrease the perceived strength of other artifacts, while the presence of blurriness has the opposite effect. In their study, no temporal artifacts were considered and no relationship could be established between video content and quality. Up to our knowledge, besides the work by Farias *et al.*, little work has been performed to study and characterize the appearance and perception of combined artifacts.[7] As a consequence, currently there is no clear knowledge on how different artifacts combine perceptually and how their impact depends on the physical properties of the video.

In this work, we study the characteristics of two spatial artifacts (blockiness and blurriness) and one temporal artifact (packet-loss), which are among the most commonly found artifacts in digital transmission scenarios. More specifically, we analyze the relationship between the “perceptual strengths” of these artifacts and the overall annoyance. We also analyze the relationship between physical and perceptual artifact strengths and study masking effects between artifacts. With this goal, we perform a set of three psychophysical experiments in which subjects estimated the strength of blockiness, blurriness, and packet-loss artifacts, either in isolation or in combinations. Then, we performed an analysis of the subjective data obtained from these experiments and tested a set of combination models with the goal of predicting overall annoyance from the perceptual strengths of these three artifacts. This work is a follow-up on a previous work,[21,22] in which we investigated the impact of physical strength parameters of blockiness, blurriness, and packet-loss on overall annoyance.

In summary, in this work we are interested in understanding how the perceptual strengths of spatial and temporal artifacts combine to produce the overall annoyance, whilst in our previous work we were interested in studying the visibility and annoyance of these artifacts. More specifically, while in the previous work we collected and studied the overall annoyance data corresponding to sequences with combinations of three artifacts, in this work we collected perceptual strength data for each of these three artifacts for the same degraded sequences. With this strength data, we study the contribution of each artifact for the overall quality. First, we test the effect on the overall annoyance of the presence of each artifact and analyze possible facilitation and masking effects among artifacts. Second, we fit several annoyance models, which combine the individual artifact perceptual strengths to predict overall annoyance, to understand the contribution of each artifact to the overall annoyance. We compare all models in terms of their performance and complexity.

The paper is divided as follows: Sec. 2 presents the experimental methodology, which includes the stimuli generation, the type of equipment, the subjective method, and the statistical analysis. Sec. 3 introduces the experimental results and the annoyance models. Conclusions are detailed in Sec. 4.

2 Experimental Methodology

To understand the relationship between the perceptual strengths of blockiness, blurriness, and packet-loss artifacts and how they can be combined to estimate the overall annoyance, we performed a set of three psychophysical experiments using test sequences with combinations of these artifacts at different strengths.[22]

2.1 Stimuli

We used seven high definition original videos, chosen with the goal of generating a diverse content, with spatial resolution of 1280×720, temporal resolution of 50 frames per second (fps), and duration of 10 seconds. We followed the recommendations detailed in the Final Report of VQEG on the validation of objective models multimedia quality assessment (Phase I),[23] which suggest using a set of video sequences with a good distribution of spatial and temporal properties.[24]

To add artifacts to the originals, we used a system for generating artifacts[20] that allowed a control of the artifact combination, visibility, and strength, which would be impossible when using, for example, a H.264 codec. To add blockiness to each video frame in our dataset, we calculated the average value of each 8×8 block of the frame and of the 24×24 surrounding block, then added the difference between these two averages to the block. To generate blurriness, we used a simple low-pass filter, as suggested by Recommendation P.930.[25] Although we can vary the filter sizes and the cut-off frequencies to control the amount of blurriness, we used a simple 5×5 moving average filter.

To generate packet-loss artifacts, we first compressed the videos at high compression rates, what avoids inserting additional artifacts. Then, packets from the coded video bitstream were randomly deleted using different loss percentages (the higher the percentage, the lower the quality).[22] To vary the time interval between consecutive artifacts, we changed the number of frames (M) between I-frames. We generated test sequences with combinations of blockiness, blurriness, and packet-loss, we first linearly combined the original video with blockiness and blurriness artifact signals in different proportions (i.e. 0.4, 0.6, and 0.8).[26] Then, we added packet-loss artifacts using the same procedure used to generate these artifacts in original content.

2.2 *Methodology and Equipment*

The experiments were performed using a PC computer with test sequences displayed on a Samsung LCD monitor of 23 inches (Sync Master XL2370HD) with resolution 1920×1080 @60hz (FullHD 1080p). The dynamic contrast of the monitor was turned off, the contrast was set at 100, and the brightness at 50. The monitor measured gamma values for luminance, red, green, and blue were 1.937, 1.566, 1.908, and 1.172, respectively. We set a constant illumination of approximately 70 lux. Participants were kept at a fixed distance of 0.70 meters from the monitor using a chin-rest. The experimental methodology was the single-stimulus with hidden reference, with a *100-point* continuous-scale.[22,27]

The participants were mostly graduate students. They were considered naive of most kinds of digital video defects and the associated terminology. No vision test was performed, but participants were asked to wear glasses or contact lenses if they needed them to watch TV. The experiment started after a brief oral introduction. Then, participants performed a training stage that consisted of watching highly impaired and pristine sequences to get acquainted with the typical artifact combinations and strengths. The sequences presented during the training were not scored and were meant to be visual anchors (references) for the annoyance scoring.

After the training, the actual scoring session started. After each test sequence was played, participants were asked to give a strength score to each individual type of artifact. Artifacts as strong as those seen in the training session should be given a *100* strength score, artifacts half as strong a *50* strength score, and so on. In each experiment, the number of artifacts present in the test sequences varied. To avoid fatigue, experimental sessions were broken into sub-sessions, between which participants could take a break for as long as they wanted to. All experimental sessions lasted between 45 and 60 minutes.

It is worth pointing out that in a previous set of experiments,[21,22] instead of rating the strengths of individual artifacts, subjects were asked to give an overall annoyance score to each of the test sequence. More specifically, to estimate the annoyance caused by the artifacts in the test sequences, subjects were asked to give a score between *0* and *100*. Artifacts as annoying as the worst artifacts shown in the training session should have be given a *100* annoyance score, artifacts half as annoying a *50* annoyance score, and so on. The same set of test sequences used in this previous set of experiments is used in this work, what makes it possible to compare the data collected in both sets of experiments.

2.3 *Statistical Analysis*

Data gathered from the three experiments provided up to three Mean Strength Values (MSV) for each test sequence: MSV_{bloc}, MSV_{blur}, and MSV_{pck}, which correspond to MSVs for blockiness, blurriness, and packet-loss, respectively. For each video and artifact type, we computed the MSVs by averaging the strength values over all subjects:

$$MSV_a = \frac{1}{T}\sum_{i=1}^{T} S_a(i), \tag{1}$$

where $S_a(i)$ is the strength value reported by the i^{th} subject, T is the number of subjects, and *a* refers to the type of artifact, i.e. *bloc*, *blur*, or *pck*.

As mentioned earlier, in a previous experiment,[21,22] for each test sequence we obtained a Mean Annoyance Value (MAV), given by:

$$MAV = \frac{1}{T}\sum_{i=1}^{T} A(i), \tag{2}$$

where $A(i)$ is the annoyance value reported by the i^{th} subject. To study how the artifact strengths combine to predict the perceived annoyance of videos impaired by overlapping artifacts, we fit a set of linear and non-linear models to the MSV and MAV data collected.[21,22]

To estimate the performance of the models, we calculate the Pearson correlation coefficient (PCC) and the Spearman Rank Order Correlation Coefficient (SCC) between the subjective and predicted scores. Also, we use the Akaike Information Criterion (AIC)[28] to analyze the trade-off between accuracy of fitting and the number of degrees of freedom in the model, thereby controlling the bias/variance trade-off and overfitting. To test the effect of the artifact parameters on models, we perform a repeated-measure ANOVA (RM-ANOVA) with a significance level of 95%.

