Perceptual Annoyance Models for Videos with Combinations of Spatial and Temporal Artifacts

Alexandre F. Silva and Mylène C.Q., *Member, IEEE,* Judith A. Redi

***Abstract*— Understanding the perceptual impact of compression artifacts in video is one of the keys for designing better coding schemes and appropriate visual quality control chains. Although compression and transmission artifacts, such as blockiness, blurriness and packet-loss, appear simultaneously in digital videos, traditionally they have been studied in isolation. In this paper, we report the results of three subjective quality assessment experiments aimed at studying perceptual characteristics of a set of artifacts common in digital videos. With this goal, first we study the annoyance of each of three artifacts (blockiness, blurriness and packet-loss) in isolation and then in combination. Based on the subjective evaluations, we design several models of the annoyance caused by the joint presence of these three artifacts on digital video.**

***Index Terms*— Video quality assessment, compression artifacts, subjective quality, human visual system modeling.**

1. INTRODUCTION

In modern digital imaging systems, the quality of the visual content can undergo a drastic decrease due to impairments introduced during capture, transmission, storage and/or display, as well as by any signal processing algorithm that may be applied to the content along the way (e.g., compression). Impairments are defined as visible defects (flaws) and can be decomposed into a set of perceptual features called *artifacts* [1]. Being able to detect artifacts and reduce their strength can improve the quality of the visual content prior to its delivery to the user [2], [3].

Visual quality assessment methods can be divided into two categories: subjective and objective methods. In subjective methods, the quality of a video is measured by performing psychophysical experiments with human subjects [1]. Objective quality methods, on the other hand, are algorithms (metrics) that aim at predicting visual quality as perceived by human observers. Subjective methods are considered most reliable and are frequently used to provide "ground truth" quality scores. These methods also provide insights into mechanisms of the human visual system related to the user quality of experience, inspiring, not only the design of objective quality metrics, but of all kinds of multimedia applications [4]. Nevertheless, subjective methods are expensive, time-consuming and, contrary to objective methods, cannot be easily incorporated into an automatic quality of service control system.

Objective quality metrics that take into account aspects of the human visual system usually have the best performance [5], [6]. This type of algorithm is often computationally expensive and, therefore, hardly applicable in real-time contexts [7]. Alternatives include artifact metrics [8]–[10], which estimate the strength of individual artifacts and, then, combine them to obtain an overall annoyance or quality model. The assumption here is that, instead of trying to estimate overall annoyance, it is easier to detect individual artifacts and estimate their strength because we 'know' their appearance and the type of process that generates them. These metrics have the advantage of being simple and not necessarily requiring the reference. They can be useful for post-processing algorithms, providing information about which artifacts need to be mitigated. Their design requires a good understanding of the perceptual characteristics of each artifact, as well as the knowledge of how each artifact contributes to the overall quality.

Little work has been done on studying and characterizing the individual artifacts [11]–[13], as pointed out by Moorthy and Bovik [6]. Farias et al. [14], [15] studied the appearance, annoyance, and detectability of common digital video compression *spatial* artifacts by measuring the strength and overall annoyance of these artifact signals when presented alone or in combination in interlaced Standard Definition (SD) videos (480i). The presence of noisiness in videos seemed to decrease the perceived strength of other artifacts, while the addition of blurriness had the opposite effect. Moore et al. [16] investigated the relationships among visibility, content importance, annoyance, and strength of *spatial* artifacts in interlaced SD videos. Their results show that the artifacts' annoyance are closely related to their visibility, but only weakly related to the video content.

Huynh-Thu and Ghanbari [17] examined the impact of *spatio-temporal* artifacts in video and their mutual interactions. They verified that spatial degradations affected the perceived quality of temporal degradations (and vice-versa). Moreover, the contribution of spatial degradations to the quality is greater than the contribution of temporal degradations. Nevertheless, as shown by Reibman et al. [18], temporal artifacts, like packet-loss, have an important contribution to quality and can be successfully used to predict it. Zhai et al. [19] have studied the perceptual quality of low bit-rate videos considering multiple dimensions. Differently from the previous works, their work does not focus on specific types of artifacts, but on different settings for video codecs, such as encoder type, video content, bit rate, frame size, and frame rate. More specifically, the authors performed a series of experiments that allowed to establish which codec settings had the greatest impact on quality. Naccari et al. [20], on the other hand, modeled the effects of spatial and temporal error concealment, the loss of

A.F. Silva and M.C.Q. Farias are with the University of Brasilia, Brasilia, Brazil e-mail: (alexandrefieno@gmail.com, mylene@ieee.org). J.A. Redi is with the Delft University of Technology, Delft, The Netherlands (e-mail: J.A.Redi@tudelft.nl). This work was supported by CAPES-Brazil and by Nuffic-Netherlands.

prediction residuals, and the temporal distortion propagation due to the motion-compensation loop.

Despite these earlier studies, there is no clear knowledge on how different *spatial* and *temporal* artifacts combine perceptually and how their joint impact depends on the properties of high definition videos and of the artifacts themselves. In this paper, we investigate how spatial and temporal artifacts combine to determine quality [21], [22]. With this goal, we present results of three psychophysical experiments that measure annoyance and detection characteristics of two spatial artifacts (blockiness and blurriness) and a very important temporal artifact (packet-loss). The artifacts appear in isolation or in combination. Up to our knowledge, there is no study in the literature that performs an analysis of the influence spatial-temporal artifacts (in isolation and in combinations) have on the perceived annoyance. Most importantly, there is no study on how spatial and temporal artifacts interact to produce overall annoyance. To quantify the contribution of each artifact to the overall annoyance and of the interactions among the different artifacts, we test linear and non-linear annoyance models. Also, as a contribution of this project, a diverse high-definition (720p) video database is made publicly available [23].

The paper is divided as follows. Sections 2 and 3 describe the method used to generate the test sequences and the experimental methodology. Sections 4 to 7 describe the experiments and discuss their results. Section 8 presents the conclusions.

2. Generation of Test Sequences

Our goal is to find a perceptual model that describes how the physical strengths of individual artifacts determine the overall annoyance or quality of the video. To achieve this goal, we performed psychophysical experiments using a set of test sequences with several combinations of blurriness, blockiness, and packet-loss artifacts at different strengths. Although this set of artifacts is not exhaustive, these three artifacts are known to be among the most commonly encountered in video applications. Seven ten-seconds high-definition videos with spatial resolution 1280×720 (50fps) were used in the experiments. The videos were chosen following the recommendations of VQEG that stated that the set of originals must had a good distribution of spatial and temporal activity [24].

It is difficult to generate artifacts in isolation using compression algorithms, such as H.264 or H.265 codecs. When videos are compressed, several types of artifacts (i.e., blockiness, blurriness, ringing, etc.) are simultaneously introduced, at different levels of strengths. Even if we used artifact metrics to measure the strengths of each individual artifact, it would not be possible to control the combination of artifacts or the strengths of each artifact in the compressed video. It is worth pointing out that the outputs of the artifact metrics would not be directly (or linearly) related to the artifact perceptual strengths, due to (1) the fact that artifact metrics tend to measure annoyance rather than strength and (2) the relative low accuracy of current artifact metrics, specially in multiple artifact scenarios. Therefore, it would be complicated to produce the videos with the combinations of artifacts necessary to obtain an annoyance model, which describes how artifacts combine to produce annoyance and what are the perceptual interactions among artifacts.

