On the impact of packet-loss impairments on visual attention mechanisms

Judith Redi, Ingrid Heynderickx
Intelligent Systems
Delft University of Technology
Delft, The Netherlands

Bruno Macchiavello[1], Mylene Farias[2]
[1]Department Computer Science
[2]Department Electrical Engineering
University of Brasilia
Brasilia, Brazil

***Abstract*— This paper reports the results of a psychometric experiment aimed at investigating the visibility and annoyance of packet-loss artifacts, with a focus on understanding to what extent their presence influences viewing behavior. The rationale behind this study is that packet loss artifacts might "distract" visual attention, creating visual saliency themselves, thereby becoming more visible. In turn, higher artifact visibility might impact their annoyance. Our study involved seven videos, compressed at a very high bitrate (thus free of spatial artifacts), impaired by discarding packets at different packet loss ratios. We tracked the observers' eye-movements while (1) they assessed the annoyance of impaired videos and (2) looked freely at pristine videos. Our results show that the viewing behavior significantly changes from pristine to impaired videos. This change is related to both properties of the video and to the annoyance of the artifacts presented.**

I. Introduction

When digitally compressed, videos become more susceptible to artifacts originating from the loss of bitstream packets during transmission. These artifacts are spatially localized, and when combined with other spatial artifacts introduced by the compression process (such as blurriness and blockiness), can become very annoying [1]. To ensure high Quality of Experience (QoE), it is thus of primary importance to deploy objective quality metrics that can automatically detect the appearance of such artifacts, quantify their annoyance in a way that is consistent with visual perception, and steer corrective or quality improvement actions [2].

As a first step to design such objective quality metrics, we investigate the annoyance provoked exclusively by "packet loss" artifacts. Our focus is on understanding to what extent packet loss artifacts influence viewing behavior, and whether this influence has an impact on the resulting annoyance. Multiple studies [3-6] showed indeed that a relationship exists between viewing behavior, impairment visibility and annoyance, and that unveiling this relationship can be beneficial to the design of objective quality metrics [7,8]. This relationship has been shown to hold for packet-loss artifacts [6], which resulted more annoying when appearing in the region of interest of a video than in less salient areas. What has been overlooked so far is the potential of packet loss impairments for "distracting" visual attention and creating visual saliency themselves. This would make them more visible and have an effect on their annoyance. To better understand this aspect, we designed an eye-tracking study during which subjects were asked to (1) freely look at pristine videos and (2) judge annoyance of a set of impaired versions of those videos. To detect changes in the viewing behavior, we conducted an analysis of the saliency maps [9], which are visual representations of the probability that a location (pixel) in a scene is attended by the average observer. Changes in saliency between pristine and impaired videos can point out a "distraction" of visual attention from the natural region of interest to other areas in the scene (e.g., where packet-loss artifacts appear). We show in the remainder of this paper that these changes take place and that they are related to the annoyance of the impairments appearing in the videos

II. Experimental setup

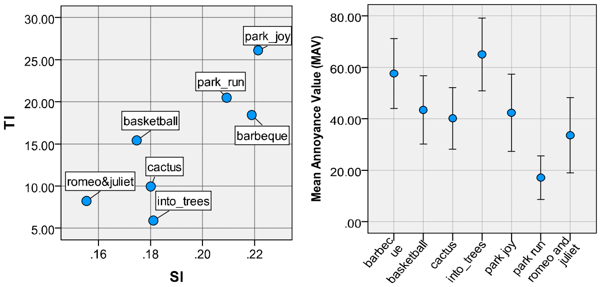
A. Video Material

We selected seven 720p videos from the Consumer Digital Video Library [10], representing different types of content and covering different motion characteristics. The first frame of each video is depicted in Figure 1. Their Spatial (*SI*) and Temporal (*TI*) perceptual information measures (computed as per [11]) are illustrated in Figure 2.a. All videos were encoded using the H.264/AVC codec at a very high bitrate

'Park Joy' 'Into Trees' 'Park Run' 'Romeo and Juliet'

'Cactus' 'Basketball' 'Barbecue'

Figure 1. Screenshots of the first frame of the sequences included in the packet loss visibility study.



(a) *(b)*

Figure 2. (a) Temporal and spatial characteristics of the videos included in the experiment; (b) Mean Annoyance Values averaged across all the distorted versions of each video.