We also use a Support vector regression (SVR) technique to predict annoyance from the subjective data. To train the SVR, we use a k-fold cross-validation setup. First, we split the dataset in k equally sized non-overlapping sets. Then, we run the training k times. Each time, we use a different fold as a test set, while using the remaining $(k-1)$ folds for training.[29] In our experiments, we set k equal to 10. We use a radial kernel for SVR, since it maps samples into a higher dimensional space, handling well cases in which the relationship between the class labels and the attributes is nonlinear.[30] SVR has the following parameters: C, ϵ, γ. The parameter *C* determines the trade-off between the training error and the model complexity, while the parameter ϵ determines the level of accuracy of the approximated function. Finally, γ is a parameter of the Gaussian radial basis function, with a small γ corresponding to a low bias/high variance and a large γ corresponding to a higher bias/low variance.

2.4 Experiments

The three experiments shared identical experimental methodology, interface, protocol, and viewing conditions. The stimuli were different per experiment but derived from a common set of 7 original contents. Next, we describe briefly each experiment.

Experiment 1: Fourteen participants performed strength tasks on test sequences containing only packet-loss artifacts. The artifact strength varied by changing the percentages of deleted packets (PDP = 0.7%, 2.6%, 4.3%, and 8.1%) and the number of M frames between the I-frames (M = 4, 8, and 12). A total of 7 originals and 12 combinations were used, resulting in $12{\times}7{+}7 = 91$ test sequences.[31,32]

Experiment 2: Fifteen participants performed strength tasks on test sequences containing different strengths of blockiness and blurriness artifacts, presented at isolation or in combination. We represent the artifact strength combinations as a vector (bloc;blur), where *bloc* corresponds to the blockiness strength and *blur* to the blurriness strength. Three strengths (0.0, 0.4, and 0.6) were used in the experiment, which consisted of a full factorial experimental design ($3^2 = 9$ combinations) and included unimpaired videos (0.0;0.0). Two further combinations, pure blockiness (0.8;0) and pure blurriness (0.8;0) at a high strength of 0.8 were also added to the set. In total, $11 \times 7 = 77$ test sequences were evaluated by the participants.[22]

Experiment 3: Fifteen participants performed strength tasks on test sequences containing different strengths of blockiness, blurriness, and packet-loss artifacts, presented in combinations. We represent the strength combinations as a vector (PDP;bloc;blur), where *PDP* corresponds to the packet-loss strength, *bloc* to the blockiness strength, and *blur* to the blurriness strength. These artifacts were combined at 3 different strengths: bloc $\in [0.0, 0.4, 0.6]$, blur $\in [0.0, 0.4, 0.6]$, and PDP $\in [0.0, 0.7\%, 8.1\%]$, resulting in $20 \times 7 = 140$ test sequences.[1,22]

3 Experimental Results

In this section, we present the statistical analysis of the experimental data collected from experiments 1-3.

3.1 Experiment 1: Packet-Loss

As mentioned earlier, in Experiment 1 we used test sequences with only packet-loss. Fig. 1 shows a graph of the average MSV_{pck} versus PDP, grouped according to the M value. In this graph, we also show the average MSV_{pck} calculated only for the original videos (blue point in the left side of

Table 1 Exp. 1: Pairwise comparisons between average MSV_{pck}, with M = 12 and different PDP values. (* Significant at 0.05 level.)

Pairs of PDP values	Diff. Mean	Std. Error
(0.7, 2.6)	-21.541*	2.434
(0.7, 4.3)	-29.684*	2.280
(0.7, 8.1)	-35.918*	3.077
(2.6, 4.3)	-08.143*	2.357
(2.6, 8.1)	-14.378*	2.584
(4.3, 8.1)	-06.235	2.492

the graph). Notice that the MSV_{pck} values are not equal to zero for the original (pristine) videos, indicating that subjects perceived impairments in unimpaired videos. For M = 4, 8, and 12, the highest MSV_{pck} always correspond to the strongest artifact (i.e. PDP = 8.1$\%$). Although MSV_{pck} increases with both PDP and M, PDP seems to have a bigger effect on MSV_{pck} than M. Visually, packet-loss artifacts in videos with large smooth regions (e.g. skies) were easier to detect, while in videos with a high spatial and/or temporal activity they are harder to detect.

We perform an RM-ANOVA to check the influence of the M and PDP parameters on MSV_{pck}. Results show that the differences between the MSV_{pck} average values obtained for any pair of M values are statistically significant. When we analyze the influence of PDP on MSV_{pck}, we notice that there are significant statistical differences between the MSV_{pck} values for most PDP pairs, with the exception of the pair PDP = 4.3$\%$ and PDP = 8.1$\%$ for M = 12, as showed in Table 1.

With the goal of studying if MSV can be used to predict the annoyance, we test the following simple linear model without any interaction term:

$$PA_{E1,L1} = \alpha \cdot MSV_{pck}, \tag{3}$$

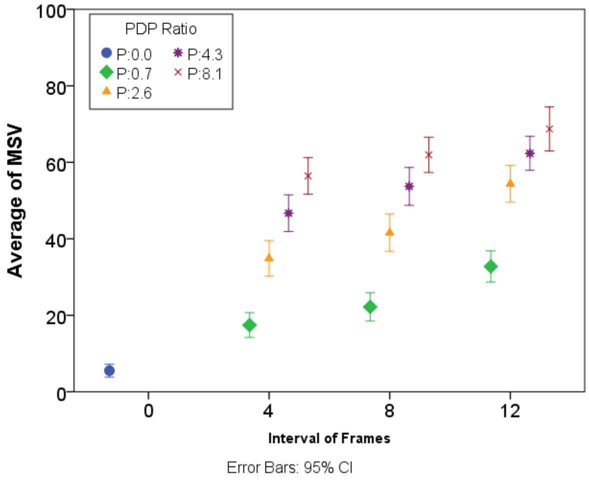


Fig 1 Exp. 1: MSV_{pck} plots for clustered error for M = 4, 8, and 12.