Although most psychophysical studies varies the presence and strength of artifacts by changing the bitrate and/or the codec implementation [1], [25], some studies generate each artifact individually and control their strength by linearly changing their amplitude [16], [26]. For example, Libert et al. [26] compared the presence and strength of artifacts by changing the bitrate and/or the codec implementation and they concluded that the later can validly approximate the changes produced by varying the bit-rate goal. In this work, we used a previously developed system to impair our video stimuli [14], [15], [27], which was based on ITU Recommendation P.930 [28]. We decided to simulate artifacts, rather than obtaining them by applying compression algorithms or transmission systems, because we wanted to be able to control the artifacts individually and to combine them arbitrarily, i.e. to control both the appearance and the strength of each artifact. This allows us to measure artifact psychophysical characteristics, when presented in isolation or in combination, and test several annoyance models [27].

A. Test sequences

Blockiness is the appearance of the underlying block encoding structure of typical compression schemes. It is often caused by coarse quantization of the spatial frequency components during the encoding process [28]. The algorithm used for generating blockiness calculates the average value of each 8×8 block of the frame and the average of the 24×24 surrounding block and adds the difference between these two averages to the block [14], [15]. Blurriness is characterized by a loss of spatial details and a reduction in sharpness [28]. Recommendation P.930 suggests generating blurriness using a simple low-pass filter [14], [15]. To control the amount of blurriness, we can vary the filter sizes and the cut-off frequencies. In this work, we used a 5×5 moving average filter to generate blurriness.

As the name suggests, packet-loss artifacts are caused by a complete loss of the packet being transmitted, as a consequence of digital transmission errors. As a consequence, parts (blocks) of the video are missing for several frames. To generate test sequences with packet-loss, we used the reference H.264 codec. To avoid inserting additional artifacts, we compressed the original videos at high compression rates, generating high quality videos. To vary the "packet-loss" strength levels, we randomly deleted packets from the coded video bitstream with different loss percentages (the higher the percentage, the lower the quality) and changed the interval between I-frames (time interval among artifacts).

The algorithm for generating test sequences consists of the following steps [14], [15]. First, we generated videos with one type of artifact signal at a high level of annoyance. These maximum levels of artifact annoyance were established in a previous experiment, in which the artifact strengths were matched to the perceptual strengths of real digital video artifacts [14]. Then, the test sequences (Y) with a combination of blockiness and blurriness artifacts were generated by linearly combining the original video with the videos with

(a) Joy Park (b) Into Trees (c) Crowd Run (d) Romeo & Juliet (e) Cactus (f) Basketball (g) Barbecue

Figure 1: Sample frames of original videos.

very strong blockiness and blurriness artifacts (X_{bloc} and X_{blr}, respectively). To create a test sequence Y with both blockiness and blurriness, we used the following expression:

$$Y = X_0 + s_1 \cdot (X_0 - X_{bloc}) + s_2 \cdot (X_0 - X_{blur}), \quad (1)$$

where X_0 was the original video and s_1 and s_2 ($0 \leq s_1, s_2 \leq 1$) were the relative strength parameters corresponding to the blockiness and blurriness signals, respectively. In general $s_1 + s_2 \leq 1$, but, in some cases, we allowed $s_1 + s_2 \geq 1$ to make artifacts stronger. To generate a sequences with combinations of blockiness, blurriness, and packet-loss, we first generated videos with a combination of blockiness and blurriness and, then, we inserted packet-loss artifacts using the packet-deletion procedure described earlier.

3. Experimental Setup

We designed an experimental procedure composed of three psychophysical experiments. In the first experiment, we examined packet-loss artifacts in isolation. In the second experiment, we studied blockiness and blurriness artifacts in isolation and in combination. Finally, in the third experiment, we combined blockiness, blurriness, and packet-loss to study their joint perceptual impact on quality. Although these three experiments were performed at different times and with different participants, the experimental protocol and the environmental conditions were the same for all of them.

The experiments were performed using a PC computer with a Samsung LCD monitor of 23 inches (Sync Master XL2370HD), with the dynamic contrast of the monitor turned off, the contrast set to 100, and the brightness set to 50. The measured gamma of the monitor for luminance, red, green, and blue were approximately 1.937, 1.566, 1.908, and 1.172, respectively. The room had the lights dimmed to avoid reflection and the experiments were run with one subject at a time. The subjects were seated straight ahead of the monitor, centered at or slightly below eye height for most subjects. The distance between the subject's eyes and the video monitor was 3 video monitor screen heights [1]. We used a chin rest to guarantee that the distance between the subject's eye and the monitor remained the same.

Subjects were volunteers from the Delft University of Technology, The Netherlands, and from the University of Brasília, Brazil. Most subjects were considered naive of most digital video defects and the associated terminology. No vision test was performed, but subjects were asked to wear glasses or contact lenses if they needed them. Before starting the experiment, the experimenter made sure the subject was properly seated at the adequate distance. Subjects were then explained the tasks to be performed and told to disregard the content of the videos and judge only the impairments they saw.

Since initial judgments are generally erratic, ITU recommends that the first five to ten trials be discarded [1]. Instead of discarding the first trials, we included *practice trials* with a set of at least 5 test sequences. First, subjects watched videos with strong impairments and, then, they rated the annoyance of a separate set of videos (not included in the main session). Besides eliminating erratic answers, practice trials exposed subjects to a good range of impairments and gave them a chance to try the scoring interface.

Experimental trials were performed with the complete set of test sequences presented in a random order. Videos were played once and subjects were not allowed to go back and watch them again. The experiment used the ITU BT.500 Single Stimulus (SS) with hidden reference methodology [1]. Subjects were instructed to search each video for impairments and to perform detection and annoyance tasks. The detection task consisted of detecting impairments in the video sequence. After each test sequence was played, the question "Did you see a defect or an impairment?" appeared in the monitor. The subject was supposed to choose a 'yes' or 'no' answer. Only if they detected impairments, they performed the annoyance task. The annoyance task consisted of giving a score of how annoying the detected impairment was. The scoring was performed on a continuous numerical annoyance scale, ranging between 0 and 100. Any defect as annoying as the worst impairments in the practice stage should be given '100', half as annoying '50', ten percent as annoying '10' and so forth. Videos with no detected impairments were automatically assigned an annoyance score of 0.

The mean opinion score (MOS) across subjects is used as a measure of the annoyance caused by the impairments. In the following sections, we refer to it as Mean Annoyance Value (MAV). Also, the Probability of detection (P_{det}) is computed taking the ratio of subjects who saw the impairment over the total number of subjects.

4. Experiment 1: Packet-Loss

In Experiment 1, 16 subjects rated the annoyance of test sequences containing only packet-loss artifacts. As mentioned earlier, to vary the strength of the artifacts, we randomly deleted packets from the coded video bitstream. The Percentages of Deleted Packets (PDP) used were 0.7%, 2.6%, 4.3%, and 8.1%. To vary the time interval among introduced artifacts, we varied the number of frames between the I-frames. Three frame intervals (M) were used: 4, 8 and 12. The set of PDP and M parameters used in this experiment are given in Table I. A total of 7 originals and 12 parameter combinations were used, resulting in $12 \times 7 + 7 = 91$ test sequences. To avoid fatigue, these videos were evaluated in a single experimental session, divided in three sub-sessions with two 10-minutes breaks.

We analyze the probability of detection (P_{det}) for all test sequences of Experiment 1. Results show that P_{det} increases

TABLE I: Exp. 1: Combinations of the parameters PDP and M used for each of the 7 originals.

Comb	M	PDP	Comb	M	PDP	Comb	M	PDP
1	4	0.7	5	8	0.7	9	12	0.7
2	4	2.6	6	8	2.6	10	12	2.6
3	4	4.3	7	8	4.3	11	12	4.3
4	4	8.1	8	8	8.1	12	12	8.1

TABLE II: Exp. 1: Pairwise comparisons between average MAVs for different M values. (* Significant at 0.05 level.)

