DETECTABILITY, ANNOYANCE, AND STRENGTH OF PACKET LOSS IMPAIMENTS

Mylène C.Q. Farias, Judith A. Redi, Ingrid Heynderickx

Address - Line 1
Address - Line 2
Address - Line 3

ABSTRACT

We present some results on studying the visibility and annoyance of packet loss artifacts in isolation of other digital artifacts.

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, 2, 3]. Being able to detect artifacts and improve the quality of the visual content prior to its delivery to the user is therefore crucial to ensure a good quality of experience. At the basis of such a quality control mechanism, is an (automated) visual quality assessment system.

The most accurate way to determine the quality of a video is by using psychophysical experiments with human subjects [3]. Unfortunately, these are very expensive, time-consuming and hard to incorporate into a design process or an automatic quality of service control. Therefore, there is a great need for objective quality metrics, i.e., algorithms that can predict visual quality as perceived by human observers. Quality metrics that analyze visible differences between a test and a reference signal, taking into account aspects of the human visual system (HVS), usually have the best performance [4, 5], but are often computationally expensive and therefore hardly applicable in real-time contexts. Alternatives to these metrics are artifact metrics, which estimate the strength of individual artifacts and, then, combine the artifact strengths to obtain an overall annoyance or quality model. The assumption here is that, instead of trying to detect and estimate the strength of an 'unknown' impairment that consists of a combination of artifacts, it is easier to detect individual artifact signals and estimate their strength because we 'know' their appearance and the type of process which generates them [6, 7, 8, 9].

Artifact metrics have the advantage of being simple and not necessarily requiring the reference signal. Also, they can be useful for post-processing algorithms, providing information about which artifacts need to be mitigated. One disadvantage of artifact metrics is that their design requires a good understanding of the perceptual characteristics of each artifact. A second disadvantage is that the artifact metrics need to be combined to an overall quality estimate [8, 9]. Therefore, in order to design good artifact metrics it is important to find a model that describes how the individual artifact measurements (signal strengths) can be combined to provide the overall annoyance or quality. Unfortunately, little work has been done on studying and characterizing the individual artifacts [10, 11, 12, 13], as pointed out by Moorthy and Bovik in a recent paper [14]. We believe that an extensive study of the most relevant artifacts is still necessary, since until today we still do not have a good understanding of how artifacts depend on the physical properties of the video and how they combine to produce the overall annoyance.

In a previous work, the visibility, annoyance, and interaction of blockiness, ringing, noisiness, and blurriness and their relation with spatial content was studied [15, 16]. In this paper, we extend this work to include temporal artifacts. In particular, we study the transmission artifact "packet loss". With the goal of understanding the main psychophysical characteristics of the "packet loss" artifact, we generate "packet loss" artifacts with diffent strengths, durations, and spatial and time distributions in the videos. Then, we perform two psychophysical experiments that measure the visibility, annoyance, and strength of these artifacts. We also compare our results to studies that have evaluated packet loss artifacts in highly compressed videos (i.e., containing a mixture of digital artifacts).

This paper is divided as follows. In Section 2 we describe the algorithms used to generate the "packet loss" artifacts. In Section 3, we describe the psychophysical experiment methodology used in both experiments, which includes the type of

equipment used in the test, the tasks, and choice of the test sequences. ADD SECTION OF DISCUSSION. Finally, in Section 5 we give the conclusions.

2. GENERATION OF PACKET LOSS ARTIFACTS

In the video transmission over IP networks, the network variability and the lack of service guarantees represent a big challenge. Transmission errors may occur due to network congestion and path loss. Typical impairments caused by these errors are "packet loss", jitter, and delays. Typically, for block-based video compression schemes (e.g. MPEG-1/2/4, H-261/2/3/4), consecutive macroblocks in a frame are transmitted as a slice in a single network packet. Therefore, the loss of network packets results in a loss of macroblocks. Because the compression process removes a lot of spatial and temporal redundancies from the original video, and because of the use of motion-compensated temporal prediction, a single loss of a packet can affect many subsequent frames. As the name suggests, "packet loss" artifacts are caused by a complete loss of the packet being transmitted.

"Packet loss" artifacts are visually characterized by the presence of rectangular areas distributed over the video frames, whose content differs from the surrounding areas. Their visibility and annoyance depend heavily on how the video stream has been coded, how it has been mapped into flows and packetized, and what type of error concealment algorithm is being used. In the literature, there is a considerable amount of work on the visibility of the packet-loss, as summarized by Boulos *et al* [17]. Some literature also investigated the effect of scene characteristics on the visibility of packet loss [18]. Some studies have attempted to address the visibility and annoyance of packet loss artifact [19, 20].

In most studies, packet loss artifacts are generated by varying parameters of compression algorithms (codec type, bitrate, etc.) and digital transmissions (loss rate, channel model, etc.). As a consequence, the generated videos contain compression artifacts (e.g., blockiness, blurriness, and ringing) besides the packet loss artifacts. **In [19], the authors show that the annoyance of packet loss artifacts is correlated with their length (propagation throughout frames) and with the severity of the losses (PSNR), whereas their visibility seems not to be related to the length of the loss itself, but rather to the overall degradation of the video. However, these results are based on the subjective evaluation of degraded versions of a single video content and both visibility and annoyance are not analyzed in relation to the spatial and temporal characteristics of the video. Furthermore, a loss is considered visible if it generates a drop in visual quality; whereas it might be argued that a loss might be visible and yet not generate annoyance (and quality loss as a consequence). The study in [20] relates the visibility of packet loss artifacts to the percentages of slices lost, the type of frames where the loss happened (I, B or P), the duration of the loss and the amount of motion in the video. Unfortunately, the analysis is not extended to the annoyance of visible artifacts.**

The overall goal of this reasearch is to study and characterize the most relevant artifacts present in digitally compressed and transmitted videos. More specifically, we want to obtain a good understanding of the characteristics of various individual artifacts, their mutual interference, and their interference with the content of the image material. With this goal, we use the specifications for an adjustable video reference system detailed in ITU-T Recommendation P.930 [21]. This system generates synthetic artifacts that look like "real" artifacts, yet are simpler, purer, and easier to describe. Therefore, it offers advantages for experimental research on video quality because it makes it possible to control the amplitude, distribution, and mixture of different types of artifacts making it possible to study the different types of artifacts. According to the ITU-T Recommendation P.930, the created artifacts must be relatively pure, easily adjusted, and must be combined to match the appearance real impairments. Also, the algorithms for generating them must be well defined in a way that the artifacts can be easily reproduced. Previous studies on the matter show that the the artifacts must produce *psychometric* and *annoyance* functions that are similar to those of real artifacts [12].