Table 2 Exp. 1: Fitting parameters for linear model without intercept ($PA_{E1,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
α	0.904	0.016	56.150	$< 2e - 16*$	0.953	0.949

Table 3 Exp. 1: Fitting parameters for linear model with intercept ($PA_{E1,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>\|t\|$)	PCC	SCC
δ	-4.396	1.607	-2.736	0.007*	0.953	0.950
α	0.983	0.033	29.816	$< 2e - 16*$		

and the following linear model with an intercept term δ:

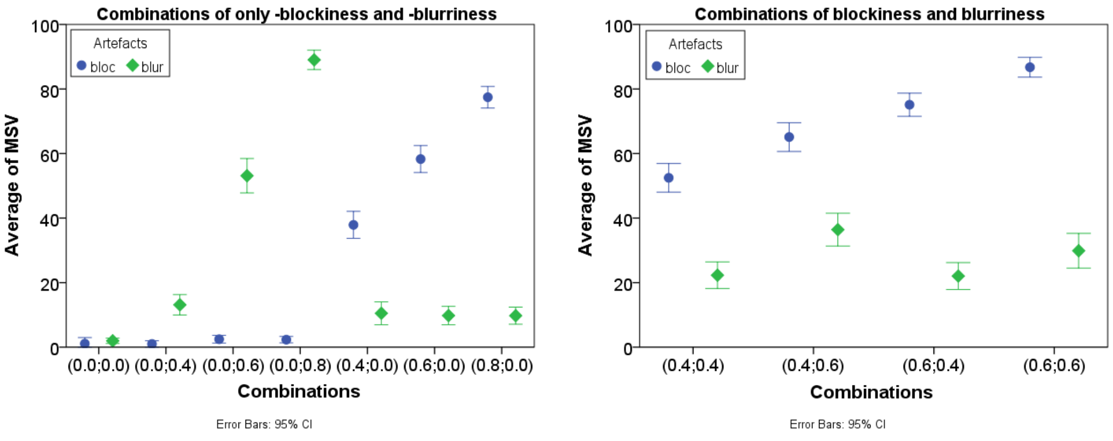
$$PA_{E1,L2} = \delta + \alpha \cdot MSV_{pck}. \tag{4}$$

Tables 2 and 3 show the fitting results for both models. All coefficients are statistically significant.

We also use the Support vector regression (SVR) technique to predict annoyance from the strength data using MSV_{pck}. We refer to this model as $PA_{E1,SVR}$. The second line of Table 4 summarizes the SVR results, with columns 2-5 showing the estimated parameters and columns 6-7 showing the PCC and SCC values for the fit. For best results, we use a radial kernel for the SVR, obtaining PCC and SCC values equal to 0.953 and 0.927, respectively.

Table 4 Exp. 3: Fitting parameters for SVR model by Experiment.

Experiment	K	C	γ	ϵ	PCC	SCC
Exp. 1: $PA_{E1,SVR}$	radial	64	1	0.0	0.953	0.927
Exp. 2: $PA_{E2,SVR}$	radial	8	0.5000	0.0	0.982	0.948
Exp. 3: $PA_{E3,SVR}$	radial	4	0.3333	0.1	0.963	0.957



(a) (b)

Fig 2 Exp. 2: MSV plots for clustered error for combinations (bloc;blur): (a) only -blockiness and -blurriness, and (b) blockiness and blurriness.

3.2 *Experiment 2: Blockiness and Blurriness*

As mentioned earlier, test sequences used in Experiment 2 had two different types of artifacts: blockiness and blurriness. These artifacts were presented in different strengths, either in isolation or in combination. Fig. 2 shows a graph of the average MSV_{blur} (green) and the average MSV_{bloc} (blue) for test sequences containing combinations of only-blurriness and only-blockiness. The first combination of the graph corresponds to the original (pristine) videos. Notice that, again, the MSVs for pristine videos are not equal to zero, indicating that participants perceived impairments in these unimpaired videos ($\overline{MSV}_{blur} = 1.95$ and $\overline{MSV}_{bloc} = 1.09$). But, in general, highest MSVs are obtained for the combinations with higher artifact strengths.

An RM-ANOVA test shows that there are significant statistical differences between the MSV_{blur}

Table 5 Exp. 2: Pairwise comparisons between average MSV_{blur} for sequences with only-blurriness (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.4)	(0.0;0.6)	-39.990*	2.631
(0.0;0.4)	(0.0;0.8)	-75.905*	2.125
(0.0;0.6)	(0.0;0.8)	-35.914*	2.641

Table 6 Exp. 2: Pairwise comparisons between average MSV_{bloc} for sequences with only-blockiness (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(bloc,blur)			
(0.4;0.0)	(0.6;0.0)	-20.390*	2.597
(0.4;0.0)	(0.8;0.0)	-39.552*	2.440
(0.6;0.0)	(0.8;0.0)	-19.162*	1.888

Table 7 Exp. 2: Pairwise comparisons between average MSV_{bloc} and MSV_{blur} for any pair of blurriness and blockiness (*. Significant at 0.05 level.)

Combinations		MSV_{bloc}		MSV_{blur}	
(bloc,blur)		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.4;0.4)	(0.4;0.6)	-12.629*	2.915	-14.133*	3.068
(0.4;0.4)	(0.6;0.4)	-22.638*	2.414	0.248	2.709
(0.4;0.4)	(0.6;0.6)	-34.267*	2.330	-7.590*	3.307
(0.4;0.6)	(0.6;0.4)	-10.010*	2.220	14.381*	2.713
(0.4;0.6)	(0.6;0.6)	-21.638*	2.108	6.543*	2.881
(0.6;0.4)	(0.6;0.6)	-11.629*	1.525	-7.838*	2.906

corresponding to any pair of videos with only-blurriness (see Table 5) and the MSV_{bloc} corresponding to any pair of videos with only-blockiness (see Table 6). These results indicate that participants correctly perceived the different artifact strengths introduced in the videos.

For all combinations of blockiness and blurriness ((0.4;0.4), (0.4;0.6), (0.6;0.4), and (0.6;0.6)), MSV_{bloc} were higher than MSV_{blur}. Fig. 2 (b) shows a plot of MSV_{blur} and MSV_{bloc} for all videos. Again, an RM-ANOVA test shows that differences between MSVs obtained for any two combinations of blockiness and blurriness are statistically significant. The only exception is the combination pair (0.4;0.4) and (0.6;0.4), for which MSV_{blur} differences are not statistically significant. Results of this RM-ANOVA are reported in Table 7.

Table 8 Exp. 2: Fitting parameters for linear model without intercept ($PA_{E2,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
α	0.797	0.016	49.220	$< 2e-16*$	0.971	0.958
β	0.721	0.023	30.840	$< 2e-16*$		

Table 9 Exp. 2: Fitting parameters for linear model with intercept ($PA_{E2,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
δ	0.386	1.725	0.224	0.824		
α	0.793	0.025	32.302	$< 2e-16*$	0.971	0.958
β	0.716	0.032	22.103	$< 2e-16*$		

To verify if we can predict annoyance using the MSVs of blockiness and blurriness, we test a set of linear and non-linear models using MSV_{bloc}, MSV_{blur}, and MAV data. The first model is a simple linear model, as given by:

$$PA_{E2,L1} = \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur}, \tag{5}$$

and the second model is a linear model with an intercept term δ, as given by:

$$PA_{E2,L2} = \delta + \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur}. \tag{6}$$

For both models, fitting results returned coefficients, α and β, that are statistically significant (Column 5 in Tables 8 and 9). However, the intercept term (δ) in Eq. 6 is not statistically significant. In fact, adding an intercept does not change the values of the correlation coefficients. An ANOVA test showed that differences between $PA_{E2,L1}$ and $PA_{E2,L2}$ models are not statistically significant.