M values		Diff. Mean	Std. Error
4	8	-0.170	1.512
	12	-9.134*	1.664
8	12	-8.964*	1.946

with MAV. In particular, some videos gathered all P_{det} values equal to one (e.g., 'Into Tree' and 'Barbecue'). This means that for these two originals, all subjects saw impairments in all test cases. It is worth pointing out that these two scenes have large smooth regions (e.g. skies) that make impairments easier to detect. 'Park Joy', 'Cactus' and 'Basketball' have values of P_{det} that grow (and saturate) very fast as MSE increases. On the other hand, for 'Park Run' and 'Romeo & Juliet' P_{det} increases at a slower rate. This indicates that, for these scenes, it is harder to detect packet-loss artifacts. 'Romeo & Juliet', although having small spatial and temporal activity, is relatively dark and has a very clear focus of attention (the couple). On the other hand, 'Park Run' has lots of spatial and temporal activity and not a lot of camera movement. All of this makes it harder to spot packet-loss artifacts.

Figure 2 shows a plot of the average MAVs for the three values of M and the four values of PDP. Notice that MAV increases with both PDP and M, but PDP has a bigger effect on MAV than M. The effect of PDP on MAV is clearly significant. But, we analyzed the influence of M on MAV by performing a repeated-measure ANOVA (RM-ANOVA) with significance level of 95% ($\alpha = 0.05$). Table II shows the pairwise comparisons between average MAVs for different M parameters. Notice that there are significant statistical differences between average MAVs for any pair of M values, with exception of the pair M = 4 and M = 8 for PDP=0.7%.

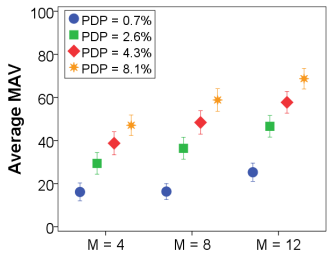

Figure 2: Exp. 1: Average MAV plots for different values of PDP: 0.7%, 2.6%, 4.3% and 8.1%.

5. Experiment 2: Blockiness and Blurriness

In Experiment 2, 16 subjects performed annoyance and detection tasks on test sequences containing combinations of blockiness and blurriness artifacts. Strength combinations are represented by a vector (bloc; blur), where 'bloc' is the blockiness strength and 'blur' is the blurriness strength. The experiment contained a set of videos with all possible combinations of the two artifact types at strengths 0.0, 0.4, and 0.6 (full factorial design: $3^2 = 9$). Two additional combinations, consisting of pure blockiness and pure blurriness at strength 0.8, were added to the experiment. Table III shows all combinations used in the experiment. A total of 7 originals and 11 combinations were used in this experiment, resulting in $11 \times 7 + 7 = 84$ test sequences. To avoid fatigue, these videos were evaluated in a single experimental session, divided in three sub-sessions by two 10-minutes breaks.

TABLE III: Exp. 2: Set of combinations used for each of the 7 originals: 'bloc' and 'blur' correspond to the blockiness and blurriness strengths, respectively.

Comb	(bloc;blur)	Comb	(bloc;blur)	Comb	(bloc;blur)
1	(0.0;0.0)	5	(0.4;0.4)	9	(0.6;0.6)
2	(0.0;0.4)	6	(0.4;0.6)	10	(0.0;0.8)
3	(0.0;0.6)	7	(0.6;0.0)	11	(0.8;0.0)
4	(0.4;0.0)	8	(0.6;0.4)		

TABLE IV: Exp. 2: Pairwise comparisons of MAVs for videos with only blockiness ($\hat{F}$=85.62, $\alpha \leq$0.01) and only blurriness ($\hat{F}$=334.75, $\alpha \leq$0.01). (* Significant at 0.05 level)

		Blockiness		Blurriness	
Strengths		Diff. Mean	Std. Error	Diff. Mean	Std. Error
0.4	0.6	-22.982*	1.863	-22.295*	2.796
	0.8	-33.125*	3.179	-66.107*	2.526
0.6	0.8	-10.143*	2.571	-43.813*	2.464

The values of P_{det} for all original videos are smaller than 0.2, except for the original 'Into Tree', which has large smooth regions. Similarly to Experiment 1, test sequences with low P_{det} values got lower MAVs, while test sequences with higher P_{det} values got higher MAVs. Figures 3(a) and (b) show plots of the average MAVs for sequences with only blockiness and blurriness, respectively, at strengths 0.4, 0.6, and 0.8. As expected, average MAVs increase with the artifact strength. We performed an RM-ANOVA to check if MAV differences for different blockiness and blurriness strengths are significant. Table IV displays the results, showing that there are significant statistical differences for all pairs of different strengths in only-blockiness and only-blurriness sequences.

Figure 3(c) shows a plot of the average MAVs for sequences with combinations of blockiness and blurriness. Table V shows the results of an RM-ANOVA test that performed pairwise comparisons of the average MAVs of these sequences. Results show that there are significant statistical differences between average MAVs obtained for any pair of blockiness and blurriness combinations ($\hat{F} = 124.68$, $\alpha \leq 0.01$), with the only exception of the pair (0.4;0.6) and (0.6;0.4). This means that a change in the artifact strength was perceived by subjects.

We also performed an analysis to check if there are differences among sequences with one or with two artifacts. Results of the pairwise comparisons of average MAVs between blockiness+blurriness and of only-blockiness sequences are shown in Table VI ($\hat{F} = 141.78$, $\alpha \leq 0.01$ for (0.4;0.0) and $\hat{F} = 151.13$, $\alpha \leq 0.01$ for (0.6;0.0)), while results of the pairwise comparisons of average MAVs between block-

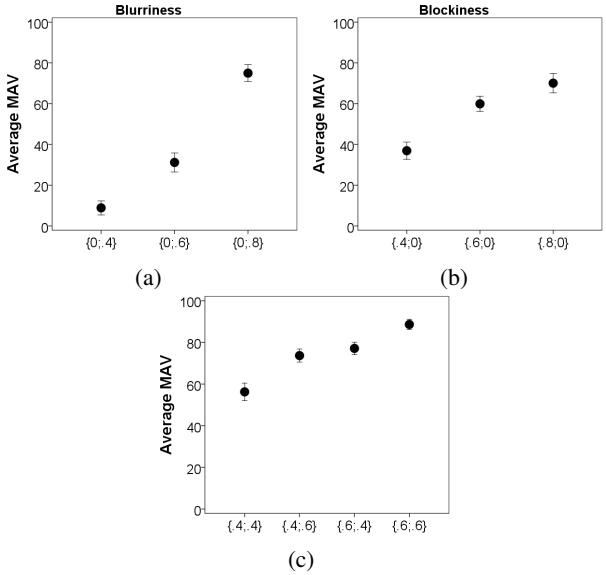

Figure 3: Exp. 2: Average MAVs for: (a) blurriness, (b) blockiness, and (c) combinations of blockiness and blurriness.

iness+blurriness and only-blurriness sequences are shown in Table VII ($\hat{F} = 520.12,\ \alpha \leq 0.01$ for (0.0;0.4) and $\hat{F} = 358.22,\ \alpha \leq 0.01$ for (0.0;0.6)). Notice that there are significant statistical differences between average MAVs for any pair of these combinations. In other words, on average, adding an extra artifact affected the MAV.

