Linking Distortion Perception and Visual Saliency in H.264/AVC Coded Video Containing Packet Loss

Ulrich Engelke[a], Romuald Pepion[b], Patrick Le Callet[b], and Hans-Jürgen Zepernick[a]

[a]Blekinge Institute of Technology, PO Box 520, 372 25 Ronneby, Sweden

[b]IRCCyN UMR no 6597 CNRS, Ecole Polytechnique de l'Universite de Nantes, rue Christian Pauc, La Chantrerie, 44306 Nantes, France

ABSTRACT

In this paper, distortions caused by packet loss during video transmission are evaluated with respect to their perceived annoyance. In this respect, the impact of visual saliency on the level of annoyance is of particular interest, as regions and objects in a video frame are typically not of equal importance to the viewer. For this purpose, gaze patterns from a task free eye tracking experiment were utilised to identify salient regions in a number of videos. Packet loss was then introduced into the bit stream such as that the corresponding distortions appear either in a salient region or in a non-salient region. A subjective experiment was then conducted in which human observers rated the annoyance of the distortions in the videos. The outcomes show a strong tendency that distortions in a salient region are indeed perceived as much more annoying as compared to distortions in the non-salient region. The saliency of the distorted image content was further found to have a larger impact on the perceived annoyance as compared to the distortion duration. The findings of this work are considered to be of great use to improve prediction performance of video quality metrics in the context of transmission errors.

Keywords: Visual saliency, distortion perception, video transmission, packet loss, subjective experiment.

1. INTRODUCTION

Recent advances in wireless and wireline communication networks facilitated the transition from conventional voice services to more elaborate multimedia services, including packet based video streaming over IP. This has also been enabled through a rapid progress of video coding standards, such as H.264/AVC[1,2] which allows for encoding with half the bit rate as compared to its predecessors, while maintaining a similar level of visual quality. However, the stringent bandwidths restrictions of the networks and the strong compression of the source signal result in the video content being highly prone to packet loss during transmission. Unlike pure source coding distortions that usually appear on a global scale in a frame, packet loss can cause strong, localised distortions, both spatially over a frame and also temporally over a number of frames due to error propagation.[3] Objectively measuring packet loss and its resulting visual distortions is of great importance for service providers to provide a certain level of quality of experience to the end user. To this end, the temporal effects of packet loss have been evaluated by different research groups.[4–7] However, there has not been many efforts so far to evaluate impact of distortions in relation to the spatial saliency of their location within a video frame.

In this respect, one property of the human visual system (HVS), that is well known as visual attention (VA),[8] is of particular interest. This property allows to reduce the complexity of visual scene processing by selecting only a subset of the available information by rapidly scanning the visual scene and focusing only on the most salient regions. Thus, one can expect that the viewer may be more likely to detect distortions in salient regions as compared to non-salient regions. Furthermore, given that the HVS is highly space variant in sampling and processing of visual signals, with the highest accuracy in the central point of focus, the fovea,

Further author information: (Send correspondence to U.E)
U.E: E-mail: ulrich.engelke@bth.se, Telephone: +46 (0)768 845 877
R.P.: E-mail: romuald.pepion@irccyn.ec-nantes.fr, Telephone: + 33 (0)2 4068 3065
P.L.C: E-mail: patrick.lecallet@univ-nantes.fr, Telephone: +33 (0)2 4068 3047
H.J.Z: E-mail: hans-jurgen.zepernick@bth.se, Telephone: +46 (0)708 782 680

distortions in a salient region may also be perceived as more annoying than in a non-salient region. The effect of the saliency driven, bottom-up attention[9,10] on the perception of distortions in a visual scene, is assumed to be particularly high in video signals, where the visual scene changes constantly, unlike with images, where the visual scene is static. Thus, in video there is typically not enough time to inspect the whole scene and as such, distortion perception varies considerably more with the observer's focus of attention. In this respect, the gain of incorporating VA into objective quality models can also be expected to be higher in case where distortions are localised in certain areas of the scene, such as in case of packet loss. The reason for this being that global distortions may not as much distract the attention from the content of the scene.

Thus far, only few works have considered visual saliency to improve prediction performance of image quality metrics[11–14] and even less attention has been given to relating visual saliency with perceived quality for video.[15,16] Only little (if any) improvement of quality prediction performance by considering visual saliency is usually reported in works that are focusing on the context of source coding artifacts[12,14] where distortions are usually globally distributed. On the other hand, larger improvements were found in case of localised distortions due to bit errors and packet loss distortions.[11,13]

These previous works have not reported any qualitative or quantitative analysis regarding the perceptual impact of localised distortions in relation to video content saliency. In this work, we therefore present a subjective experiment that we conducted, with the aim to gain further insight into the impact of visual saliency on the perceived annoyance of localised packet loss distortions during video transmission. For this purpose, we utilised eye tracking data from an earlier experiment as a ground truth to identify saliency in a number of videos. We then introduced packet loss into the bit stream of the video sequences such that the corresponding distortions appear either in a salient region or in a non-salient region. In the subjective experiment we then asked human observers to rate the annoyance of the distortions in the videos. We found that the annoyance of the distortions depends strongly on the saliency of the region that they appear in. It was further revealed that the length of the distortion seems to have a smaller impact on the perceived annoyance as compared to the content saliency.