To understand how perceptual artifact strengths interact with one another, we also test a linear

Table 10 Exp. 2: Fitting parameters for the linear metric with interactions ($PA_{E2,L3}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr $(>\|t\|)$	PCC	SCC
α	0.874	0.029	30.059	$< 2e-16*$	0.975	0.966
β	0.747	0.024	31.551	$< 2e-16*$		
γ	-0.004	0.001	-3.105	$0.004*$		

Table 11 Exp. 2: Fitting parameters for the linear metric with interactions and intercept term ($PA_{E2,L4}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr $(>\|t\|)$	PCC	SCC
δ	-1.553	1.733	-0.896	0.373	0.975	0.966
α	0.899	0.040	22.396	$< 2e-16*$		
β	0.770	0.035	22.116	$< 2e-16*$		
γ	-0.005	0.001	-3.219	$0.002*$		

model with interactions, as given by:

$$PA_{E2,L3} = (\alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur} + \gamma \cdot MSV_{bloc} \cdot MSV_{blur}), \tag{7}$$

and, the same model with an intercept coefficient (δ), given by:

$$PA_{E2,L4} = (\delta + \alpha \cdot MSV_{bloc} + \beta \cdot MSV_{blur} + \gamma \cdot MSV_{bloc} \cdot MSV_{blur}). \tag{8}$$

Tables 10 and 11 show results for both fittings. For both models, the coefficients (α, β, and γ) are all statistically significant (Column 5 in Tables 10 and 11). Also, for both models, the correlation coefficients are slightly higher than those for the linear models with no interactions (see Eq. 5). However, the interaction term (γ) is negative. These results seem to indicate that there are masking effects among artifacts.

We also test the following Minkowski metric model:

$$PA_{E2,M1} = (\alpha \cdot MSV_{blo}^{m} + \beta \cdot MSV_{blu}^{m})^{\frac{1}{m}}, \tag{9}$$

Table 12 Exp. 2: Fitting parameters for Minkowski model ($PA_{E2,M1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($> \|t\|$)	PCC	SCC
m	1.341	0.132	10.190	9.99e-16*		
α	0.870	0.029	29.590	$< 2e - 16*$	0.975	0.965
β	0.693	0.030	22.820	$< 2e - 16*$		

where $PA_{E2,M1}$ is the predicted annoyance value and m is the Minkowski power obtained from the fit. Table 12 shows the fitting results for this model. Notice that the coefficients m, α and β are statistically significant (Column 5 in Table 12).

We also tested a Minkowski model with an intercept term ($PA_{E2,M2}$), but we found that the intercept term was not statistically significant. In other words, adding an intercept term to the Minkowski model did not change the values of the coefficients or the correlation coefficients. Since an ANOVA test showed that the differences between $PA_{E2,M1}$ and $PA_{E2,M2}$ models are not statistically significant, we decided not to show the results of this second model.

We also use an SVR algorithm ($PA_{E2,SVR}$) to predict annoyance from MSV_{bloc} and MSV_{blur}. The SVR fitting parameters and results are summarized in the third row of Table 4. Again, our tests show that using a radial kernel for the SVR provides the best performance. PCC and SCC values obtained from the trained SVR are 0.982 and 0.948, respectively.

3.3 Experiment 3: Packet-loss, Blockiness and Blurriness

In Experiment 3, we used test sequences with up to three different types of artifacts: packet-loss, blockiness, and blurriness. Again, results show that MSVs for the combination (0.0;0.0;0.0) (original video) are not equal to zero, indicating that subjects perceived impairments in unimpaired videos. Also, in general, participants correctly identified artifacts, giving highest MSVs to the corresponding strongest artifact and smaller MSVs to the other two artifacts (see Fig. 3).

For combinations with only one artifact, the highest MSVs correspond to the only artifact in

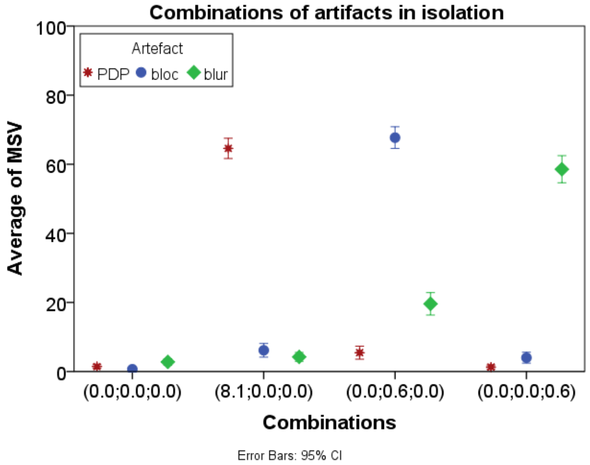


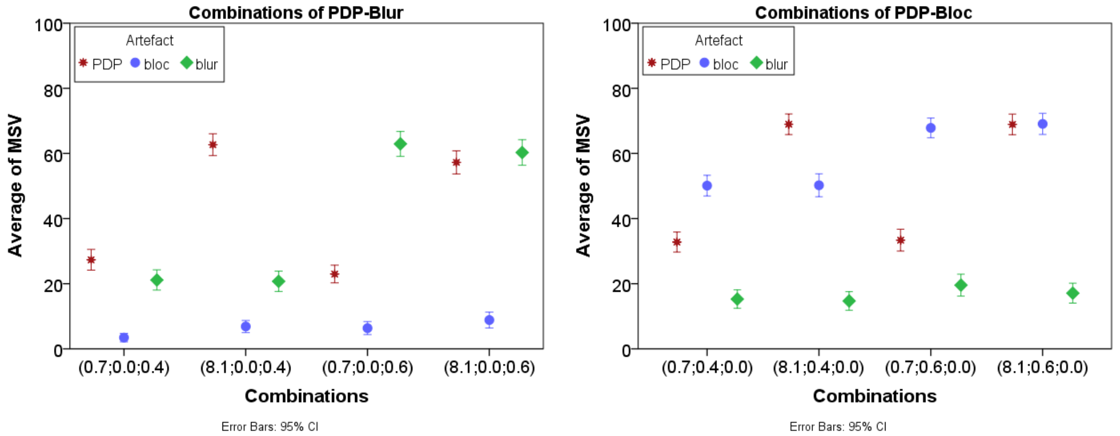
Fig 3 Exp. 3: MSV plot for clustered error for combinations (PDP;bloc;blur) for (0.0;0.0;0.0), (8.1;0.0;0.0), (0.0;0.6;0.0), and (0.0;0.0;0.6).

Table 13 Exp. 3: Pairwise comparisons between average MSVs for sequences with only -packet-loss, -blockiness, and -blurriness (*. Significant at 0.05 level.)

Combinations (PDP;bloc;blur)		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(0.0;0.0;0.6)	6.029*	2.401
(8.1;0.0;0.0)	(0.0;0.6;0.0)	-3.118	1.801
(0.0;0.6;0.0)	(0.0;0.0;0.6)	9.147*	2.206

the video (see Table 13). An RM-ANOVA shows that there are significant statistical differences in MSV, for any pair of combinations, with exception of the combination pair (8.1;0.0;0.0) and (0.0;0.6;0.0). The average MSV is slightly higher for blockiness, followed by packet-loss, and blurriness.

For combinations with two types of artifacts ((PDP;bloc;0.0), (PDP;0.0;blur), or (0.0;bloc;blur)), in most cases, the artifact signal corresponding to the highest signal strength receives the highest MSV. Nevertheless, an increase in the strength of a particular artifact signal does not always result in a proportional increase in this artifact perceived strength. For example, for (PDP;0.0;blur) combinations, an increase in the strength of blurriness causes a decrease in the perceived strength of the packet-loss (see Fig. 4 (a)).



(a) (b)

Fig 4 Exp. 3: MSV plots for clustered error for combinations (PDP;bloc;blur) for (a) (PDP;0;blur), and (b) (PDP;bloc;0).

Table 14 Exp. 3: Pairwise comparisons between average MSVs for (PDP;blur) sequences (*. Significant at 0.05 level.)