(approximately 120 Mbps), with 8 packets per frame. This avoided the introduction of spatial artifacts in the compressed videos. The coding process generated three sequences per video, with a Group of Pictures (*GOP*) size set to 4, 8 and 12, respectively. Packet loss artifacts were then generated by dropping data packets from the bitstream. In video transmission there are several techniques that can be used to avoid undecodable bitstreams, like for example re-transmission of the parameter packets. To avoid generating unrealistically strong artifacts, a simple error concealment technique was used that consisted of replacing a lost packet by the co-located packet from the previous frame during decoding. The packet loss ratios, *p_loss*, for all videos were 0.7%, 2.6%, 4.3% and 8.1%. These parameters were considered to replicate settings commonly found in real-world video streaming applications. In total, we generated 7 videos x 3 GOP x 4 *p_loss* ratios = 84 sequences. The 7 pristine videos were added to the set, making a total of 13 different settings per original video and resulting in 91 test sequences.

B. Methodology

Fifteen observers were asked to view the sequences and indicate whether they perceived any impairment in the videos; if so, they had to express how annoying the artifacts were on a continuous annoyance scale ranging between 0 to 100. The number of observers was chosen in accordance with ITU recommendations [12] and other studies in the field [4-7]. A single stimulus setup with implicit reference [12] was adopted for the task. All observers went through a thorough training stage in order to help them understanding their task and the characteristics of the artifacts included in the test sequences. After the training, the actual experiment started, which was divided in two sessions to avoid fatigue effects.

The stimuli were displayed on a 23" Samsung LCD monitor (Sync Master XL2370HD). The distance between the subject's eyes and the video monitor was kept at 3 times the screen height using a chinrest. Illumination settings were compliant to ITU-T BT.500-11 specification [12].

A SensoMotoric Instruments iView X RED Eye Tracker was used throughout the experiment to record the eye-movements of the participants. A free viewing session was also performed at the beginning of every experiment, where participants were asked to freely look at the 7 pristine videos. The eye tracker had a sampling rate of 50/60 Hz, a pupil tracking resolution of 0.1° and a gaze position accuracy of 0.5 - 1°.

III. Preliminary analysis: Mean annoyance values and fixation duration

Mean Annoyance Values (MAV) were computed for each of the 91 sequences, following the procedure advised in [12]. As we were primarily interested in viewing behavior in presence of packet-loss artifacts, a thorough analysis of MAV in relation to the video characteristics is considered outside the scope of this paper. On the other hand, we report the Mean MAV (MMAV) computed across all 13 versions of a specific video, as shown in Figure 2.b. It is interesting to notice how the MMAV changes with video content: the videos *into_trees* and *barbecue* obtain on average a high annoyance score (MMAV of 65.07 and 57.61 on a 100 point scale, respectively), whereas impaired versions of the video *park run* seem to present overall just slightly annoying artifacts (MMAV of 17.15). These data suggest that some specific video content might be responsible for masking packet-loss impairments. This observation is in line with previous results obtained by Mantel *et al.* [5] and should be taken into account in the following analysis.

To analyze the viewing behavior we rely throughout the rest of the paper on fixation data only, since these data are most informative with respect to the location of relevant areas in the image [9]. As a preliminary analysis of the viewing behavior we looked into the duration of the fixations recorded during both the free looking and the scoring task. Figure 3 shows the average fixation duration per video content for both tasks. The videos are ordered according to increasing MMAV. Differently from previous results in the field [4, 5], we did find a significant difference (F = 24.72, df = 1, P = 1.38 e-006, computed over the average fixation duration per video and observer, i.e. 7x15 = 105 cases) between the fixation duration when freely watching videos (390 ± 22 ms) and when scoring (476 ± 25 ms). Furthermore, it seems that the duration of fixation rather than depending on MMAV (as found in [5]), depends on the spatial and temporal properties of the video content. To further look into this aspect, we define $Diff\_fd_{i,k}$ as the difference in fixation duration between the two tasks for video content i distorted with setting k:

$$Diff\_fd_{i,k} = fd\_SC(v_{i,k}) - fd\_FL(v_i) \qquad i = 1,...,7 \qquad k = 1,...13$$

where $fd\_FL(v_i)$ is the average fixation duration (across all fixations of all observers) when freely looking at video v_i and $fd\_SC(v_{i,k})$ is the average fixation duration when scoring the k-th impaired version of video v_i. Table I shows a significant correlation of *Diff_fd* with the perceptual characteristics of the pristine videos (SI and TI) and a weaker correlation with parameters that regulate the impairments (p_loss and GOP_size). These results suggest that together with a strong