To generate test sequences with several levels of "packet loss" artifacts with different strengths, durations, and spatial and time positions, we used the reference H.264 codec. To avoid inserting additional artifacts (such as ringing, blurring, and blockiness) we compressed the original videos with high compression rates, generating videos with Peak Signal to Noise Ratio (PSNR) well above 70dB. To vary the strength of the artifacts at different spatial and time distributions, we randomly deleted packets from the coded video bitstream. The percentage of deleted packets varied from 0.5% to 9. To vary the time interval of the introduced artifacts, we varied the frame interval between the I-frames. Three frame intervals were used: 4, 8, and 12.

Seven high definition videos with spatial resolution 1920x720 (50fps) were used in the experiment. The videos were all eight seconds long and were chosen with the goal of generating a diverse content, as can be seen in Fig. 1. To have an idea of the spatial and temporal content of the videos, in Fig. 2 we show a graph of the spatial and temporal measures of the originals [22]. In summary, we had the following parameters for the experiment:

- 7 originals;
- 3 time durations (Frame intervals: M = 4, 8 or 12);

- 4 artifact strengths (percentage of deleted packets varied from 0.5% to 9%)

At total, we had for each original $4 \times 3 = 12$ parameters, generating $13 \times 7 = 91$ test sequences.

(a) (b) (c) (d)

(e) (f) (g)

Fig. 1. Frames of videos: (a) Joy Park, (b) Into Trees, (c) Crowd Run, (d) Romeo & Juliet, (e) Cactus, (f) Basketball, and (g) Barbecue.

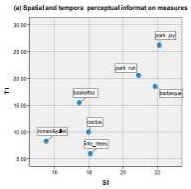


Fig. 2. Temporal and spatial characteristics of the videos included in the experiment. [GENERATE A NEW FIGURE!!!]

3. PSYCHOPHYSICAL EXPERIMENT

In this section we describe the physical conditions, the experimental methodology, and the statistical analysis methods used for Experiments I and II.

3.1. Apparatus and Physical Conditions

The experiments were performed using a PC computer with a Samsung LCD monitor of 23 inches (Sync Master XL2370HD). The dynamic contrast of the monitor was turned off and the contrast was set at 100 and the brightness at 50. The measured gamma of the monitor was approximately 0.99, 0.97, 1.00, and 0.92 for luminance, red, green, and blue, respectively. The room had the lights dimmed to avoid it to be reflected on the monitor.

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 is 3 video monitor screen heights. Three screen heights is a conservative estimate of the viewing distance according to the ITU-T Recommendation BT.500 [23]. We used a chin rest to guarantee that the distance between the subject's eye and the monitor remained the same.

Our subjects were voluntaries from the Department of Mediametics in the Delft Univeristy, The Netherlands. Most subjects were graduate students of the department. They were considered naïve of most kinds of digital video defects and the associated terminology. No vision test was performed on the subjects, but they were asked to wear glasses or contact lenses if they need them to watch TV. In each experiment, at least 15 subjects were used to guarantee robust results [24]. The software Presentation from Neurobehavioral Systems Inc. was used to run the experiment and record the subjects' data.

3.2. Experimental Methodology

A experimental (test) session was broken into the following five stages: (1) Instructions, (2) Training, (3) Practice Trials, (4) Experimental Trials, and (5) Interview. Before starting the experiment, the experimenter needed to make sure the subject was properly seated at the adequate distance. Subjects were then explained the tasks to be performed in the experimental trials. They were told to disregard the content of the videos and judge only the impairments they see.

In each experiment the subject was asked to perform a task which consists of entering a judgement about an impairments seen in the video. In order to complete this task subjects needed to have an idea of how the artifacts looked like and how videos with no impairments (originals) compared with videos with strong artifacts. With this goal, we included a **training** session in the experimental session which consisted of displaying the original videos followed by examples of videos with the strongest impairments found in the experiment. The subjects were instructed to watch these videos carefully and assign a maximum value of 100 to the worst or strongest impairments in this subset.

The initial judgements of a test subject are generally erratic. It takes time for a subject to get used to the task of judging/detecting impairments. The ITU Recommendation suggests that the first five to ten trials to be thrown away [23]. In our methodology, instead of discarding the first trials, we included **practice trials**. Before beginning this stage, subjects were told that this is a practice stage and that no data is being recorded. Besides eliminating erratic answers, the practice trials had other benefits. It exposed the subjects to sequences with a good range of impairments and gave subjects a chance to try out the data entry procedure. They also allowed subjects to gain confidence in their judgements. In this work, at least five practice trials were used.

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

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

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

The strength task consisted of asking the subjects for an estimate of how strong or visible a set of artifacts were in the detected impairment. As mentioned earlier, this type of task required that subjects be taught how each artifact looked like. Therefore, in the training stage subjects were shown a set of sequences illustrating the set of artifacts being measured. In the trials, after the video was played, the subject was asked to enter a number in a scale with range from ‘0’ to ‘100’ corresponding to the strength of that artifact or feature. If no impairments were seen, subjects were instructed not to enter any number and just click ‘Next’ to go on to the next trial.

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

3.3. Statistical Analysis Methods

The logarithm of the total squared error (TSE) is used as the objective metric in the statistical analysis of the psychophysical experiment. The TSE is given by:

$$\mathrm{TSE} = \sum_k \sum_i \sum_j \left(Y(i,j,k) - X_o(i,j,k)\right)^2, \tag{1}$$

where Y is the impaired video and X_o is the original video, i and j are the spatial coordinates, and k corresponds to the frame number.

To analyze the subjects' answers to detection tasks, we first convert the 'yes/no' answers to binary scores. The 'yes' is saved as '1', while 'no' is saved as '0'. Probability of Detection (P_{det}) of an impairment is estimated by counting the number of subjects who detect this impairment and dividing by the total number of subjects. Using the probability of detection data, we can estimate the *visibility detection threshold* of impairments. The probability as a function of the log(TSE) (*psychometric function*) is fitted using the Weibull function, which has an S-shape similar to the experimental data and is defined as:

$$P_{det}(x) = 1 - 2^{(S_T \cdot x)^\kappa}, \tag{2}$$

where $P_{det}(x)$ is the probability of detection, x is the $\log(TSE)$ of the sequence, S_T is the sensitivity, and κ is a constant that determines the steepness of the function. The 50% detection threshold in logarithmic error energy, x_T, is given by $1/S_T$.