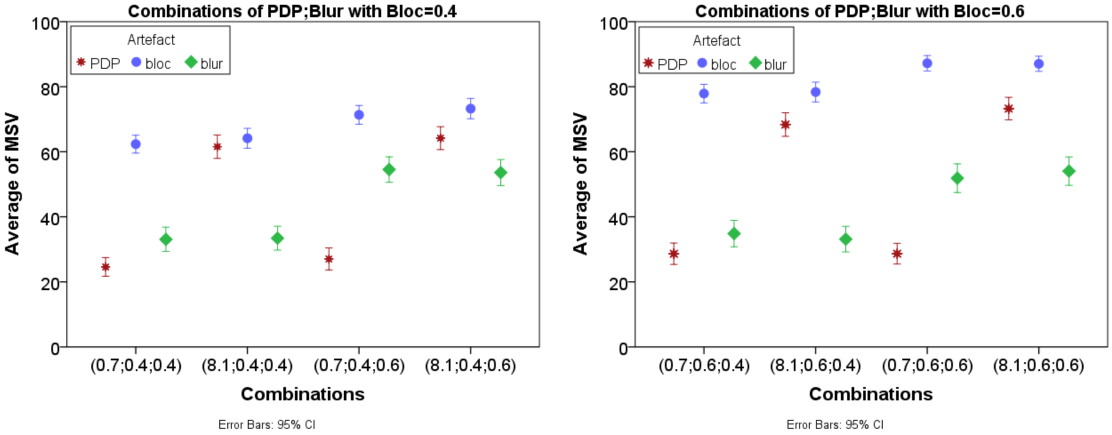
Combinations		MSV_{pck}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.0;0.4)	(8.1;0.0;0.4)	-35.351*	1.645	0.400	1.741
(0.7;0.0;0.4)	(0.7;0.0;0.6)	4.371*	1.636	-41.800*	2.097
(0.7;0.0;0.4)	(8.1;0.0;0.6)	-29.914*	1.753	-39.159*	2.193
(8.1;0.0;0.4)	(0.7;0.0;0.6)	39.722*	1.637	-42.200*	2.137
(8.1;0.0;0.4)	(8.1;0.0;0.6)	5.437*	1.614	-39.559*	2.116
(0.7;0.0;0.6)	(8.1;0.0;0.6)	-34.286*	1.685	2.641	1.843

An RM-ANOVA test shows that there are significant statistical MSV differences between all combinations of (PDP;0.0; blur). The only exceptions are the combination pairs ((0.7;0.0;0.4), (8.1;0.0;0.4)) and ((0.7;0.0;0.6), (8.1;0.0;0.6)), whose MSV_{blur} differences are not statistically significant (see Table 14). Notice that, for these two combinations, only the packet-loss strength changes while the blurriness strength is kept constant. This result suggests that blurriness may be masking the perceived strength of packet-loss.

The presence of packet-loss in the (PDP;bloc;0.0) combinations changes the perceived strength of the blockiness artifact (see Fig. 4 (b)). This indicates that increasing the packet-loss strength

Table 15 Exp. 3: Pairwise comparisons between average MSVs for (PDP;0;bloc) sequences (*. Significant at 0.05 level.)

Combinations		MSV_{pck}		MSV_{bloc}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.4;0.0)	(8.1;0.4;0.0)	-36.167*	1.652	-0.114	1.885
(0.7;0.4;0.0)	(0.7;0.6;0.0)	-0.576	1.757	-17.718*	1.613
(0.7;0.4;0.0)	(8.1;0.6;0.0)	-36.127*	1.795	-18.959*	1.855
(8.1;0.4;0.0)	(0.7;0.6;0.0)	35.592*	1.779	-17.604*	1.847
(8.1;0.4;0.0)	(8.1;0.6;0.0)	0.041	1.760	-18.845*	1.927
(0.7;0.6;0.0)	(8.1;0.6;0.0)	-35.551*	1.890	-1.241	1.515



(a) (b)

Fig 5 Exp. 3: MSV plots for clustered error for combinations (PDP;bloc;blur): (a) (PDP;blur) with bloc=0.4, (b) (PDP;blur) with bloc=0.6.

in a (PDP;bloc;0.0) combination can intensify the perceived strength of blockiness. This may be caused by the similarity of blockiness and packet-loss artifacts, which are both characterized by the presence of rectangular areas distributed over the video frames. An RM-ANOVA test (see Table 15) shows that there are significant statistical differences in MSV_{pck} for all combinations pairs (PDP;bloc;0.0). The only exceptions are the combination pairs ((0.7;0.4;0.0), (0.7;0.6;0.0)) and ((8.1;0.4;0.0), (8.1;0.6;0.0)). Another RM-ANOVA test shows that there are significant statistical differences in MSV_{bloc} values for the combination pairs ((0.7;0.4;0.0), (8.1;0.4;0.0)) and ((0.7;0.6;0.0), (8.1;0.6;0.0)).

Table 16 Exp. 3: Pairwise comparisons between average MSVs for (PDP;0.4;blur) sequences and changing packet-loss and blurriness strengths (*. Significant at 0.05 level.)

Combinations		MSV_{pck}		MSV_{bloc}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.4;0.4)	(8.1;0.4;0.4)	-36.976*	1.768	-1.788	1.501	-0.351	1.712
(0.7;0.4;0.4)	(0.7;0.4;0.6)	-2.445	1.781	-9.000*	1.437	-21.478*	2.049
(0.7;0.4;0.4)	(8.1;0.4;0.6)	-39.608*	1.833	-10.902*	1.664	-20.531*	2.161
(8.1;0.4;0.4)	(0.7;0.4;0.6)	34.531*	1.969	-7.212*	1.491	-21.117*	2.028
(8.1;0.4;0.4)	(8.1;0.4;0.6)	-2.633	1.699	-9.114*	1.767	-20.180*	2.171
(0.7;0.4;0.6)	(8.1;0.4;0.6)	-37.163*	1.916	-1.902	1.570	0.947	2.198

For combinations that correspond to videos with the three types of artifact signals, the average MSV_{bloc} is higher than the average MSV_{pck} and MSV_{blur}. Figs. 5 (a) and (b) show plots of combinations with different values of packet-loss, blockiness, and blurriness strengths. An RM-ANOVA showed that there are significant statistical differences between MSVs for most combinations of (PDP;bloc;blur). The combination pairs ((0.7;0.4;0.4), (0.7;0.4;0.6)) and ((8.1;0.4;0.4), (8.1;0.4;0.6)) are not statistically significant in MSV_{pck}, respectively. Although only the strength of blurriness vary in both combination pairs, MSV_{bloc} also increases as MSV_{blur} increases. This result suggests that the blockiness is affected by increasing the blurriness. For the combination pairs ((0.7;0.4;0.4),(8.1;0.4;0.4)) and ((0.7;0.4;0.6), (8.1;0.4;0.6)), the MSV_{bloc} and MSV_{blur} differences are not statistically significant. Notice that, for these combinations, the MSV variations are higher for MSV_{bloc} than for MSV_{blur} (see Table 16 columns 5 and 7). These results support the assumption that packet-loss artifacts can intensify the perception of blockiness artifacts.

When comparing MSVs for sequences with bloc = 0.6 and different PDP and blur values, an RM-ANOVA test shows that, for most combination pairs, the differences are statistically significant. For MSV_{pck}, only the difference for the pair ((0.7;0.6;0.4), (0.7;0.6;0.6)) is not statis-

Table 17 Exp. 3: Pairwise comparisons between average MSVs for (PDP;0.6;blur) sequences and changing packet-loss and blurriness strengths (*. Significant at 0.05 level.)