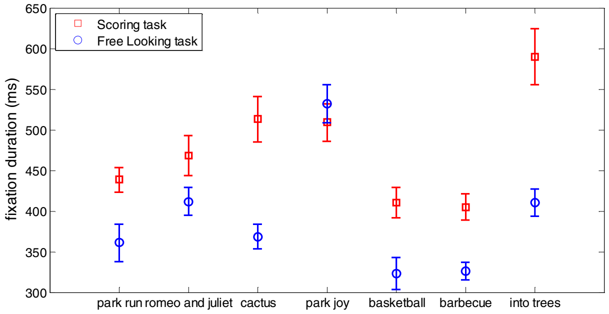


Figure 3. Average fixation duration (in ms) per video for the scoring (red squares) and the free looking (blue circles) tasks. Videos are sorted according to increasing MMAV

TABLE I. CORRELATION BETWEEN THE DIFFERENCE IN FIXATION DURATION AMONG TASKS (SCORING VS. FREE-LOOKING) AND THE SPATIAL AND TEMPORAL CHARACTERISTICS [11] OF THE VIDEOS

		Impairment characteristics		**Video characteristics**	
		GOP_size	***P_loss***	***SI***	***TI***
Diff_fd	Correlation	.215*	.229*	-.334**	-.600**
	significance	.041	.029	.001	.000

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

influence of the video content, packet loss artifacts might also affect viewing behavior.

IV. ANALYSIS OF ATTENTION DEPLOYMENT

To look deeper into the changes in viewing behavior on the appearance of packet loss impairments, we conducted an analysis according to the methodology suggested in [9]. In videos, saliency maps are estimated over time windows in order to take into account the possibility that the content of the scene (and therefore its region of interest) changes over time. Typically, the length of a time window is set to the length of a video frame (in our case, 20 ms, as our videos were rendered at 50 fps). We argue that such a short time frame gives a too fine granularity, yielding too much variability among saliency maps. We define a time window of 400 ms, being this value close to the average duration of a fixation, as found in our experiment (see Section III).

To compute the saliency map for timeslot *t* of video *v* we first gathered all related fixations deployed by all observers into a fixation map. Then, we converted the fixation map into a saliency map by applying a Gaussian patch with a width approximating the size of the fovea (about 2° visual angle) to each fixation (as per [9]).

The above procedure was applied to the following sets of videos, creating for each of them 10000/400(ms) = 25 saliency maps:

- FL_G1: the 7 pristine videos watched under a free looking task;
- SC_HQ: the 7 pristine videos watched under the annoyance scoring task;
- SC_LQ: the 84 impaired videos watched under the annoyance scoring task.

As an indicator of changes in viewing behavior due to packet-loss artifacts, we looked into (dis)similarities among saliency distributions for the three groups of maps mentioned above. Several measures exist to quantify similarities among saliency maps [9]; in this work we used:

(1) the Linear Correlation Coefficient (LCC) ∈ [-1, 1], which quantifies the strength of the linear relationship between two saliency maps. A value of LCC = 1 indicates perfect similarity among maps, while LCC = 0 indicates uncorrelated maps.

(2) The Structural Similarity Index (SSIM, [13]) ∈ [-1, 1], which indicates the extent to which the structural information of a map is preserved in another one. Also here, SSIM = 1 indicates perfect similarity among maps.