For annoyance and strength tasks, the judgements given by the subjects to each test sequence are called subjective scores. This data is first processed by calculating the *mean observer score* (MOS) by averaging the scores over all observers for each test sequence:

$$\mathrm{MOS} = \overline{S} = \frac{1}{L} \cdot \sum_{i=0}^{L} S(i), \tag{3}$$

where $S(i)$ is the score reported by the *i*-th subject, and *L* is the total number of subjects. Depending on the task, the MOS will represent different subjective magnitudes and will be named accordingly. For annoyance tasks, the MOS is called MAV (Mean Annoyance Value) since the subjective scores in this case correspond to 'annoyance' scores. For strength tasks, the MOS is called MSV (Mean Strength Value) since, in this case, they correspond to 'strength' scores.

We also calculate the sample standard deviation of the scores:

$$STD = \left(\frac{1}{L} \cdot \sum_{i=0}^{L} \left(S(i) - \overline{S}\right)^2 \right)^{1/2}, \tag{4}$$

and the internal standard error of $\overline{S}$:

$$\overline{STD} = \frac{STD}{\sqrt{L}}. \tag{5}$$

This is under the assumption that the scores are independent. The confidence interval for the 'true' MOS of a test sequence is given by $\overline{S} \pm t_{L,\alpha/2}\, \overline{STD}$ where $t_{L,\alpha/2}$ corresponds to the Student t coefficient.

With the MAVs data, we can estimate the mid-annoyance values. The MAV, as a function of the log(TSE) (*annoyance function*), is fitted with the standard logistic function:

$$y = y_{min} + \frac{(y_{max} - y_{min})}{1 + exp\left(-\frac{(x - \overline{x})}{\eta}\right)}, \tag{6}$$

where y is the predicted annoyance and $x = \log(TSE)$. The parameters y_{max} and y_{min} establish the limits of the annoyance value range. The parameter $\overline{x}$ (mid-annoyance value) translates the curve in the x-direction and the parameter η controls the steepness of the curve.

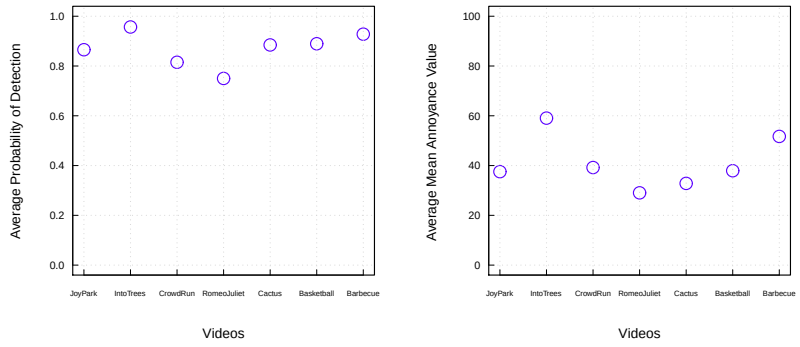
4. DATA ANALYSIS

4.1. Experiment I: Visibility and Annoyance

A total of 16 subjects performed the annoyance and detection task in Experiment I. In Fig. 3, the average values for the probability of detection (P_{det}) and mean annoyance values (MAV) are depicted for every group of originals. From Fig. 3.(a),

we can notice that the average values of P_{det} were generaly high, indicating that subjects were able to detect most artifacts in the videos. It is also possible to see that, for the videos 'Park Joy', 'Crowd Run', and 'Romeo & Juliet', the average values of P_{det} are slightly smaller than for the other videos. From Fig. 3.(b), notice that the average values of MAV have a larger variation than than the average P_{det}. Notice also that the videos 'Into Trees' and 'Barbecue' got the highest scores, while the videos 'Romeo & Juliet' and 'Cactus' got the lowest scores.

To take a closer look, in Figure 4 we depict the P_{det} values for each of the videos. The x axis in the graphs corresponds to the Mean Squared Error (MSE) corresponding to the four 4 artifact strengths, while the y axis corresponds to the probability of detection (P_{det}). The different curves in the graphs correspond to the 3 different frame intervals between the I frames (M = 4, 8, or 12). Each curve, therefore, corresponds to a *test group* which consists of 4 artifact strengths, 1 frame interval, 1 original.



(a) (b)

Fig. 3. Experiment I: Average values for the probability of detection (P_{det}) and mean annoyance values (MAV).

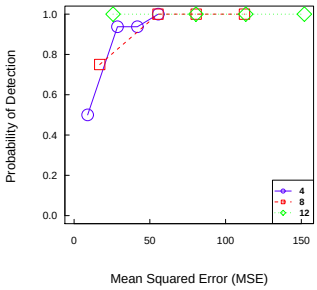
Notice from the graphs in Figures 4 (b) and (c) (videos 'Into Tree' and 'Barbecue') that P_{det} has all values equal to one, which means that all subjects reported seeing impairments in all test cases. It is worth pointing out that these two scenes, which showed the highest average P_{det}, present large smooth regions (sky) that make impairments easier to detect. Videos like 'Park Joy', 'Cactus', and 'Basketball' (Figures 4 (a), (e), and (f)) have values of P_{det} that grow (and saturate) very fast as the MSE increases. For the videos 'Romeo and Juliet' and 'Crowd Run' (Figures 4 (c) and (d)), on the other hand, have curves of P_{det} that increase at a slower rate. This indicates that for these scenes it is harder to detect the artifacts. The video 'Romeo and Juliet', although it is a video with small spatial and temporal activity, it is relatively dark and has a very clear central of attention (the couple in the middle of the scene). All of this, makes it harder to spot the artifacts. In the case of 'Crowd Run', this is a video with lots of spatial (crowd) and temporal activity and not a lot of camera movement. Therefore, it is again not easy to spot packet loss impairments.

One of the goals of this experiment was to understand how the content (originals), the artifacts (packet loss), and their characteristics affect the probability of detection (P_{det}). Unfortunately, many of the test groups had very high P_{det} and we were only able to calculate the 50% detection threshold for 5 (out of 21) test groups: 'Joy Park' with M = 4, 'Crowd Run' with M = 4, 'Crowd Run' with M = 8, 'Romeo & Juliet with M = 4, and 'Romeo & Juliet with M = 8. Because most of the P_{det} values were equal to '1', it is hard to fit a valid model for P_{det}.