Combinations		MSV_{pck}		MSV_{bloc}		MSV_{blur}	
		Diff. Mean	Std. Error	Diff. Mean	Std. Error	Diff. Mean	Std. Error
(0.7;0.6;0.4)	(8.1;0.6;0.4)	-39.710*	1.942	-0.482	1.485	1.714	2.147
(0.7;0.6;0.4)	(0.7;0.6;0.6)	-0.020	2.085	-9.327*	1.249	-17.029*	2.312
(0.7;0.6;0.4)	(8.1;0.6;0.6)	-44.616*	1.990	-9.151*	1.310	-19.208*	2.339
(8.1;0.6;0.4)	(0.7;0.6;0.6)	39.690*	1.921	-8.845*	1.306	-18.743*	2.327
(8.1;0.6;0.4)	(8.1;0.6;0.6)	-4.906*	1.605	-8.669*	1.323	-20.922*	2.119
(0.7;0.6;0.6)	(8.1;0.6;0.6)	-44.596*	1.854	0.176	1.031	-2.180	2.430

tically significant. Again, the MSV_{bloc} and MSV_{blur} differences for combinations ((0.7;0.6;0.4), (8.1;0.6;0.4)) and ((0.7;0.6;0.6), (8.1;0.6;0.6)) are not statistically significant (see Table 17 columns 5 and 7).

We test a set of linear and non-linear models, fitting them on the MSV_{pck}, MSV_{bloc}, MSV_{blur}, and MAV data. The first linear model is a simple linear model, without any interaction term:

$$PA_{E3,L1} = \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur}. \tag{10}$$

Next, we adapt Eq. 10 to include an intercept coefficient (δ):

$$PA_{E3,L2} = \delta + \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur}. \tag{11}$$

Tables 18 and 19 show the fitting results for both models. Notice that all coefficients (i.e. δ, α, β, and γ) are statistically significant (see Columns 5 in Tables 18 and 19, respectively).

Since we are also interested in understanding if the perceptual strengths interact with one an-

Table 18 Exp. 3: Fitting parameters for linear model without intercept ($PA_{E3,L1}$) (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
α	0.340	0.022	18.330	$< 2e-16*$	0.937	0.936
β	0.470	0.020	23.210	$< 2e-16*$		
γ	0.413	0.026	16.04	$< 2e-16*$		

Table 19 Exp. 3: Fitting parameters for linear model with intercept ($PA_{E3,L2}$). (* Significant at 0.05 level.)

Coefficient	Estimate	Std. Error	t-value	Pr ($>|t|$)	PCC	SCC
δ	3.846	1.870	2.057	0.042*	0.937	0.937
α	0.370	0.026	14.313	$< 2e-16*$		
β	0.456	0.021	21.448	$< 2e-16*$		
γ	0.371	0.033	11.326	$< 2e-16*$		

other, we test a linear model with interactions, as given by:

$$\begin{aligned} PA_{E3,L3} = \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur} + \rho_1 \cdot MSV_{pck} MSV_{bloc} \\ + \rho_2 \cdot MSV_{pck} MSV_{blur} + \rho_3 \cdot MSV_{bloc} MSV_{blur} + \rho_4 \cdot MSV_{pck} MSV_{bloc} MSV_{blur}. \end{aligned} \tag{12}$$

We also adapt Eq. 12 to include an intercept coefficient (δ):

$$\begin{aligned} PA_{E3,L4} = \delta + \alpha \cdot MSV_{pck} + \beta \cdot MSV_{bloc} + \gamma \cdot MSV_{blur} + \rho_1 \cdot MSV_{pck} MSV_{bloc} \\ + \rho_2 \cdot MSV_{pck} MSV_{blur} + \rho_3 \cdot MSV_{bloc} MSV_{blur} + \rho_4 \cdot MSV_{pck} MSV_{bloc} MSV_{blur}. \end{aligned} \tag{13}$$

Tables 20 and 21 show the fitting results for both models. Notice that most first, second, and third order coefficients are statistically significant (Columns 5 in Tables 20 and 21, respectively). The exceptions are ρ_3 and ρ_4 in $PA_{E3,L3}$ (see Table 20), which correspond to the interaction of (bloc;blur) and (PDP;bloc;blur), respectively. Notice also that most second order coefficients are negative, what may indicate masking effects, i.e. when two artifacts are present, one of them may

Table 20 Exp. 3: Fitting parameters for the linear metric with interactions $PA_{L3,E3}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr $(>\|t\|)$	PCC	SCC
α	5.476e-01	3.572e-02	15.327	$< 2e-16*$	0.956	0.947
β	5.470e-01	4.535e-02	12.062	$< 2e-16*$		
γ	4.432e-01	3.530e-02	12.558	$< 2e-16*$		
ρ_1	-2.918e-03	1.054e-03	-2.768	0.006*		
ρ_2	-3.414e-03	1.321e-03	-2.585	0.011*		
ρ_3	-1.855e-04	1.277e-03	-0.145	0.885		
ρ_4	1.908e-05	2.834e-05	0.673	0.502		

Table 21 Exp. 3: Fitting parameters for the linear metric with interactions and an intercept term $PA_{L3,E4}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr $(>\|t\|)$	PCC	SCC
δ	-1.857e+01	2.768e+00	-6.710	5.22e-10*	0.965	0.957
α	8.516e-01	5.488e-02	15.516	$< 2e-16*$		
β	8.411e-01	5.888e-02	14.286	$< 2e-16*$		
γ	7.670e-01	5.713e-02	13.424	$< 2e-16*$		
ρ_1	-7.729e-03	1.161e-03	-6.654	6.93e-10*		
ρ_2	-8.740e-03	1.393e-03	-6.274	4.66e-09*		
ρ_3	-5.488e-03	1.360e-03	-4.036	9.17e-05*		
ρ_4	1.062e-04	2.778e-05	3.821	0.000*		

attenuate the strength of the other artifact(s). The interaction coefficient with highest magnitude corresponds to the interaction (PDP;blur). This suggests that packet-loss artifacts affect how blurriness artifacts are perceived.

Next, we test the weighted Minkowski metric, which includes weights for each individual artifact strength, as given by the following equation:

$$PA_{E3,M1} = (\alpha \cdot MSV_{pck}^{m} + \beta \cdot MSV_{bloc}^{m} + \gamma \cdot MSV_{blur}^{m})^{\frac{1}{m}}, \tag{14}$$

Table 22 Exp. 3: Fitting parameters for the Minkowski model $PA_{L3,M1}$ (* Significant at 0.05 level).

Coefficient	Estimate	Std. Error	t-value	Pr $(>\|t\|)$	PCC	SCC
m	1.993	0.143	13.960	$< 2e-16*$	0.969	0.963
α	0.387	0.023	17.130	$< 2e-16*$		
β	0.565	0.021	27.760	$< 2e-16*$		
γ	0.321	0.029	11.280	$< 2e-16*$		

where α, β, and γ are the weights for MSV_{pck}, MSV_{bloc}, and MSV_{blur}, respectively, and m is the Minkowski power. Table 22 shows the fitting results. Notice that all coefficients are statistically significant (Columns 5 in Table 22). Blockiness is the artifact with the highest impact on MAV, followed by packet-loss and blurriness.