Similarity among maps strongly depends on the scene content and is highly affected by inter-observer variability [9]. As a consequence, we adopted here the notion of upper empirical similarity limit (UESL). UESL is defined as the maximum achievable similarity between the saliency maps derived from two different (groups of) humans under the same experimental conditions. We chose the free looking task as a reference condition, and to compute the UESL we compared the saliency maps obtained for the FL_G1 videos with a set of maps derived from the eye-tracking data of a second group of 15 observers (distinct from the 15 that took part in the present experiment), who freely looked at the 7 pristine videos in the exactly same experimental conditions. We refer to this set of video as FL_G2. The UESL for video content *i* (*i* = 1,…, 7) and similarity measure S (S∈{LCC, SSIM}) is, therefore, computed as:

$$UESL(S,i)=\frac{1}{T}\sum_{t=1}^{T}S(SM(v_{i,t}^{G1}),SM(v_{i,t}^{G2})) \qquad (1)$$

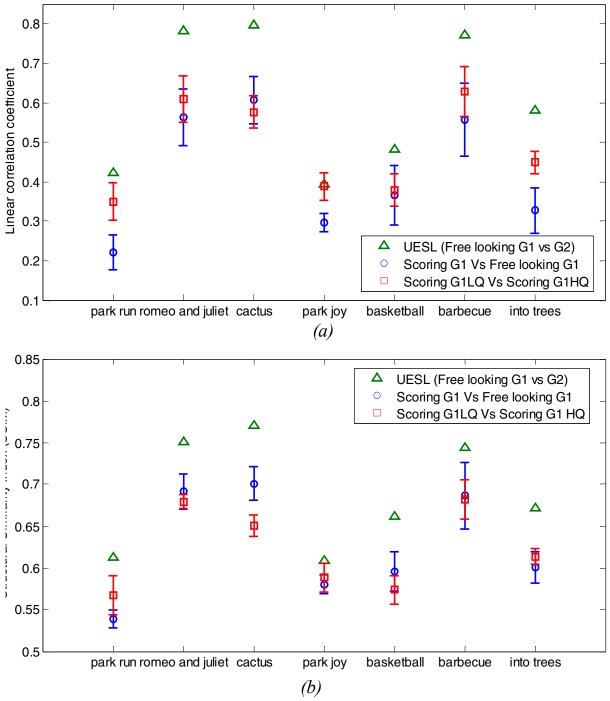
Where $SM(v_{i,t}^{G})$ indicates the saliency maps obtained for timeslot *t* of video *i* for the observer group G.

We then computed the similarity among the saliency maps obtained for freely looking at pristine videos (FL_G1) and scoring impaired videos (SC_LQ), across all C=12 impairment conditions:

$$\begin{aligned}&SC\_FL^{G1}(S,i)=\\&=\frac{1}{T*C}\sum_{k=1}^{C}\sum_{t=1}^{T}S(SM(v_{i,k,t}^{SC\_LQ}),SM(v_{i,t}^{FL\_G1}))\end{aligned} \qquad (2)$$

The resulting values are shown in figure 4 for both LCC (a) and SSIM (b). It is straightforward to notice that the *SC_FL*^{G1} values are significantly lower than the UESL. This implies that when scoring impaired videos, the visual attention diverges from the one it would follow when freely looking at pristine videos. This result is in contrast with that obtained by Le Meur et al. [4]. A possible reason for this is that videos in [4] presented visible coding artifacts, evenly spread across the whole video content, whereas our videos were impaired only by packet loss artifacts, which are strongly locally and temporally localized, and as such could create saliency on their own.

Before validating this hypothesis, it should be noticed that the change in saliency might also result from the change in viewing task (scoring instead of free looking). To evaluate the impact of task on our results, we computed the similarity



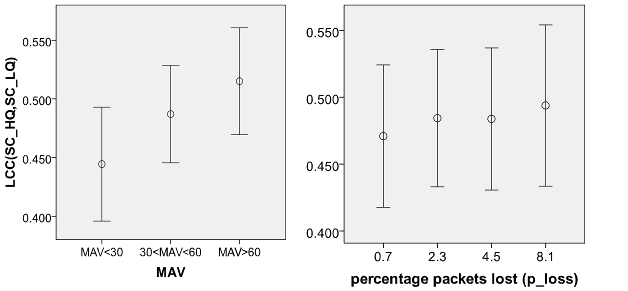
(a)

(b)

Figure 4. Similarity among saliency maps obtained for free looking of pristine videos and scoring of impaired videos, computed through (a) LCC and (b) SSIM. Videos are ordered according to increasing MMAV.

among the saliency maps obtained for scoring pristine videos and scoring impaired videos, by computing the quantity $SC^{HQ}\_SC^{LQ}$ by substituting SC_HQ to FL_G1 in equation (2). The resulting values are represented by red squares in Figure 4. If the dissimilarities in saliency between SC_LQ and FL_G1 were only due to task, we would expect $SC^{HQ}\_SC^{LQ}$ (similarity among maps obtained under the same scoring task) to have values comparable to the UESL. Figure 4 shows instead that this is not the case, and that saliency for impaired videos is distributed in a different way than for pristine videos, even under the same scoring task. Therefore, the appearance of packet-loss artifacts seems to distract attention from its natural deployment.