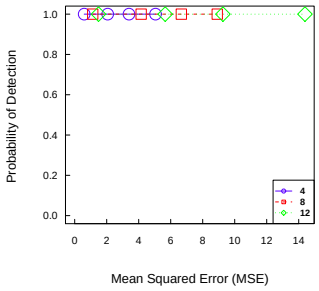
To get around this problem, we selected only the originals for which we were able to detect the detection thresholds ('Joy Park','Crowd Run', and 'Romeo & Juliet). The following factors were tested: mean squared error (MSE), frame interval (M), spatial and the temporal activity (SI and TI). For these videos, we fitted the following linear model with interactions:

$$\begin{aligned}\hat{P}_{det} =& a_1 \cdot ST + a_2 \cdot SI + a_3 \cdot M + a_4 \cdot MSE + \\ & a_{12} \cdot SI \cdot ST + a_{13} \cdot SI \cdot M + a_{14} \cdot SI \cdot MSE + \\ & a_{23} \cdot ST \cdot M + a_{24} \cdot ST \cdot MSE + a_{34} \cdot M \cdot MSE + \\ & a_{123} \cdot SI \cdot ST \cdot M + a_{234} \cdot ST \cdot M \cdot MSE.\end{aligned}$$

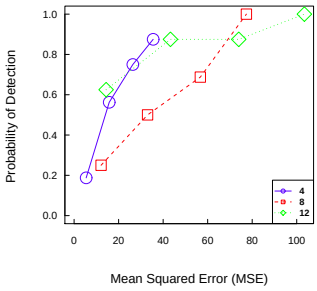
This model allowed us to analyze the main effects (M, SI, ST, and MSE) and the interactions among these factors. The results of the significance tests are shown in Table 1. The factors M (duration of the artifact), SI (spatial activity), and MSE (error) had statistically significant effects ($P < 0.05$) on the probability of detection of the originals tested. Also, the interactions between ST:MSE and M:MSE were also found to be statistically significant. These results are in agreement with our observations. For



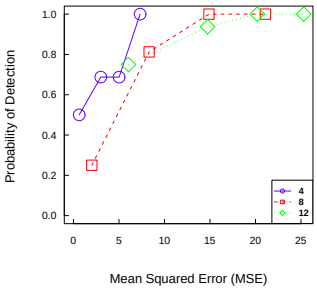
(a) ‘Joy Park’



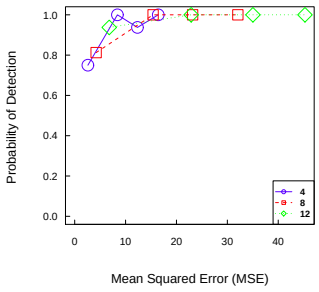
(b) ‘Into Tree’



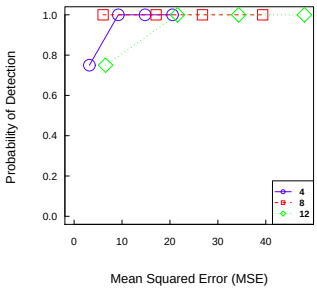
(c) ‘Crowd Run’



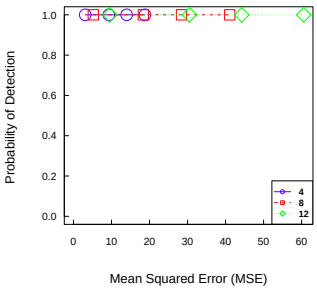
(d) ‘Romeo & Juliet’



(e) ‘Cactus’



(f) ‘Basketball’



(g) ‘Barbecue’

Fig. 4. Experiment I: Probability of Detection (P_{det}) for all videos.

example, the highest the M value the easiest it is to detect the artifact. The same, of course, is true for the MSE. Concerning the spatial activity, the higher the spatial

In Figure 5, the graphs of MAV for each original is shown. The higher the packet loss ratio (MSE), the higher the MAV. The graphs show three curves, each one corresponding to a different frame interval (M). As expected, the bigger M (4, 8 or 12), the higher the MAV. But, alhtough for M = 12 the MAV is always higher, the reverse is not always true, i.e. for M = 4 the MAV are not always smaller.

For some of the videos (‘Crowd Run’ and ‘Romeo and Juliet) the MAV curves for M = 4 and 8 are very similar (see Fig. 7), i.e. subjects did not notice a difference in quality between these two types of artifacts with different time intervals. Notice that these two videos are the same videos for which subjects had difficulty in detecting the artifacts (see Fig. 5). Notice also that, the video ‘Barbecue’, which had probability of detection equal to ‘1’, had annoyance scores higher than the annoyance scores given to other videos (compare Fig.8 with Figs.6-7), This may indicate that there must be a correlation between visibility and annoyance.

Figures 8-9 show the annoyance functions for all test groups. Unfortunately, the fits of the functions were not always good. Columns 2 and 3 of Table 2 show the annoyance function fitting parameters ($\bar{x}$ and η).

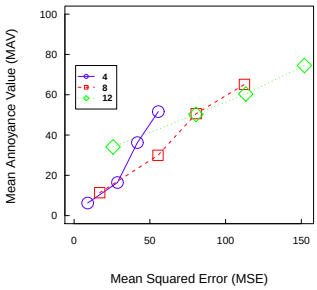
We also performed an ANOVA test to analyze the MAV model and estimate the main effects and their interactions. The factors tested were: MSE, M, SI and TI. The results of the ANOVA are shown on Table 3, where the statistically significant

Table 1. Experiment I: Table of the fitting of the linear model with interactions used for testing the effects of MSE, SI, ST, M on P_{det}. Statistically significant terms ($P < 0.05$) are in bold.

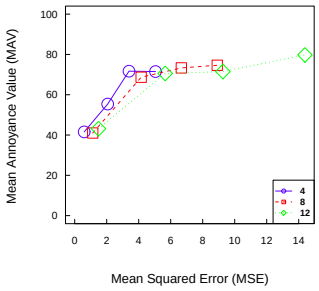
	Df	Sum Sq	Mean Sq	F value	$P < 0.05$
ST	1	0.0306	0.030590	1.320400	0.260605
SI	1	0.2875	0.287520	12.410400	**0.001540**
M	1	1.0545	1.054470	45.514800	**3.038e-07**
MSE	1	0.3283	0.328340	14.172300	**0.000822**
ST:M	1	0.0495	0.049470	2.135500	0.155463
SI:M	1	0.0000	0.000000	0.000000	0.994683
ST:MSE	1	0.3543	0.354330	15.294100	**0.000560**
SI:MSE	1	0.0081	0.008130	0.350800	0.558563
M:MSE	1	0.4406	0.440590	19.017600	**0.000169**
ST:M:MSE	1	0.0214	0.021430	0.924900	0.344724
SI:M:MSE	1	0.0570	0.057030	2.461500	0.128316
Residuals	27	0.6255	0.023170		

Table 2. Fitting parameters for annoyance functions.