Finally, we use SVR to predict annoyance (i.e. $PA_{E3,SVR}$) from MSV_{pck}, MSV_{blo} and MSV_{blu}. The fourth line of Table 4 summarizes the SVR results, with columns 2-5 showing the estimated parameters and columns 6-7 showing the PCC and SCC values for the fit. Once again, our tests show that using a radial kernel for the SVR provides the best performance. PCC and SCC values obtained from the trained SVR are 0.963 and 0.957, respectively.

3.4 Discussion

Results show that subjects were able to properly identify the strengths of the individual artifacts in all experiments. As expected, MSVs were affected by each of the artifact parameters. For example, in Experiment 1, PDP and M both affected the perceived strength, although PDP (percentage of packet loss) had a higher impact than M (duration). Also, subjects perceived blockiness as being stronger than packet-loss and blurriness. Finally, experimental data showed that there are masking and facilitation effects between the artifacts. For example, blurriness seems to mask packet-loss and intensify blockiness, while packet-loss seems to intensify blockiness.

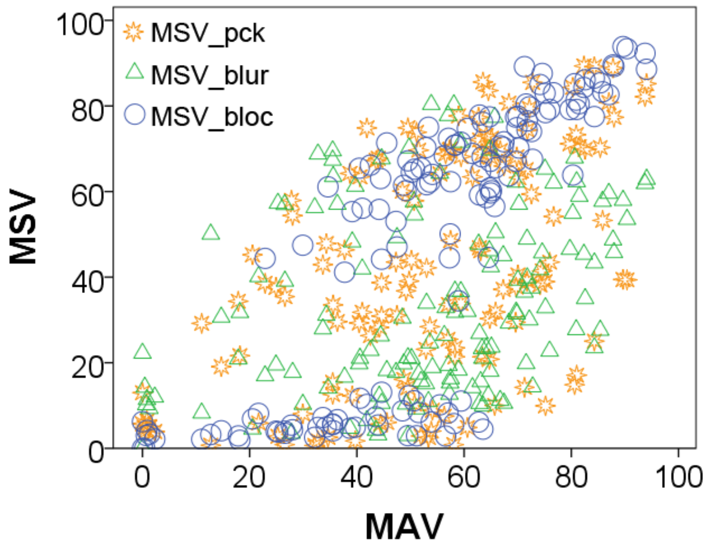


Fig 6 MAV versus MSV_{pck}, MSV_{bloc}, and MSV_{blur} for all three experiments.

It is worth pointing out that our set of stimuli contains sequences with combinations of up to three artifacts. For each sequence, we have one MAV (collected in a previous work), which corresponds to the overall annoyance of the set of artifacts in the sequence, and up to three MSVs (MSV_{bloc}, MSV_{blur}, and MSV_{pck}), which correspond to the perceptual strength of each type of artifact. Considering the data from the three experiments, we noticed that there is a correlation between the MAV and the individual MSVs. This can be observed in Fig. 6 that shows a graph of MSVs versus MAVs for each type of artifact (MSV_{bloc}, MSV_{blur}, and MSV_{pck}). The Pearson correlation coefficients between MAV and MSV_{bloc}, MSV_{blur}, and MSV_{pck} are 0.788, 0.337, and 0.536, respectively. The high correlation between the MAV and MSV_{bloc} confirms the importance of the blockiness artifact when predicting MAV.

In Fig. 6, the concentration of points close to the x-*axis* correspond to combinations that do not contain a specific artifact and, therefore, have small MSVs for at least one artifact. Points on the right side of the graph (higher MAVs) correspond, in general, to combinations with at least two artifacts. Notice that the MSV_{bloc} points (blue circles) on the top part of the graph ($MSV_{bloc} > 40$)

Table 23 Exp. 3: Average correlation across the 10-fold cross-validation runs between model predictions and MAVs

Model	PCC	SCC	Model	PCC	SCC	Model	PCC	SCC
$PA_{E1,L1}$	0.955	0.935	$PA_{E2,L1}$	0.968	0.912	$PA_{E3,L1}$	0.938	0.917
$PA_{E1,L2}$	0.955	0.935	$PA_{E2,L2}$	0.968	0.912	$PA_{E3,L2}$	0.938	0.918
			$PA_{E2,L3}$	0.975	0.929	$PA_{E3,L3}$	0.951	0.929
			$PA_{E2,L4}$	0.975	0.926	$PA_{E3,L4}$	0.975	0.926
			$PA_{E2,M1}$	0.975	0.965	$PA_{E3,M1}$	0.969	0.963
$PA_{E1,SVR}$	0.953	0.927	$PA_{E2,SVR}$	0.982	0.948	$PA_{E3,SVR}$	0.963	0.957

all have high MAVs, which explain their relatively good correlation with MAV. The data shows that the highest MAVs always have a high MSV_{bloc}, what is not necessarily true for MSV_{pdp} and MSV_{blur}. Although, as mentioned earlier, blockiness has a high impact on annoyance, MAV cannot be modeled as a function of a single artifact. As shown in this work, the annoyance model is a multidimensional function that must take into account the strengths of the most 'important' or relevant artifacts.

We tested several annoyance models, which combine the individual perceptual strengths of the artifacts to predict the overall annoyance. Overall, the models (linear models with interaction terms, SVR, and Minkowski) presented a good fit with the experimental data. Table 23 shows a summary of the correlation coefficients for all models tested in all three experiments. Notice that more complex models, i.e. models with interactions terms and the SVR, have the best correlation values.

As shown in Table 23, the different models achieved different degrees of accuracy, yet in most cases a higher accuracy came at the expense of an increased complexity. For example, models with interaction terms have more parameters to be fit than models without them and, although they may be more accurate, they can be more prone to overfitting. To compare models in terms of the trade-off between complexity and accuracy, we use the Akaike Information Criterion (AIC). Table 24

Table 24 Exp. 3: Akaike Information Criterion (AIC) for the linear and Minkowski models. A lower value indicates a better trade-off between model complexity and accuracy.

Experiment 1			Experiment 2			Experiment 3		
Model	df	AIC	Model	df	AIC	Model	df	AIC
$PA_{E1,L1}$	2	627.473	$PA_{E2,L1}$	3	517.883	$PA_{E3,L1}$	4	984.615
$PA_{E1,L2}$	3	622.126	$PA_{E2,L2}$	4	519.831	$PA_{E3,L2}$	5	982.327
			$PA_{E2,L3}$	4	510.451	$PA_{E3,L3}$	8	949.302
			$PA_{E2,L4}$	5	511.609	$PA_{E3,L4}$	9	910.218
			$PA_{E2,M1}$	4	509.032	$PA_{E3,M1}$	5	907.273

summarizes the AIC values computed for all models, where a model with lower AIC is preferred.

For Experiment 1, $PA_{E1,L2}$ has the best performance. For Experiments 2 and 3, $PA_{E2,M1}$ and $PA_{E3,M1}$ have the lowest AIC. Nevertheless, it was not possible to use a 10-fold cross-validation setting to compare these models to the other ones. Therefore, in this work, we compare only the linear models. Among the linear models in Experiment 2, $PA_{E2,M1}$ outperforms all models in terms of AIC, however its performance is very similar to the performance of $PA_{E2,M1}$, both in terms of AIC and correlation. Similar results were obtained for Experiment 3, where, among the linear models, $PA_{E3,L4}$ has the smallest AIC. Overall, this is a very interesting result, which indicates that interactions need to be taken into consideration when estimating the overall annoyance (or quality) of video sequences containing different types of distortions.