This paper is organised as follows. In Sec. 2 the creation of the test sequences that were used in the subjective experiment is discussed. Section 3 then explains the subjective experiment that we conducted to identify the perceived annoyance of the localised loss patterns. A detailed analysis of the experiment outcomes is provided in Sec. 4 and conclusions are finally drawn in Sec. 5.

2. CREATION OF TEST SEQUENCES

Within the scope of this work, we considered 30 reference video sequences in standard definition (SD) format available from the Video Quality Experts Group (VQEG).[17] Out of these 30 sequences we selected 20 sequences with respect to the saliency of the content and with regards to the spatial and temporal characteristics of the content. The sequence selection and creation of the distorted test sequences is explained in the following sections.

2.1 Identification of content saliency

We utilised gaze patterns from a previously conducted eye tracking experiment[18] to identify the visual saliency for each of the 30 video sequences. The gaze patterns were recorded using a dual-Purkinje eye tracker from Cambridge Research Systems.[19] In this eye tracking experiment, the sequences were presented to 37 participants under task free condition and as such, the recorded gaze patterns represent the saliency of the visual content.

The gaze patterns were post-processed to eliminate saccades (rapid eye movements that carry the focus of attention) leaving the fixations and smooth pursuit eye movements that contribute to VA. A Gaussian filter was then deployed to create the final saliency maps for each sequence based on the visual fixations of all 37 observers. These saliency maps were visually inspected to identify frames that contain regions of particular high saliency.

2.2 Source encoding and creation of loss patterns

The video sequences were encoded in H.264/AVC format[1] using the JM 16.1 reference software.[21] As we are interested in evaluating the perceptual impact of transmission errors rather than source coding distortions, we encoded the sequences in high quality with a constant quantization parameter of QP=28. The fixed QP further minimizes quality differences between the various sequences, unlike for instance a constant bit rate. The

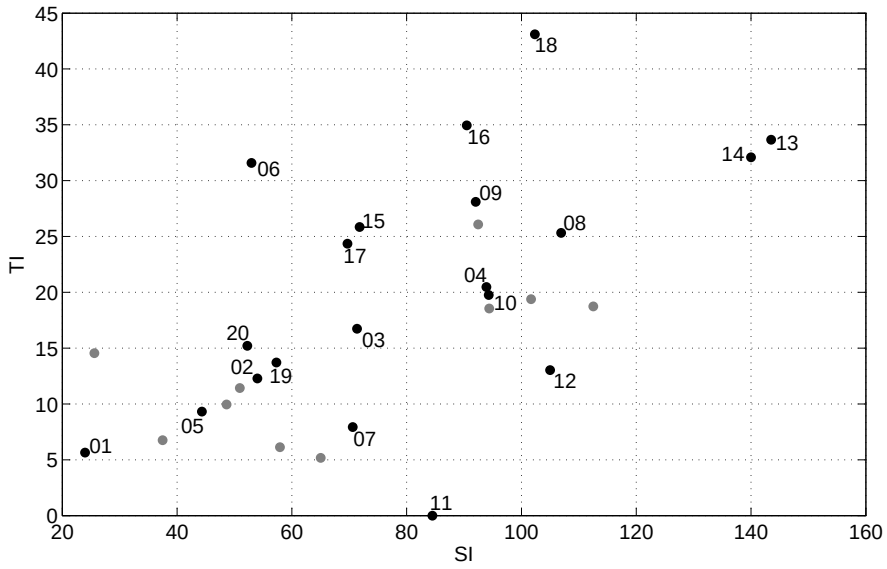


Figure 1. Spatial information (SI) and temporal information (TI) indicators[20] for all 30 sequences. The numbered dots represent the 20 sequences that have been selected for the experiment.

sequences were encoded in High profile with an IBBPBBP... GOP structure and two different lengths; 30 frames (GOP30) and 10 frames (GOP10). The frame rate was set to 25 and as such the two GOP lengths correspond to 1.2 sec and 0.4 sec, respectively.