TABLE V: Exp. 2: Pairwise comparisons of MAVs of sequences with combinations of blockiness and blurriness. (* Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.4;0.4)	(0.4;0.6)	-17.420*	2.044
	(0.6;0.4)	-20.866*	1.841
	(0.6;0.6)	-32.375*	2.044
(0.4;0.6)	(0.6;0.4)	-3.446	1.499
	(0.6;0.6)	-14.955*	1.445
(0.6;0.4)	(0.6;0.6)	-11.509*	1.097

TABLE VI: Exp. 2: Pairwise comparisons of MAVs between sequences with only blockiness and sequences with combinations of blockiness and blurriness. (* Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.4;0.0)	(0.4;0.4)	-19.330*	2.027
	(0.4;0.6)	-36.750*	2.453
(0.6;0.0)	(0.6;0.4)	-17.214*	1.844
	(0.6;0.6)	-28.723*	1.921

TABLE VII: Exp. 2: Pairwise comparisons of MAVs between sequences with only blurriness and sequences with combinations of blockiness and blurriness. (* Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.4)	(0.4;0.4)	-47.393*	2.492
	(0.6;0.4)	-68.259*	2.124
(0.0;0.6)	(0.4;0.6)	-42.518*	2.507
	(0.6;0.6)	-57.473*	2.553

6. Experiment 3: Blockiness, Blurriness and Packet-Loss

In Experiment 3, 23 subjects performed annoyance and detection tasks on test sequences containing different combinations of blockiness, blurriness, and packet-loss artifacts. The strength combinations are represented as a vector (PDP;bloc;blur), where 'PDP' is the level of packet-loss strength, 'bloc' is the level of blockiness strength, and 'blur' is the level of blurriness strength.

Considering the results from the previous experiments, we selected a subset of artifact strength values to limit the number of artifact combinations. For packet-loss ratio, we chose M = 12 because this was the most realistic setting for the GOP size, which is recommended to be at most half of the frame rate. Also, we chose PDP = 0.7% and 8.1% because these values corresponded to the highest differences in annoyance (as shown in the analysis of Experiment 1). With respect to blockiness and blurriness, we chose strength values equal to 0.4 and 0.6, which were considered to be more representative of these artifacts (as shown in the analysis of Experiment 2). Table VIII shows all combinations used in this experiment, which include three strengths for each artifact type. Again, 7 originals and 19 combinations were used, resulting in $19 \times 7 + 7 = 140$ test sequences. To avoid fatigue, these videos were evaluated in a single experimental session, divided in three sub-sessions by two 10-minutes breaks.

All original videos got P_{det} values below 0.09, with the exception of the video 'Park Run' ($P_{det} = 0.17$). 'Park Run' has a lot of spatial and temporal activity and not a lot of camera movement, what could have led some subjects to think they saw impairments in the originals. Similarly to Experiment 1 and 2, test sequences with low P_{det} values got lower MAVs, while test sequences with higher P_{det} values got higher MAVs.

Figure 4 show plots of average MAV over all test sequences with pure strong blockiness (0.0;0.6;0.0), blurriness (0.0;0.0;0.6), and packet-loss (8.1;0.0;0.0). For comparison purposes, the plot also shows the average MAVs for the original sequences. As expected, the average MAV for originals is close to zero and when the artifact strengthens, MAV increases. Average MAV values are higher for blockiness (average MAV for bloc= 0.6 is 48.56), followed by packet-loss (average MAV for PDP=8.1% is 37.99), and blurriness (average MAV for blur=0.6 is 32.45). This is in agreement with results of Experiment 2, where blockiness artifacts are the most annoying artifacts. To check if these average MAVs differences between different artifacts were statistically significant, we performed an RM-ANOVA. Results in Table IX show that

TABLE VIII: Exp. 3: Combinations for each original: 'bloc' corresponds to the blockiness strength, 'blur' to the blurriness strength, and 'PDP' to the packet-loss ratio.

Comb.	(PDP;Bloc;Blur)	Comb.	(PDP;Bloc;Blur)	Comb.	(PDP;Bloc;Blur)
1	(0.0;0.0;0.0)	8	(8.1;0.0;0.6)	15	(0.7;0.6;0.0)
2	(0.0;0.6;0.0)	9	(0.7;0.4;0.0)	16	(8.1;0.6;0.0)
3	(0.0;0.0;0.6)	10	(8.1;0.4;0.0)	17	(0.7;0.6;0.4)
4	(8.1;0.0;0.0)	11	(0.7;0.4;0.4)	18	(8.1;0.6;0.4)
5	(0.7;0.0;0.4)	12	(8.1;0.4;0.4)	19	(0.7;0.6;0.6)
6	(8.1;0.0;0.4)	13	(0.7;0.4;0.6)	20	(8.1;0.6;0.6)
7	(0.7;0.0;0.6)	14	(8.1;0.4;0.6)		

TABLE IX: Exp. 3: Pairwise comparisons for sequences with only packet-loss, blockiness and blurriness. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(0.0;0.6;0.0)	-10.590*	2.006
	(0.0;0.0;0.6)	5.534	2.701
(0.0;0.6;0.0)	(0.0;0.0;0.6)	16.124*	2.203

TABLE X: Exp. 3: Pairwise comparisons for sequences with packet-loss and either blockiness or blurriness. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(8.1;0.0;0.0)	(8.1;0.0;0.4)	-8.180*	1.624
	(8.1;0.0;0.6)	-17.950*	1.838
	(8.1;0.4;0.0)	-21.199*	1.509
	(8.1;0.6;0.0)	-28.994*	1.653
(8.1;0.0;0.4)	(8.1;0.0;0.6)	-9.770*	1.659
	(8.1;0.4;0.0)	-13.019*	1.668
	(8.1;0.6;0.0)	-20.814*	1.620
(8.1;0.0;0.6)	(8.1;0.4;0.0)	-3.248	1.555
	(8.1;0.6;0.0)	-11.043*	1.488
(8.1;0.4;0.0)	(8.1;0.6;0.0)	-7.795*	1.418

MAVs differences between blockiness and the other two artifacts are significant ($\hat{F} = 24.906,\ \alpha \leq 0.01$). But, the difference between average MAVs between packet-loss and blurriness are not statistically significant.

Figure 4(b) shows a plot of the average MAV for test sequences with combinations of strong packet-loss artifacts (PDP=8.1%) and either blockiness (bloc=0.4 or 0.6) or blurriness (blur=0.4 or 0.6). Table X shows the RM-ANOVA test performed on the average MAVs of these sequences. Results of these pairwise comparisons show that there are significant statistical differences between the average MAVs obtained for

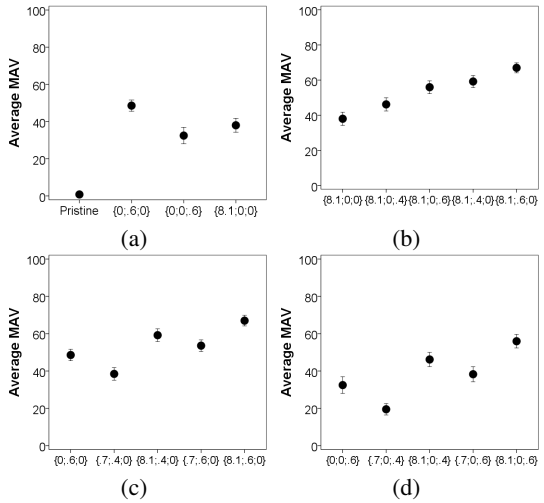

Figure 4: Exp 3: (a) Average MAVs for blockiness, blurriness and packet-loss, (b) MAVs for packet-loss by itself (PDP) and in combination with blurriness (+blur) and blockiness (+bloc), (c) MAVs for blockiness by itself (bloc) and in combination with packet-loss (+PDP), and (d) MAVs for blurriness by itself (blur) and in combination with packet-loss (+PDP).

combinations of packet-loss and either blockiness or blurriness artifacts ($\hat{F} = 99.542,\ \alpha \leq 0.01$). The only exception is the pair of combinations (8.1;0.0;0.6) and (8.1;0.4;0.0). In other other words, combining packet-loss with either blockiness and blurriness, on average, affects the MAV.