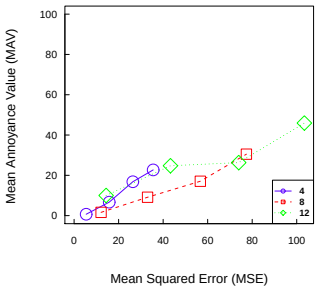
Group	Xmean	Beta	Residuals
1	6.74	0.18	0.67
2	6.91	0.26	0.69
3	6.79	0.48	0.76
4	5.63	0.64	0.43
5	5.94	0.46	0.46
6	6.03	0.44	0.43
7	6.91	0.22	0.21
8	7.18	0.24	0.18
9	7.28	0.42	0.76
10	6.45	0.32	0.77
11	6.54	0.16	0.72
12	6.48	0.20	0.33
13	6.44	0.44	0.38
14	6.67	0.36	0.52
15	6.68	0.43	0.30
16	6.59	0.51	0.94
17	6.45	0.44	0.39
18	6.52	0.29	0.14
19	5.94	0.76	0.40
20	6.37	0.31	0.15
21	6.32	0.38	0.52



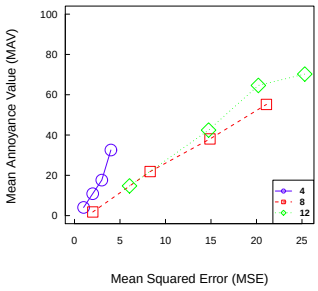
(a) 'Joy Park'



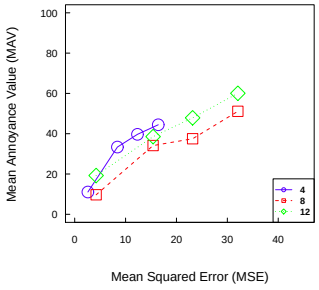
(b) 'Into Tree'



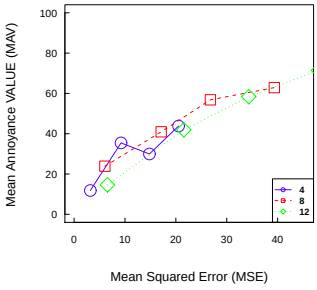
(c) 'Crowd Run'



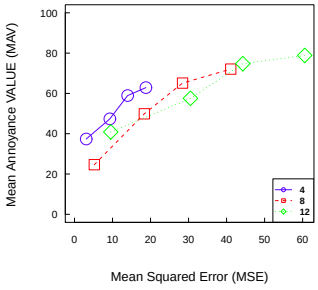
(d) 'Romeo & Juliet'



(e) 'Cactus'



(f) 'Basketball'



(g) 'Barbecue'

Fig. 5. Experiment I: MAV for all videos.

effects are shown in bold ($P < 0.05$). It can be noticed that all single factors have a significant effect in determining MAV. The model only found the following statistically significant interactions: SI*MSE and ST*SI*MSE.

Besides the most obvious factors like error energy (MSE) and artifact time interval (M), we found that spatial activity (SI) has a major impact on determining MAV. Although temporal activity (ST) has also a statistically significant effect on MAV, the overall effect seems to be smaller. Further studies are needed to determine what other factors (luminance, contrast, attention) affect quality.

4.2. Experiment II

4.2.1. Strength versus Error

4.2.2. Strength versus Annoyance

5. SUMMARY AND CONCLUSIONS

We presented the description, statistical analysis, and conclusions of two psychophysical experiments. The goals of these experiments were to study the appearance, visibility, and annoyance of four artifacts (blockiness, blurriness, ringing, and

Table 3. Anova Table for testing the effects of MSE, SI, ST, M, and all their interactions on MAV. Statistically significant terms (P¡0.05) are in bold.

	Df	SumSq	MeanSq	F value	Pr (>F)
ST	1	1576.4	1576.4	5.917	**0.0174**
SI	1	4511.1	4511.1	16.933	**9.83E-005**
M	1	11316.6	11316.6	42.478	**7.34E-009**
MSE	1	7371.2	7371.2	27.669	**1.31E-006**
ST:SI	1	175.1	175.1	0.657	0.42
ST:M	1	1146.7	1146.7	4.304	0.0415
SI:M	1	80.3	80.3	0.302	0.5845
ST:MSE	1	417	417	1.565	0.2148
SI:MSE	1	1626	1626	6.103	**0.0158**
M:MSE	1	684.9	684.9	2.571	0.1131
ST:SI:M	1	286.4	286.4	1.075	0.3031
ST:SI:MSE	1	1608.2	1608.2	6.036	**0.0163**
ST:M:MSE	1	212.5	212.5	0.798	0.3747
SI:M:MSE	1	6.5	6.5	0.024	0.8767
ST:SI:M:MSE	1	134.3	134.3	0.504	0.4799
Residuals	75	19980.9	266.4		

Table 4. Experiment II: Fitting parameters for strength (impairment) curve.

Group	Xmean	Beta	Residuals
1	6.77	0.33	0.61
2	6.86	0.29	0.14
3	6.83	0.53	0.24
4	5.73	0.5	0.43
5	5.92	0.6	0.84
6	5.79	0.67	0.46
7	7.03	0.35	0.83
8	7.1	0.28	0.65
9	7.19	0.39	0.8
10	6.15	0.39	1.34
11	6.45	0.27	0.09
12	6.3	0.29	0.34
13	6.35	0.48	0.35
14	6.77	0.61	0.11
15	6.57	0.39	0.46
16	6.28	0.38	0.22
17	6.26	0.51	0.86
18	6.35	0.57	0.79
19	5.84	0.46	0.46
20	6.3	0.3	0.77
21	6.26	0.34	0.76

Table 5. Experiment II: Fitting Parameters for MSV versus MAV.