We utilised an adapted version of the Joint Video Team (JVT) loss simulator[22] to introduce packet loss into the H.264/AVC bit stream. The packet loss was introduced into a single I frame in each sequence resulting in error propagation until the next I frame, due to the inter-frame prediction of the P and B frames. Thus, the two different GOP lengths (30/10 frames) relate to the maximum lengths of error propagation (1.2/0.4 sec). To have better control regarding the location and extent of the corresponding spatial loss patterns we chose a fixed number of macro blocks (MB) of 45 per slice. Given that SD video has dimensions 720×576 pixels, corresponding to 45×36 MB, each slice represents exactly one row of MB.

To identify the impact of saliency on the perception of the distortions, we introduced packet loss into the sequences such as that the corresponding visual distortions appear either in a salient region or in a non-salient region, based on the saliency as identified in Sec. 2.1. In particular, we created test sequences with packet loss introduced in 5 slices centered around the most salient region in an I frame. We then created a corresponding sequence with 5 slices of distortions introduced into a non-salient region of the same I frame. The extent of the loss pattern was intentionally kept constant to allow for a better comparison between distortions in the salient region and the non-salient region. We created such two sequences for both the GOP30 and GOP10 coded sequences, resulting in a total of four distorted sequences for each reference sequence SEQ_R. The subsets of distorted sequences will in the following be referred to as $\mathrm{SEQ}_{S,0.4}$, $\mathrm{SEQ}_{N,0.4}$, $\mathrm{SEQ}_{S,1.2}$, and $\mathrm{SEQ}_{N,1.2}$, where 0.4 relates to error propagation length for GOP10 and, accordingly, 1.2 relates to GOP30. The indices S and N refer to the salient and non-salient regions, respectively.

All sequences were shortened to 150 frames, corresponding to 6 sec duration. During the creation of the test sequences it was assured that no distorted frames were present in the first second and the last second of the video and also not immediately before or after scene cuts.

2.3 Content classification

For the final selection of the 20 test sequences for the experiment we further classified the content using spatial information (SI) and temporal information (TI) indicators.[20] As both SI and TI indicators may change significantly throughout the duration of a sequence, we used for the content classification only the frames containing

the distortions. The SI and TI for all sequences are shown in Fig. 1. The selection of the test videos was done with respect to covering a wide range of SI and TI indicators. In Fig. 1, the numbered dots represent the 20 sequences chosen for the experiment whereas the remaining 10 sequences were not included into the test set.

An example frame for each of the 20 sequences used in the experiment is shown in Fig. 2. In particular, the I frame is presented in which the packet loss was introduced to create the sequences $\text{SEQ}_{S,0.4}$ and $\text{SEQ}_{N,0.4}$. For visualisation purposes in this paper only, the distortions of both the salient region and the non-salient region are presented within the same frame. The salient distortion region is additionally highlighted with green lines (bright in grey-scale images) and the non-salient distortion region is highlighted with red lines (dark in grey-scale images). The saliency information from the task-free eye tracking experiment[18] on the reference images is additionally visualised using heat maps.

3. SUBJECTIVE EXPERIMENT

To identify the perceptual impact of the loss patterns depending on the saliency of their location and also depending on their duration, we conducted a subjective experiment in which human observers where asked to rate the annoyance of the distortions in the video sequences. The experiment procedures were designed according to ITU Rec. BT.500[23] and will be discussed in the following sections.

3.1 Laboratory setup

The laboratory in which the experiment took place was set up with grey covers on all walls and was illuminated with low light levels. The videos were presented on a LVM-401W full HD screen by TVlogic with a size of 40" and a native resolution of 1920×1080 pixels. A mid-grey background was added to the SD test sequences to be displayed on the HD screen. The observers were seated at a distance of about 150 cm corresponding to six times the height of the displayed video sequences.

3.2 Viewer panel

A total of 30 people participated in the experiment out of which 10 were female and 20 were male. The participants were mainly students and staff of the University of Nantes with an average age of about 23 years. Prior to each experiment, the visual accuracy of the participants was tested using a Snellen chart and any colour deficiencies were identified using the Ishihara test.