Figure 4(c) shows a plot of the average MAV for test sequences with blockiness and packet-loss artifacts. Table XI show the results of the RM-ANOVA tests performed on the average MAVs of these sequences. Notice that there are significant statistical differences for all pairs of combinations ($\hat{F} = 101.252,\ \alpha \leq 0.01$). Figure 4(d) shows a plot of the average MAV for test sequences with combinations packet-loss and blurriness. Table XII shows the results of the RM-ANOVA tests performed on the average MAVs of these sequences. Again, there are significant statistical differences for all pairs of these combinations ($\hat{F} = 93.310,\ \alpha \leq 0.01$). In general, combinations of packet-loss and blockiness have higher average MAVs than combinations of packet-loss and blurriness. Also, for combinations of packet-loss, blockiness, and blurriness, the presence of an additional artifact incurs in an increase of the average MAV's.

TABLE XI: Exp. 3: Pairwise comparisons for sequences with combinations of packet-loss and blockiness artifacts. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.6;0.0)	(0.7;0.4;0.0)	10.137*	1.554
	(8.1;0.4;0.0)	-10.609*	1.787
	(0.7;0.6;0.0)	-4.994*	1.343
	(8.1;0.6;0.0)	-18.404*	1.536
(0.7;0.4;0.0)	(8.1;0.4;0.0)	-20.745*	1.677
	(0.7;0.6;0.0)	-15.130*	1.439
	(8.1;0.6;0.0)	-28.540*	1.494
(8.1;0.4;0.0)	(0.7;0.6;0.0)	5.615*	1.506
	(8.1;0.6;0.0)	-7.795*	1.418
(0.7;0.6;0.0)	(8.1;0.6;0.0)	-13.410*	1.343

TABLE XII: Exp. 3: Pairwise comparisons for sequences with combinations of blurriness and packet-loss artifacts. (*. Significant at 0.05 level.)

Combinations		Diff. Mean	Std. Error
(0.0;0.0;0.6)	(0.7;0.0;0.4)	12.975*	2.310
	(8.1;0.0;0.4)	-13.714*	2.732
	(0.7;0.0;0.6)	-5.820*	1.749
	(8.1;0.0;0.6)	-23.484*	2.122
(0.7;0.0;0.4)	(8.1;0.0;0.4)	-26.689*	1.812
	(0.7;0.0;0.6)	-18.795*	1.983
	(8.1;0.0;0.6)	-36.460*	1.756
(8.1;0.0;0.4)	(0.7;0.0;0.6)	7.894*	2.177
	(8.1;0.0;0.6)	-9.770*	1.659
(0.7;0.0;0.6)	(8.1;0.0;0.6)	-17.665*	1.625

7. Comparison of Data from Experiments

Research shows that even results gathered from experiments using the same experimental methodology may differ considerably because of differences in physical location, viewer expectations, and especially set of stimuli [29]. It is known that subjects have a tendency to use the entire scoring scale to evaluate the quality of the test stimuli presented in an experimental session. As consequence, scores may suffer from

context effects [30]. For example, mildly impaired stimuli may get higher annoyance scores in an experiment containing only unimpaired or slightly impaired stimuli than in an experiment containing slightly to highly impaired stimuli.

In our experiments, we used different artifacts at different strengths. It is reasonable to assume that they may have spanned different ranges of MAVs that are not necessarily equivalent. In other words, the highest MAVs in the three experiments may correspond to videos impaired with artifacts of very different perceptual strengths.

For example, videos with the highest packet-loss strengths in Experiment 1 may have received the highest MAVs. But, the same MAVs in Experiment 3 may correspond to videos with much more annoying artifacts (and a lower quality), most likely presenting packet-loss in combination with blockiness and blurriness. Hence, before combining the MAVs into the same dataset, we must re-align them.

In fact, if we compare the MAVs obtained by sequences with strongest packet-loss configuration in isolation (i.e., (8.1;0;0)) in Experiment 1 and Experiment 3, we see a striking difference. In Experiment 1, this is the highest level of impairment encountered by subjects throughout the whole experiment. As such, it obtains a relatively high MAV (on average, across all contents, $MAV > 70$, see Fig. 2). On the other hand, the same videos impaired with the same combination in Experiment 3, are perceived only as mildly annoying (across all contents, $MAV \sim 40$, see Fig. 4 (a)). This is probably because, in comparison with videos that are distorted by multiple artifacts, heavy packet-loss is not as annoying. This discrepancy clearly points towards the presence of context effects in the MAVs of Experiment 1: MAVs are artificially inflated due to the relatively narrow range of quality spanned by the videos included in experiment. A re-alignment process is therefore necessary to map the MAVs of Experiment 1 to a range that is more commensurate to the annoyance values measured in Experiments 2 and 3.

Pinson et al. proposed a technique to merge data from different experiments known as the iterative nested least squares algorithm (INLSA) [29], [31]. INLSA re-scales subjective scores from different experiments using objective quality metrics as a common external variable. The procedure is performed solving two least squares problems. A single first-order correction method is used in the first problem to homogenize the heterogeneous scores of the different experiments. An approximation of the linear combinations of the parameters across the scores of the different experiments is obtained by solving the second problem. A full mapping of the scores of the different experiments into a common scale is obtained by performing an iteration of these two least-squares problems. To sample the mapping among scores of the different experiments, it is necessary to choose a common set of stimuli from all the experiments involved in the realignment.

Before comparing the data of the three experiments, we used INLSA to re-align the annoyance scores. We used SSIM [32] as the objective quality metric. Experiment 3 was used as the reference experiment because it had the highest number of artifact combination. Figures 5 show the MAV for the complete set of experiments before (top) and after (bottom) using INLSA, respectively, against the corresponding SSIM [32] value of the video. Notice that for the same SSIM values each experiment has different MAVs. In particular, and as expected, for Experiment 1, the entire MAV range is clustered on the top part of the SSIM scale. This means that videos with relatively low levels of impairments (as measured by SSIM) are judged as highly annoying (probably due to context effects, as mentioned above). This is not true for the other two experiments.

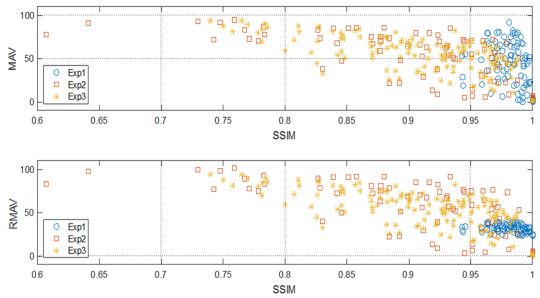

Figure 5: (top) MAVs and (bottom) RMAVs (after applying INLSA [31]) versus SSIM for Experiments 1-3.

After mapping the MAVs from Experiments 1 and 2 onto the scale of Experiment 3, the MAVs of Experiment 1 span a more comparable range of annoyance, also given the objective metric value. The range of the RMAVs of Experiment 1 ($Avg. = 32.17$, $Std = 5.52$) is smaller than the original range spanned by its MAVS ($Avg. = 42.78$, $Std. = 25.47$), and more skewed towards the lower part of the annoyance scale. In other terms, RMAVs now denote that annoyance values of videos impaired with only packet loss (as it is the case for Experiment 1) are lower as compared with those of sequences distorted by multiple artifacts. This result suggests that scores from Experiment 1 can be merged with those of the other two experiments, allowing to analyze the data from the three experiments as a whole.