Group	a	b	c	d	err	r	t value	P value
1	1.040	-5.580	0.000	0.000	8.260	0.940	3.950	0.060
2	0.980	-2.850	0.000	0.000	9.910	0.940	3.800	0.060
3	1.110	-2.840	0.000	0.000	8.940	0.900	2.900	0.100
4	0.790	14.680	0.000	0.000	8.240	0.880	2.610	0.120
5	1.180	-8.120	0.000	0.000	3.920	0.980	6.800	0.020
6	1.420	-33.260	0.000	0.000	5.490	0.960	4.790	0.040
7	0.810	-0.510	0.000	0.000	6.240	0.850	2.260	0.150
8	0.780	-2.670	0.000	0.000	0.720	1.000	29.610	0.000
9	0.730	5.350	0.000	0.000	16.240	0.320	0.480	0.680
10	0.560	-2.030	0.000	0.000	9.000	0.780	1.750	0.220
11	1.050	-12.920	0.000	0.000	4.530	0.990	8.570	0.010
12	1.270	-29.400	0.000	0.000	11.400	0.930	3.490	0.070
13	0.980	-3.980	0.000	0.000	3.040	0.990	8.250	0.010
14	1.460	-18.150	0.000	0.000	4.810	0.970	6.000	0.030
15	0.860	0.890	0.000	0.000	8.230	0.920	3.260	0.080
16	0.620	5.630	0.000	0.000	11.140	0.710	1.420	0.290
17	1.040	-10.920	0.000	0.000	8.250	0.920	3.320	0.080
18	1.460	-34.980	0.000	0.000	10.160	0.940	3.870	0.060
19	0.610	15.900	0.000	0.000	3.540	0.970	5.430	0.030
20	0.900	1.070	0.000	0.000	6.580	0.970	5.310	0.030
21	0.880	4.590	0.000	0.000	14.240	0.710	1.440	0.290
AllLin	1.000	-5.120	0.000	0.000	68.390	0.870	16.260	0.000
AllQuad	0.000	0.790	-1.630	0.000	67.700	0.880	16.480	0.000
AllCubic	0.000	0.010	0.610	-0.030	67.650	0.880	16.500	0.000

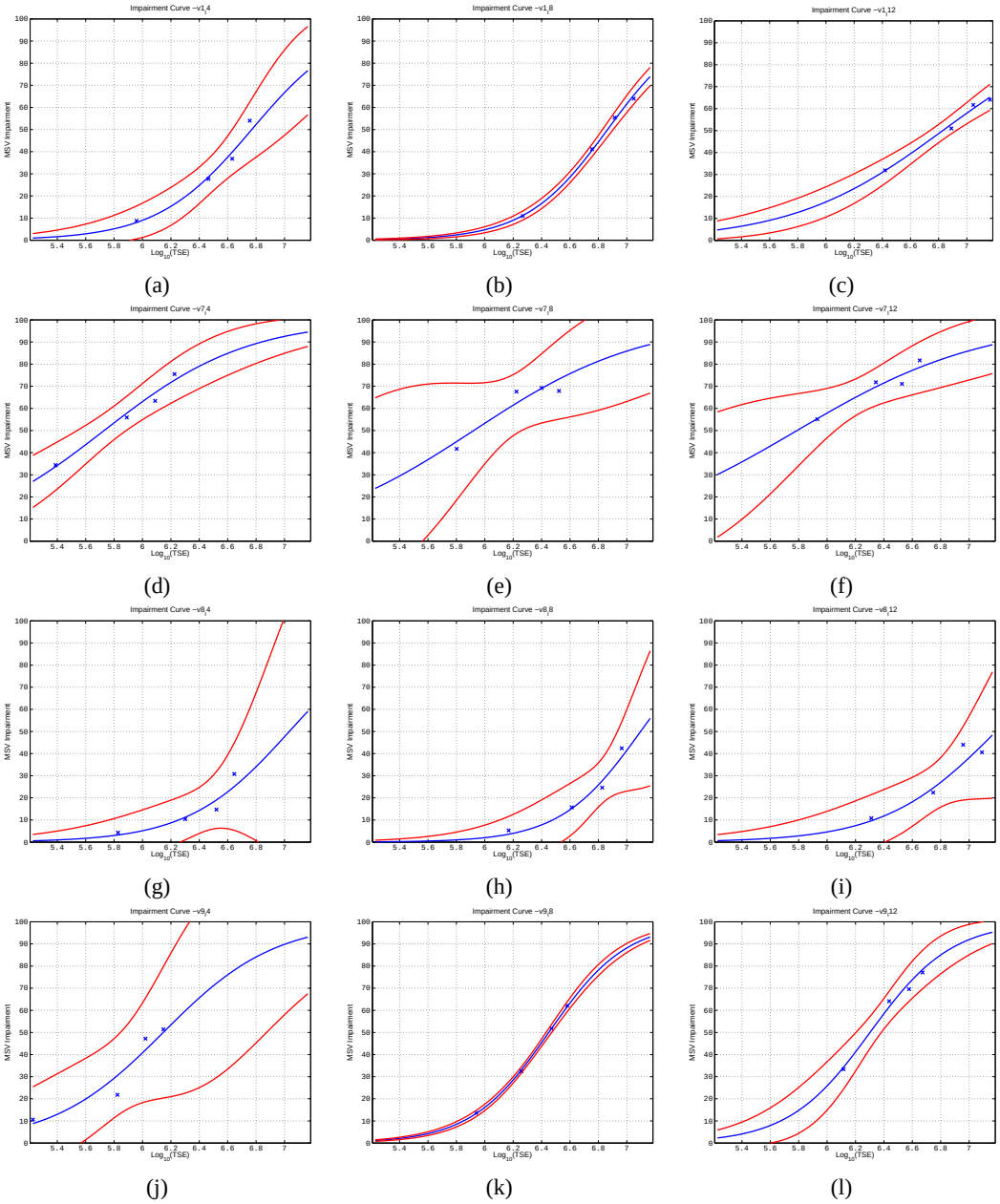


Fig. 6. Experiment II - Impairment curves for test groups (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, (f) 6, (g) 7, (h) 8, (i) 9, (j) 10, (k) 11, and (l) 12.

noisiness) commonly found in digital videos and to understand how these artifacts combine and interact to produce overall annoyance. The results showed that when the artifact signals were presented alone at a high strength, subjects were able to correctly identify them. At low strengths, on the other hand, other artifacts were reported. Annoyance increased with both the number of artifacts and their strength. The noisy artifact signals seemed to decrease the perceived strength of the other artifacts, while blurry artifact signals seemed to increase them.