4 Conclusions

We presented the methodology, statistical analysis, and conclusions of three psychophysical experiments. The goals of these experiments were to measure the perceptual strengths and overall annoyance of videos with different combinations of blockiness, blurriness, and packet-loss artifacts. Mainly, in this work, we wanted to understand how the perceived strengths of 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.

Annoyance models were obtained by combining the artifact perceptual strengths using a weighted Minkowski model, a support vector regression (SVR) model, and a linear model. Performing an RM-ANOVA test, we found that all types of artifact signal strengths had a significant effect on MAV. The RM-ANOVA test also indicated that there are interactions among some of the artifact perceptual strengths. The non-linear SVR model provided greater correlation coefficients than the other tested models. In summary, results show that annoyance can be modeled as a multidimensional function of the individual artifact signal measurements.[2, 19, 33, 34]

These results indicate that a blind image quality assessment method, which is based on artifact measurements, is indeed a valid approach. But, 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 (perceptually) significant artifacts. For example, blockiness seems to have the biggest effect on MAV. Finally, 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.

Acknowledgments

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and by the University of Brasília (UnB).

References

1 A. F. Silva, M. C. Farias, and J. A. Redi, "Assessing the influence of combinations of blockiness, blurriness, and packet loss impairments on visual attention deployment," in *IS&T/SPIE Electronic Imaging*, *Proc. SPIE* **9394**, 93940Z–93940Z–11 (2015).

2 M. C. Farias and S. K. Mitra, "Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance," *Journal of Electronic Imaging* **21**(4), 043013–043013 (2012).

3 M. C. Farias, *No-Reference and Reduced Reference Video Quality Metrics: New Contributions*. PhD thesis, University of California, Santa Barbara, California (2004).

4 W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation* **22**(4), 297–312 (2011).

5 M. Shahid, K. Pandremmenou, L. P. Kondi, *et al.*, "Perceptual quality estimation of h. 264/avc videos using reduced-reference and no-reference models," *Journal of Electronic Imaging* **25**(5), 053012–053012 (2016).

6 M.-A. Kourtis, H. Koumaras, and F. Liberal, "Reduced-reference video quality assessment using a static video pattern," *Journal of Electronic Imaging* **25**(4), 043011–043011 (2016).

7 A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms : What does the future hold?," *Intern. Journal of Multimedia Tools and Applic., Vol:* **51**, 675–696 (2011).

8 M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Visual Communications and Image Processing 2003*, 583–592, International Society for Optics and Photonics (2003).

9 A. Ahumada and C. H. Null, "Image quality: A multidimensional problem," *Digital images and human vision* , 141–148 (1993).

10 V. Kayargadde and J.-B. Martens, "Perceptual characterization of images degraded by blur and noise: model," *JOSA A* **13**(6), 1178–1188 (1996).

11 M. Nijenhuis and F. Blommaert, "Perceptual error measure for sampled and interpolated images," *Journal of Imaging Science and Technology* **41**(3), 249–258 (1997).

12 H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, 16–26, International Society for Optics and Photonics (1992).

13 H. de Ridder and G. M. Majoor, "Numerical category scaling: an efficient method for assessing digital image coding impairments," in *SC-DL tentative*, 65–77, International Society for Optics and Photonics (1990).

14 H. Liu and I. Heynderickx, "A perceptually relevant no-reference blockiness metric based on local image characteristics," *EURASIP Journal on Advances in Signal Processing* **2009**, 2 (2009).

15 R. V. Babu, A. Perkis, and O. I. Hillestad, "Evaluation and monitoring of video quality for uma enabled video streaming systems," *Multimedia Tools and Applications* **37**(2), 211–231 (2008).

16 J. Caviedes and J. Jung, "No-reference metric for a video quality control loop," *Proc* **13**, 290–5 (2001).

17 Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of

jpeg compressed images," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, **1**, I–477, IEEE (2002).

18 P. Marziliano, F. Dufaux, S. Winkler, *et al.*, "Perceptual blur and ringing metrics: application to jpeg2000," *Signal processing: Image communication* **19**(2), 163–172 (2004).

19 M. C. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, **3**, III–141, IEEE (2005).

20 M. C. Farias, J. M. Foley, and S. K. Mitra, "Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts," *IEEE Trans. on Signal Processing, v* **55**, 2954–2964 (2007).

21 A. F. Silva, M. C. Farias, and J. A. Redi, "Annoyance models for videos with spatio-temporal artifacts," In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon (2016).

22 A. F. Silva, M. C. Farias, and J. A. Redi, "Perceptual annoyance models for videos with combinations of spatial and temporal artifacts," *IEEE Transactions on Multimedia* (2016).

23 V. Q. E. G. (VQEG), "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i," tech. rep., Video Quality Experts Group (VQEG) (2008).

24 A. Ostaszewska and R. Kloda, "Quantifying the amount of spatial and temporal information in video test sequences," in *Recent Advances in Mechatronics, Springer*, 11–15 (2007).

25 International Telecommunication Union, *ITU-T Recommendation P.930: Principles of a reference impairment system for video* (1996).

26 M. C. Farias, J. M. Foley, and S. K. Mitra, "Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing synthetic artifacts," in *Electronic Imaging 2005*, 107–118, International Society for Optics and Photonics (2005).

27 International Telecommunication Union, *ITU-T Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures* (1998).

28 H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, 199–213, Springer (1998).

29 P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*, 532–538, Springer (2009).

30 A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing* **14**(3), 199–222 (2004).

31 J. A. Redi, I. Heynderickx, B. L. M. Espinoza, *et al.*, "On the impact of packet-loss impairments on visual attention mechanisms," in *ISCAS: International Symposium on Circuits and Systems*, 1107–1110, IEEE (2013).

32 M. C. Q. Farias, I. Heynderickx, B. L. M. Espinoza, *et al.*, "Visual artifacts interference understanding and modeling (varium): A project overview," in *Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, VQPM (2013).

33 H. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters* **4**(11), 317–320 (1997).

34 Z. Wang, A. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE International Conference on Image Processing*, **3**, 981–984 (2000).

Alexandre F. Silva is a professor at the Federal Institute of Triângulo Mineiro. He received his M.Sc. and Ph.D. degrees in computer science from the University of Uberlândia (UFU) in 2006 and from University of Brasília in 2017, respectively. He also spent 7 months in Delft University of Technology (TU Delft), in Netherlands as a Ph.D. researcher. His current research interests include video quality assessment, video quality metrics and visual attention.

Mylene C.Q. Farias is an Associate professor at the University of Brasilia (UnB). She received her Ph.D. in electrical and computer engineering from the University of California Santa Barbara (UCSB), in 2004. She has worked as a research engineer at CPqD (Brazil) in video quality. She has worked as an intern for Philips Research Laboratories (NL) and for Intel Corporation (USA) developing NR video quality metrics. Her current interests include multimedia signal processing, watermarking, and visual attention. She is an IEEE, ACM, and SPIE member.

List of Figures

List of Tables

22 Exp. 3: Fitting parameters for the Minkowski model $PA_{L3,M1}$ (* Significant at 0.05 level).

23 Exp. 3: Average correlation across the 10-fold cross-validation runs between model predictions and MAVs

24 Exp. 3: Akaike Information Criterion (AIC) for the linear and Minkowski models. A lower value indicates a better trade-off between model complexity and accuracy.