A. Annoyance Models

In this work we study if models that combine the artifact strength values (PDP, bloc, and blur) can predict the perceived annoyance of videos impaired by multiple and overlapping artifacts. With this goal we fitted a set of linear and non-linear models. Prior to fitting the models, all artifact strength values were normalized to the same range $[0, 1]$, with the strength value '1' corresponding to the strongest artifact level found in practice and the value '0' corresponding to the absence of the artifact. Blockiness and blurriness were already generated using this scale, but the packet-loss strength values needed to be normalized. To re-scale PDP to fit this range, we assumed that PDP values greater than 10 would be unrealistic in practical network conditions and set 10 as the maximum PDP value [33], [34]. The normalized packet-loss strength is obtained by dividing the original values by 10, i.e. pdp = PDP$/10$.

1) Linear Models: The first linear model we tested was a simple linear model without any interaction term, given by:

$$PA_{L1} = \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur}, \qquad (2)$$

TABLE XIII: Fitting of the linear models to MAV and RMAV.

Models	δ	α	β	γ	PCC	SCC
PA_{L1}		50.060	72.480	48.620	0.726	0.721
PA_{L2}	23.515	29.350	50.606	26.740	0.730	0.727
PRA_{L1}		35.770	78.404	52.602	0.844	0.867
PRA_{L2}	18.170	19.768	61.499	35.698	0.850	0.870

TABLE XIV: Fitting of the linear model with interactions (PA_{L3}) to MAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
α	85.024	3.302	25.749	$< 2e-16^a$
β	88.550	4.344	20.386	$< 2e-16^a$
γ	64.118	4.344	14.761	$< 2e-16^a$
ρ_1	-123.393	13.301	-9.277	$< 2e-16^a$
ρ_2	-120.320	13.301	-9.046	$< 2e-16^a$
ρ_3	-22.561	14.724	-1.532	0.127
ρ_4	175.670	38.860	4.521	$< 8.87e{-}06^a$

[a] Statistically significant at ($P < 0.05$) PCC = 0.860, SCC = 0.841.

where PA_{L1} corresponds to the predicted (non-realigned) MAVs and pdp, bloc, and blur correspond to the strength of each artifact. Line 2 of Table XIII shows the results of the fitting. We also adapted eq. 2 to include an intercept coefficient (δ), referring to this model as PA_{L2}:

$$PA_{L2} = \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur} + \delta. \quad (3)$$

Line 3 of Table XIII shows the fit results for MAVs prior to the re-alignment with INLSA.

We tested the above models (Eqs. 2 and 3) on the MAVs re-aligned using INLSA, hereafter referred to as RMAVs. Line 4 of Table XIII shows the results for the first linear model (PRA_{L1}) fit, while line 5 shows the results for the second linear model (PRA_{L2}, with intercept term). To evaluate the goodness of the fit of each model, we report the Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) between predicted and subjective MAVs (or RMAVs) where the fit was based on the entire dataset. For both models, a better fit was obtained using RMAV instead of MAV.

2) Linear Models with Interactions: It has be shown that interaction terms must be taken into account when modeling the annoyance caused by combinations of artifacts because masking and facilitation processes may occur when artifacts are combined [15]. To investigate if the presence of one artifact may affect the perception of the other(s) and how this impacts the overall annoyance, we fitted a linear model with interactions, (PA_{L3}), defined as:

$$\begin{aligned} PA_{L3} = {} & \alpha \cdot \text{pdp} + \beta \cdot \text{bloc} + \gamma \cdot \text{blur} + \\ & \rho_1 \cdot \text{pdp} \cdot \text{bloc} + \rho_2 \cdot \text{pdp} \cdot \text{blur} + \\ & \rho_3 \cdot \text{bloc} \cdot \text{blur} + \rho_4 \cdot \text{pdp} \cdot \text{bloc} \cdot \text{blur}. \end{aligned} \quad (4)$$

Results of this fit for MAVs are shown in Table XIV. Column 2 of this table shows the values of the model coefficients, while column 5 shows the corresponding p-values (based on t-test, two-tailed, $p < 0.05$). Notice that the first, second, and third order coefficients are statistically significant, except for the ρ_3 coefficient corresponding to the interaction of blockiness and blurriness.

We also tested the same model with the addition of an intercept term δ, denoted as PA_{L4}. The results of this fit for MAVs are shown in Table XV. Again, the first, second, and third order terms have a statistically significant effect, except (again) for the ρ_3 coefficient that corresponds to the interaction of blockiness and blurriness. Linear models with interactions (bottom rows of Tables XIII and XV) have better correlation values than the linear models without interaction terms.

TABLE XV: Fitting of the linear model with interactions (PA_{L4}) to MAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
δ	14.117	2.078	6.792	$5.95e-11^a$
α	62.207	4.557	13.650	$< 2e-16^a$
β	65.050	5.327	12.211	$< 2e-16^a$
γ	40.619	5.327	7.625	$3.24e-13^a$
ρ_1	-84.372	13.670	-6.172	$2.18e-09^a$
ρ_2	-81.299	13.670	-5.947	$7.58e-09^a$
ρ_3	15.613	14.836	1.052	0.29348
ρ_4	109.970	37.507	2.932	0.00363^a

[a] Statistically significant at ($P < 0.05$) PCC = 0.853, SCC = 0.823.

TABLE XVI: Fitting of the linear model with interactions (PRA_{L3}) for RMAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
α	57.064	2.784	20.494	$< 2e-16^a$
β	88.685	3.663	24.212	$< 2e-16^a$
γ	61.703	3.663	16.846	$< 2e-16^a$
ρ_1	-69.785	11.217	-6.222	$< 1.65e-09^a$
ρ_2	-63.363	11.217	-5.649	$< 3.74e-08^a$
ρ_3	-10.196	12.416	-0.821	0.4122
ρ_4	55.827	32.768	1.704	0.0895

[a] Statistically significant at ($P < 0.05$) PCC = 0.880, SCC = 0.886.

TABLE XVII: Fitting of the linear model with interactions and with an intercept coefficient (PRA_{L4}) for RMAVs.

Coef.	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
δ	14.420	1.689	8.540	$6.83e-16^a$
α	33.757	3.702	9.118	$< 2e-16^a$
β	64.681	4.328	14.946	$< 2e-16^a$
γ	37.698	4.328	8.711	$< 2e-16^a$
ρ_1	-29.924	11.105	-2.695	0.00744^a
ρ_2	-23.503	11.105	-2.116	0.03514^a
ρ_3	28.800	12.053	2.390	0.01749^a
ρ_4	-11.286	30.470	-0.370	0.71134

[a] Statistically significant at ($P < 0.05$) PCC = 0.871, SCC = 0.886.

Fitting the two linear models with interaction terms with and without a fixed intercept on RMAVs, we obtained the predictions (PRA_{L3}) and (PRA_{L4}). Table XVI shows the results obtained for the model without the intercept coefficient (PRA_{L3}), while Table XVII shows the results obtained for the model with the intercept coefficient (PRA_{L4}). For both models, all main effects and first order interactions are statistically significant, except for ρ_3 (interaction of blockiness and blurriness) in PRA_{L3}. The second order interactions are not statistically significant for both models. Correlation coefficients are higher when RMAVs are used.

3) Non-Linear Models: The proposed linear models, although fairly accurate, may be unable to capture the complex non-linear interactions of the artifact combinations [35]. Therefore, we tested two different types of non-linear models: a Minkowski metric model and a model based on Support Vector Regression (SVR). We tested two Minkowski metrics, one without the intercept term (PA_{M1}) and another with the

TABLE XVIII: Fitting of Minkowski models on MAV and RMAV.