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

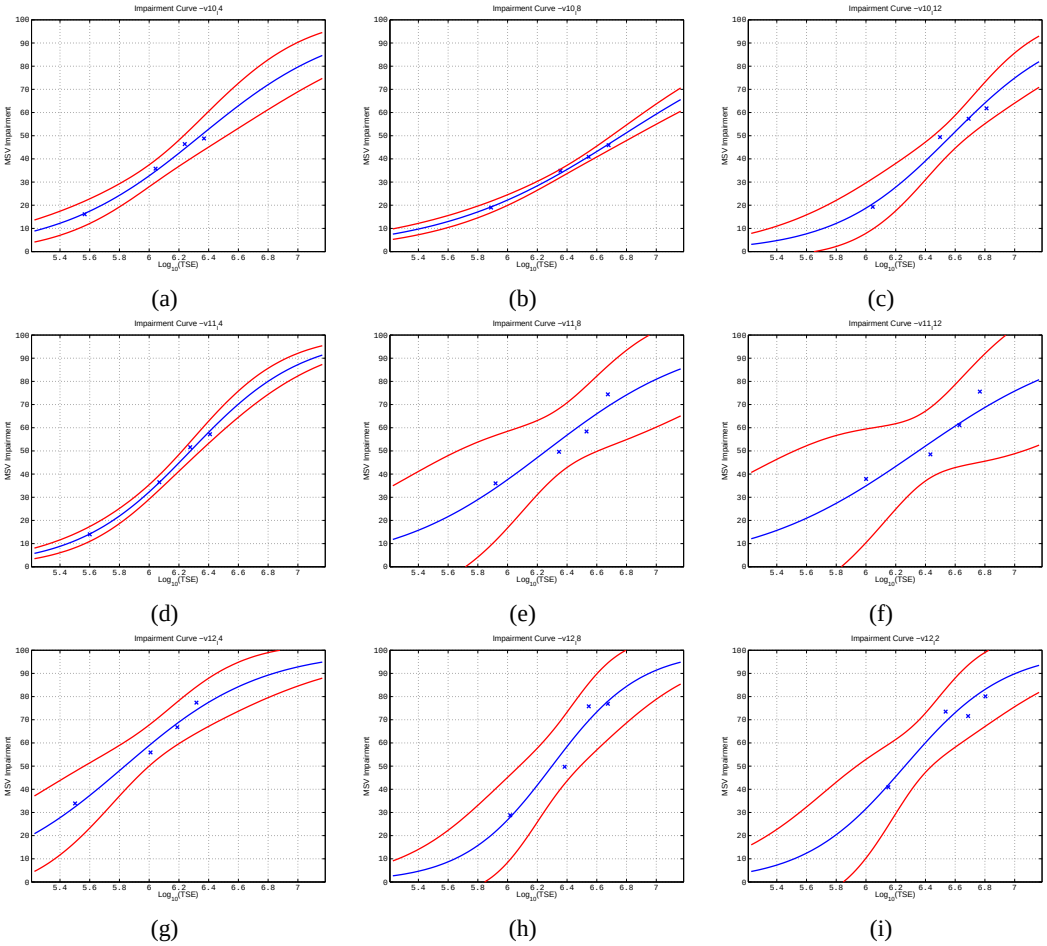


Fig. 7. Experiment II - Impairment curves for test groups (a) 13, (b) 14, (c) 15, (d) 16, (e) 17, (f) 18, (g) 19, (h) 20, (i) 21.

6. REFERENCES

[1] M. Yuen and H. R. Wu, "A survey of hybrid MC/dpcm/dct video coding distortions," *Signal Processing*, vol. 70, pp. 247–278, Oct. 1998.

[2] B. Keelan, *Handbook of image quality: characterization and prediction*. Marcel Dekker, Inc., New York, 2002.

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

[4] W. Lin and C. C., *Jay Kuo*. Perceptual Visual Quality Metrics: A Survey. J. Vis. Commun, 2011.

[5] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms : What does the future hold?" *Intern. Journal of Multimedia Tools and Applic., Vol:*, vol. 51, no. 2, pp. 675–696, Feb. 2011.

[6] H. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 317–320, 1997.

[7] Z. Wang, A. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE International Conference on Image Processing*, vol. 3, 2000, pp. 981–984.

[8] J. Caviedes and J. Jung, "No-reference metric for a video quality control loop," *Proc*, vol. 13, pp. 290–5, Jul. 2001.

[9] M. Farias and S. Mitra, "No-reference video quality metric based on artifact measurements," in *Proc. IEEE Intl. Conf. on Image Processing*, 2005, pp. III: 141–144.