As a next step, we investigated how this change of attention depended on objective parameters regulating the artifact appearance and on artifact annoyance. Figure 5.a shows how $SC^{HQ}\_SC^{LQ}$(LCC) varies depending on the MAV of the video. For visualization purposes we summarized videos into 3 categories: those having a MAV < 30, i.e. presenting just slightly annoying artifacts, those having 30<MAV<60 and those with MAV > 60, i.e. with highly annoying artifacts. Interestingly, MAV and $SC^{HQ}\_SC^{LQ}$ seem to be positively correlated ($R^2 = 0.249$, significant at the 0.05 level), that is, the similarity among saliency maps derived from scoring pristine and impaired videos increases with the annoyance of the artifacts. Also, quite interestingly $SC^{HQ}\_SC^{LQ}$ seems not to be correlated to the percentage of packets lost in the transmission. A possible explanation for this is that, depending on the spatial and temporal properties of the specific video contents, packet loss artifacts can be masked, and thus their potential for creating saliency might depend on the specific video content.



(a) (b)

Figure 5. Factors influencing the similarity among saliency maps obtained for scoring of pristine and impaired videos, computed through LCC: (a) MAV and (b) percentage of lost packets

The study reported in this paper shows that visual attention significantly shifts from its natural path with the appearance of packet loss artifacts, and that this change seems to be related to both properties of the video and to the annoyance of the artifacts presented. The spatial nature of this attention change needs to be further investigated and incorporated in models for the automated assessment of video quality.

References

[1] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, “Perceptual effects of packet loss on H.264/AVC encoded videos,” Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM), 2009

[2] W. Lin and C.-C. Jay Kuo, “Perceptual Visual Quality Metrics: A Survey,” Journal of Visual Commun. and Image Representation, vol. 22, no. 4, pp. 297-312, 2011

[3] J.A. Redi, H. Liu, R. Zunino, and I. Heynderickx, “Interactions of visual attention and quality perception,” in: IS&T/SPIE Human Vision and Electronic Imaging XVI. Vol 7865, 2011.

[4] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, “Do video coding impairments disturb the visual attention deployment?” Signal Process. Image Commun.,vol. 25, no. 8, pp. 597–609, 2010.

[5] C. Mantel, N. Guyader, P Ladret, G. Ionescu and T. Kunlin “Characterizing eye movements during temporal and global quality assessment of h.264 compressed video sequences” in Proc. SPIE 8291, Human Vision and Electronic Imaging XVII, 2012.

[6] U. Engelke, H. Kaprykowsky; H.-J. Zepernick and P. Ndjiki-Nya; “Visual Attention in Quality Assessment,” IEEE Signal Processing Magazine, vol. 6, pp. 50-59, 2011

[7] H. Liu, and I. Heynderickx, “Studying the added value of visual attention in objective image quality metrics based on eye movement data,” in Proc. of IEEE Int. Conf. on Image Processing, Nov. 2009.

[8] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba,“Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric,” Elsevier, Signal Process. Image Commun., 2010

[9] Redi, J., and Heynderickx, I., “Image Quality And Visual Attention Interactions: Towards A More Reliable Analysis In The Saliency Space,” Proceedings of IEEE QoMEX, Sep.2011

[10] M. Pinson, S. Wolf, N. Tripathi, and C. Koh, “The consumer digital video library,” in Proc. Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM), Jan. 2010.

[11] A. Ostaszewska and R. Kloda, “Quantifying the amount of spatial and temporal information in video test sequences,” in Recent Advances in Mechatronics. Berlin/Heidelberg: Springer, 2007, pp. 11–15.

[12] International Telecommunication Union, “Methodology for the subjective assessment of the quality of television pictures,” Rec. BT.500-11, ITU-R, 2002.

[13] Z. Wang and A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Proc., Vol. 13, no. 4, 2004.