Models	m	δ	PCC	SCC
PA_{M1}	0.215		0.472	0.652
PA_{M2}	0.419	4.018	0.660	0.654
PRA_{M1}	0.215		0.562	0.770
PRA_{M2}	0.397	3.424	0.770	0.744

intercept term (PA_{M2}), as given by the following equations:

$$PA_{M1} = (\text{pdp}^m + \text{bloc}^m + \text{blu}^m)^{\frac{1}{m}}, \quad (5)$$

and

$$PA_{M2} = (\delta + \text{pdp}^m + \text{bloc}^m + \text{blu}^m)^{\frac{1}{m}}, \quad (6)$$

where PA_{M1} and PA_{M2} are the predicted annoyance, and m is the Minkowski power, obtained as a result of the fitting. It is worth pointing that this is the same combination rule used by Huib de Ridder to predict annoyance caused by blockiness, ringing [12] and by Farias et al. to predict annoyance caused by blockiness, blurriness, noisiness, and ringing [15]. De Ridder's model was tested on a smaller data set of still images and returned $m > 1.6$ values, whilst Farias's model was tested on interlaced SD videos and returned $m > 0.8$ values. Our results are different from the results obtained by both authors, what is expected since our stimuli consist of HD videos with both spatial and temporal artifacts.

Lines 2 and 3 of Table XVIII show the results of the fit on non-realigned MAVs of the model without intercept term (PA_{M1}) and the model with intercept (PA_{M2}), respectively. Lines 4 and 5 show the results of fitting on re-aligned MAVs of the model without intercept term (PRA_{M1}) and the model with intercept (PRA_{M2}), respectively. We can observe that these non-linear models perform worse than the linear ones. Within non-linear models, we observe again a better performance of those fit on RMAVs.

Finally, we used Support vector regression (SVR) to predict annoyance from the artifact strength data, using both MAVs and RMAVs. Machine learning-based approaches such as SVR have been shown to be suitable to model complex non-linear perceptual processes related to artifact annoyance [35]. In these approaches, the model is not previously defined but is learned from the data (i.e. our database of impaired videos). To train SVR, we used a k-fold cross validation setup. We split the dataset in k equally sized, non-overlapping sets. We then ran the training k times, for each of which a different fold was used as test set, and the remaining $k-1$ folds were used for training. In this way, each data point has a chance of being validated against the other [36]. In our experiments, we set k to 10, thereby running 10 repetitions of the training. We then computed the correlation between subjective data and model predictions per each run, and took their average as the SVR model performance measure. The SVR trained on RMAVs returned PCC and SCC values equal to 0.855 and 0.833, respectively, whereas the model trained on MAVs returned PCC and SCC values equal to 0.850 and 0.828, respectively.

4) Model comparison: The different models considered in the previous session achieved different degrees of accuracy, yet in some cases at the expenses of increased complexity. For example, models with interaction terms have more degrees of freedom (i.e., parameters to be fit) than models without; as a consequence, although more accurate, they may be more prone to overfitting. To compare the models in terms of the trade-off between complexity and accuracy, we use the Akaike Information Criterion (AIC) [37]. AIC expresses the trade-off between accuracy of fitting and the number of degrees of freedom in the model, thereby controlling for the bias/variance trade-off and overfitting. Table XIX below summarizes the AIC values computed for all models, where a model with lower AIC is preferred. Notice that although PRA_{L4} (the linear model with interaction and bias terms fit on re-aligned data) has more parameters, it has the lowest AIC, i.e. the best trade-off between goodness-of-fit and complexity.

TABLE XIX: Akaike Information Criterion for the linear and Minkowsky models. A lower value indicates a better trade-off between model complexity and accuracy.

Model	df	AIC	Model	df	AIC
PA_{L1}	4	2776.212	PRA_{L1}	4	2607.776
PA_{L2}	5	2638.636	PRA_{L2}	5	2464.547
PA_{L3}	8	2604.215	PRA_{L3}	8	2499.193
PA_{L4}	9	2562.162	PRA_{L4}	9	2434.164
PA_{M1}	2	3207.925	PRA_{M1}	2	3144.523
PA_{M2}	3	2693.669	PRA_{M2}	3	2608.433

TABLE XX: Average correlation across the 10-fold cross-validation runs between model predictions and (R)MAVs

Model	PCC	SCC	Model	PCC	SCC
PA_{L1}	0.706	0.713	PRA_{L1}	0.836	0.849
PA_{L2}	0.711	0.719	PRA_{L2}	0.844	0.851
PA_{L3}	0.775	0.747	PRA_{L3}	0.849	0.867
PA_{L4}	0.782	0.762	PRA_{L4}	0.861	0.858
PA_{M1}	0.463	0.628	PRA_{M1}	0.560	0.745
PA_{M2}	0.640	0.630	PRA_{M2}	0.736	0.745
PA_{SVM}	0.855	0.834	PRA_{SVM}	0.851	0.829

To verify whether the PRA_{L4} model also gives the best performance in terms of correlation, we perform again its fitting and that of all the other models in a 10-fold cross-validation setting, to obtain measurements comparable to those obtained for the SVR. The outcomes are reported in table XX. Notice that PRA_{L4} outperforms all models, including SVR.

B. Discussion

Models fit on RMAVs obtained a better performance, showing that re-aligning the data before fitting the models is beneficial. When an intercept constant was added to the models, the correlation coefficients increased. One possible cause for this result is that the original content may contain pre-existing artifacts, which subjects judged as slightly annoying.

For all linear models, the coefficients corresponding to bloc had the highest magnitude, indicating that blockiness had the biggest impact on the perceived annoyance. When fitting linear models on MAVs, pdp had a stronger impact on annoyance than blur, while when the fitting was done on RMAVs, blur had a higher impact. This divergence is caused by the fact that MAVs corresponding to sequences affected by packet-loss in Experiment 1 were overestimated (probably due to context effects). Therefore, when no re-alignment was performed, the

exaggerated MAVs of sequences with packet-loss caused this artifact to have a higher impact.

The majority of the second order coefficients were statistically significant. For the models fitted on MAVs, the exception is ρ_3, indicating that the specific combination of blockiness and blurriness does not influence the annoyance scores. In fact, for models fitted on MAVs, the majority of the interaction coefficients that include pdp were statistically significant. For models fit on RMAVs, the ρ_3 in the PRA_{L3} model (without intercept) was also not statistically significant. Most second order coefficients were negative, implying that the overall annoyance caused by the presence of two artifacts is not simply an addition of the respective annoyances. The co-presence of two artifacts might, in fact, reduce their combined overall annoyance. In other words, there may be masking effects among artifacts, with artifacts mutually attenuating each other's strength. Interaction coefficients with higher magnitudes were those corresponding to the pdp·bloc and pdp·blur terms. This suggests that packet-loss affects how blockiness and blurriness are perceived.

Third order interaction coefficients (ρ_4) were significant for MAVs and non-significant for RMAVs. Again, since in the non-realigned MAV set the contribution of the pdp parameter was overestimated, any interaction term containing pdp (ρ_1, ρ_2, and ρ_4) had a statistically significant impact on MAVs. This is not true for models fit on RMAVs, for which the specific strength combination of the three artifacts did not contribute to the overall annoyance.

Correlation coefficients obtained for Minkowski models were lower than what was obtained for the linear models. The Minkowski powers found ($0.215 < m < 0.420$) were considerably lower than the values found by other authors [12], [15]. This may indicate that these models were, in fact, more sensitive to small changes in artifact strengths. For these models we obtained similar correlation coefficients for the fits on MAVs and RMAVs. Finally, the SVR-based approach achieved correlations slightly lower than those achieved by the best linear model PRA_{L4}. We can conclude, then, that in this setting, linear models better for accurately modeling artifact annoyance.