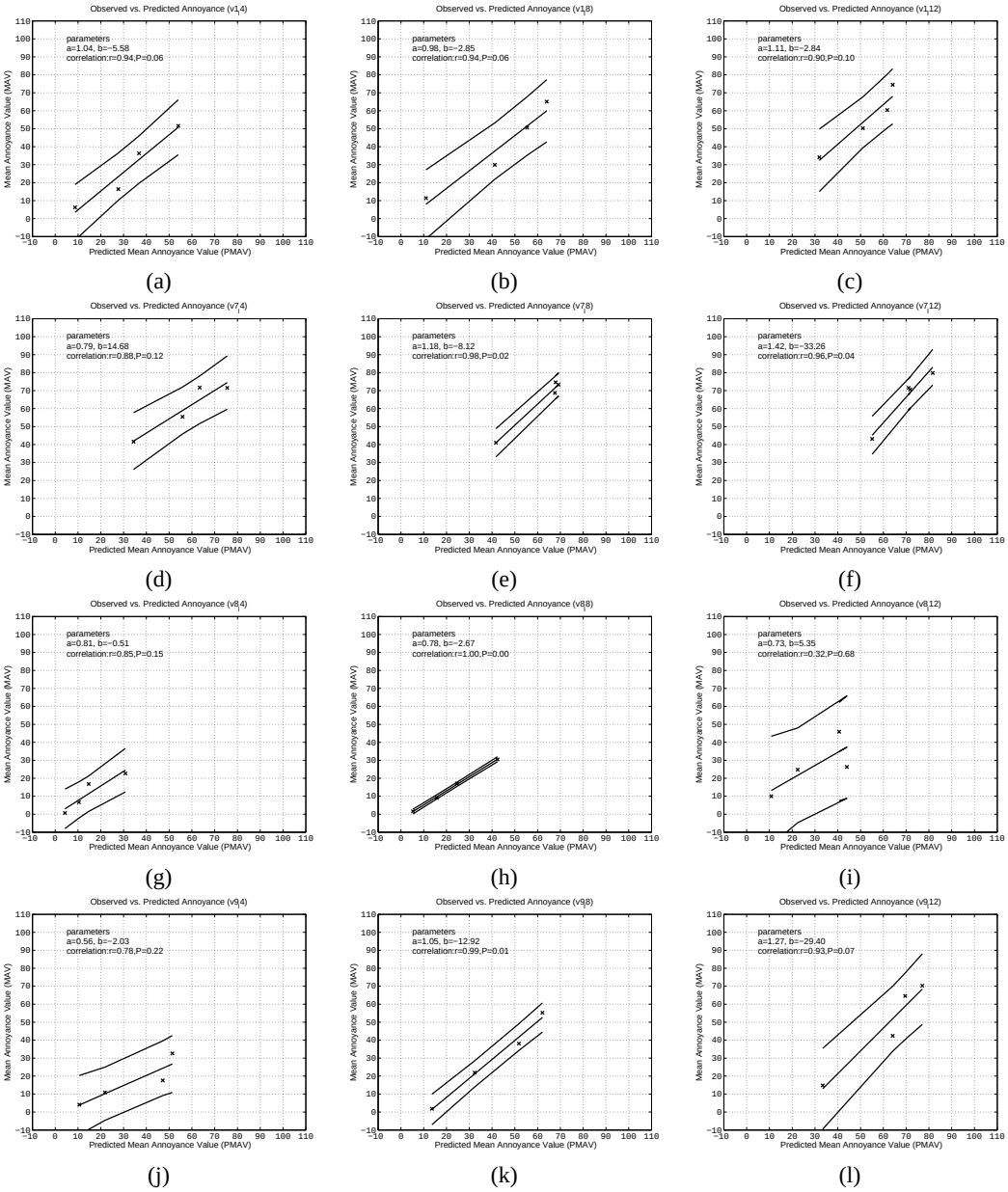


Fig. 8. Experiment II - Linear fitting curves for test groups (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, (f) 6, (g) 7, (h) 8, (i) 9, (j) 10, (k) 11, and (l) 12.

[10] 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.

[11] H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *Proc. SPIE Conference on Human Vision, Visual Processing and Digital Display III*, vol. 1666, San Jose, CA, USA, 1992, pp. 16–26.

[12] M. Farias, "No-reference and reduced reference video quality metrics: New contributions," Ph.D. dissertation, University of California, Santa Barbara, California, 2004.

[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] A. Moorthy and A. Bovik, "Visual quality assessment algorithms : What does the future hold?" *International Journal of*

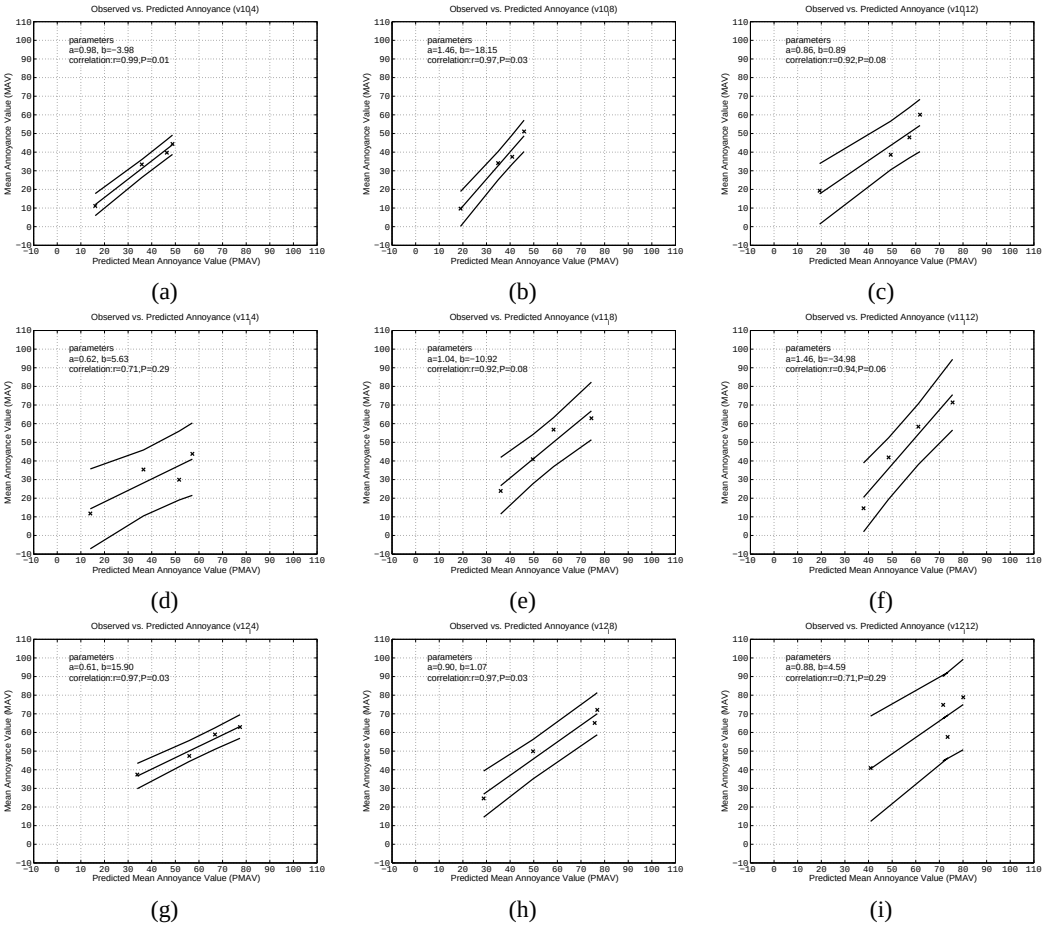


Fig. 9. Experiment II - Linear fitting curves for test groups (a) 13, (b) 14, (c) 15, (d) 16, (e) 17, (f) 18, (g) 19, (h) 20, (i) 21.

Multimedia Tools and Applications, Special Issue on Survey Papers in Multimedia by World Experts, vol. 51, pp. 675–696, 2011.

[15] M. C. Q. 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*, vol. 55, pp. 2954–2964, 2007.

[16] 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.

[17] F. Boulos, B. Parrein, P. L. Callet, and D. Hands, "*Perceptual Effects of Packet Loss on H*. 264/AVC Encoded Videos", Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2009.

[18] A. R. Reibman and D. Poole, *Predicting packet-loss visibility using scene characteristics*. Packet Video, Nov. 2007.

[19] T. Liu, Y. Wang, J. Boyce, and H. Y. Z. Wu, "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 280–293, Apr. 2009.

[20] N. Staelens, G. V. Wallendael, K. Crombecq, N. Vercammen, J. D. Cock, B. Vermeulen, R. V. de Walle, T. Dhaene, and P. Demeester, ""no-reference bitstream-based visual quality impairment detection for high definition h," *264/AVC Encoded Video Sequences, " Broadcasting, IEEE Trans. on, vol. 58, no*, vol. 58, no. 2, pp. 187–199, Jun. 2012.

[21] *ITU-T Recommendation P.930: Principles of a reference impairment system for video*, International Telecommunication Union, 1996.

[22] A. Ostaszewska and R. Kloda, “Quantifying the amount of spatial and temporal information in video test sequences,” in *Recent Advances in Mechatronics, Springer*, 2007, p. 1115.

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

[24] M. Moore, “Psychophysical measurement and prediction of digital video quality,” Ph.D. dissertation, University of California Santa Barbara, 2002.