8. Conclusions

We presented the results of three subjective experiments aimed at studying the characteristics of three artifacts (blockiness, blurriness and packet-loss) commonly found in digital videos, while investigating their interactions with each other. Results showed that annoyance increased with both the number of artifacts and their strength, with blockiness being the most annoying artifact. We proposed several models for predicting annoyance, including linear models with and without interactions and interception terms, Minkowski models, and a non-linear model based on SVR. Interactions were observed in the linear models, notably suggesting that the overlap of multiple artifacts generated masking effects, overall decreasing the annoyance perception. The correlation coefficients of fits using RMAVs were higher than for the fits using unscaled MAVs.

In this paper, we have used videos with synthetic artifacts in order to have a precise control of the parameters that affect the strength of each artifact. As future work, we intend to test the proposed models in videos with realistic distortions, i.e. generated using typical operations in a video communications pipeline (e.g. compression + transmission). We will also substitute the physical parameters used in our models (e.g., pdp, blo, and blu) with the corresponding artifact metrics. For that, we need to perform new psychophysical experiments to obtain perceptual strength measures for each artifact. With these perceptual measures, we can map the output of the artifact metrics into the physical parameters used in our models.

References

[1] "Itu-t recommendation bt.500-8: Methodology for the subjective assessment of the quality of television pictures," Internation Telecommunication Union, Tech. Rep., 1998.

[2] P. Le Callet, S. Möller, A. Perkis *et al.*, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012.

[3] J. A. Redi, "Visual quality beyond artifact visibility," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 86 510N–86 510N.

[4] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *Multimedia, IEEE Transactions on*, vol. 8, no. 5, pp. 973–980, Oct 2006.

[5] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[6] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: what does the future hold?" *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 675–696, 2011.

[7] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing attended and global stimuli in perceived video quality assessment," *Multimedia, IEEE Transactions on*, vol. 13, no. 6, pp. 1269–1285, 2011.

[8] R. V. Babu, A. S. Bopardikar, A. Perkis, and O. I. Hillestad, "No-reference metrics for video streaming applications," in *International Workshop on Packet Video*, 2004.

[9] 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*, vol. 3. IEEE, 2005, pp. III–141.

[10] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 935–949, Oct 2011.

[11] V. Kayarrgadde and J. Martens, "Perceptual characterization of images degraded by blur and noise: Model," *Journal of the Optical Society of America, A-Optics & Image Science*, vol. 13, no. 6, pp. 1178–1188, 1996.

[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*. International Society for Optics and Photonics, 1992, pp. 16–26.

[13] D. Chandler, K. Lim, and S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," in *Proc. Human Vision and Electronic Imaging*, January 2006.

[14] M. C. Farias, J. M. Foley, and S. K. Mitra, "Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts," *Signal Processing, IEEE Transactions on*, vol. 55, no. 6, pp. 2954–2964, 2007.

[15] M. C. Farias and S. K. Mitra, "Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance," *Journal of Electronic Imaging*, vol. 21, no. 4, pp. 043 013–043 013, 2012.

[16] M. S. Moore, J. M. Foley, and S. K. Mitra, "Defect visibility and content importance: effects on perceived impairment," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 185–203, 2004.

[17] Q. Huynh-Thu and M. Ghanbari, "Modelling of spatio–temporal interaction for video quality assessment," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 535–546, 2010.

[18] A. Reibman, V. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *Multimedia, IEEE Transactions on*, vol. 6, no. 2, pp. 327–334, April 2004.

[19] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1316–1324, Nov 2008.

[20] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for h.264/avc coded video," *Multimedia, IEEE Transactions on*, vol. 11, no. 5, pp. 932–946, Aug 2009.
[21] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, "Perceptual effects of packet loss on h. 264/avc encoded videos," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*, 2009.
[22] J. Redi, I. Heynderickx, B. Macchiavello, and M. Farias, "On the impact of packet-loss impairments on visual attention mechanisms," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1107–1110.
[23] "Varium project video database," http://www.ene.unb.br/mylene/databases.htm, accessed: 2016-01-30.
[24] VQEG, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment," *Video Quality Experts Group and others, Phase I*, 2008.
[25] G. Cermak, S. Wolf, E. Tweedy, M. Pinson, and A. Webster, "Validating objective measures of mpeg video quality," *SMPTE journal*, vol. 107, no. 4, pp. 226–235, 1998.
[26] J. M. Libert, C. P. Fenimore, and P. Roitman, "Simulation of graded video impairment by weighted summation: validation of the methodology," in *Photonics East'99*. International Society for Optics and Photonics, 1999, pp. 254–265.
[27] M. Farias, "No-reference and reduced reference video quality metrics: New contributions," Ph.D. dissertation, University of California, Santa Barbara, California, 2004.
[28] "Itu-t recommendation p.930: Principles of a reference impairment system for video," 1996.
[29] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Visual Communications and Image Processing 2003*. International Society for Optics and Photonics, 2003, pp. 583–592.
[30] H. de Ridder, "Cognitive issues in image quality measurement," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 47–55, 2001.
[31] S. D. Voran, "Iterated nested least-squares algorithm for fitting multiple data sets," *NASA STI/Recon Technical Report N*, vol. 3, p. 12919, 2002.
[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.
[33] A. Papadogiannakis, A. Kapravelos, M. Polychronakis, E. P. Markatos, and A. Ciuffoletti, "Passive end-to-end packet loss estimation for grid traffic monitoring."
[34] A. Vakili and J.-C. Grégoire, "Estimation of packet loss probability from traffic parameters for multimedia over ip," in *Proc. of the Seventh International Conference on Networking and Services, ICNS*, 2011, pp. 44–48.
[35] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–15, 2013.
[36] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.
[37] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.

Alexandre F. Silva received his BSc in computer science from Centro Centro Universitário de Votuporanga (CEUV), Brazil, in 2002 and his MSc in computer science from Universidade Federal de Uberlândia (UFU), Brazil, in 2006. Actually, he is a PhD student in computer science from the Universidade Estadual de Campinas (UNICAMP). His current interests include video quality metrics, image/video processing, multimedia, face recognition, academic softwares, and web programming. He is a reviewer of the IEEE Latin America Transaction, Brazilian Journal of Information Security and Cryptography (ENIGMA), and Journal of the Brazilian Computer Society (JBCS).

Mylène C.Q. Farias (M'02) received her B.Sc. degree in electrical engineering from Universidade Federal de Pernambuco (UFPE), Brazil, in 1995 and her M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas (UNICAMP), Brazil, in 1998. She received her Ph.D. in electrical and computer engineering from the University of California Santa Barbara, USA, in 2004 for work in no-reference video quality metrics. Dr. Farias 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, USA) developing no-reference video quality metrics. She is currently an assistant professor at the Department of Electrical Engineering of the University of Brasília (UnB), Brazil. Her current interests include video quality metrics, video processing, multimedia, watermarking, and machine learning. Dr. Farias is a member of IEEE.

Judith A. Redi received her PhD from the University of Genoa (Italy) in 2010, with a thesis on learning machines for objective image quality assessment, final result of a project on visual quality in displays funded by Philips research. After receiving the award for the best ICT thesis from University of Genoa, she moved to Eurecom (France) for a post doc on Digital Image Forensics and 3D face recognition. Since October 2010, she is an Assistant Professor at the Multimedia Computing group of Delft University of Technology, faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS). At TU Delft, she works on image and video understanding towards the maximization of the Quality of (multimedia) Experiences. This is a highly multidisciplinary field that connects visual perception, cognitive science and engineering, machine learning and computer vision. She is also ambassador of DEWIS (Delft Women in Science) in EEMCS, and as such she promotes activities to support gender diversity and women networking within the faculty.