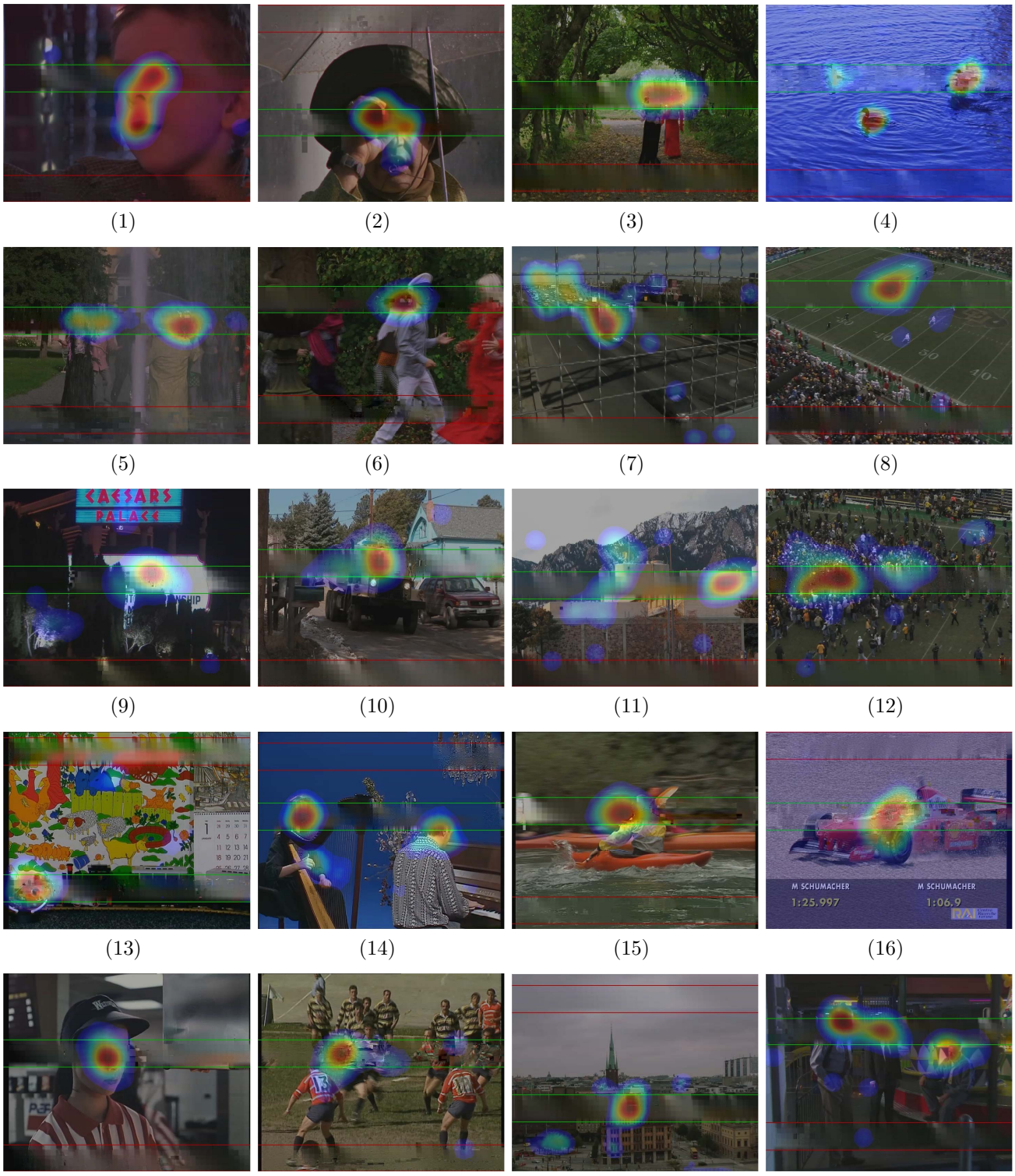
3.3 Experiment procedures

The participants were presented the 100 test sequences (20 reference sequences plus 80 distorted sequences) in a pseudo random order with a distance between the same content of at least 5 presentations. The sequences were presented using a single stimulus method, meaning, that the distorted sequences were presented without their corresponding reference sequence. The reference sequences were randomly mixed with the set of distorted sequences. The participant was not told if the currently presented sequence contained distortions or not.

Before the test sequences the participants were shown 6 training sequences in a fixed order for the participants to adapt to the impairment rating system and to get a feeling for the distortions that can be expected in the test sequences. For this purpose, training sequences were selected from the remaining 10 sequences (see Sec. 2.3 and Fig. 1) that covered the range of distortions in the test sequences.

The 5-grade impairment scale[23] was used to assess the annoyance of the distortions in the sequences. Here, the observers were asked to assign one of the following adjectival ratings to each of the sequences: 'Imperceptible (5)', 'Perceptible, but not annoying (4)', 'Slightly annoying (3)', 'Annoying (2)', and 'Very annoying (1)'. The impairment scale has the advantage over the quality scale, which is also defined in,[23] that the rating 'Imperceptible' directly allows to identify if participants actually detected the distortions in the sequences or not. The impairment scale used in the experiment is shown in Fig. 3 (with the corresponding French labels).

Given the length of 150 frames per sequence and a frame rate of 25 frames per second, each sequence lasted about 6 seconds. Including the time for the impairment rating between the sequences, each experiment lasted about 30-40 minutes. To avoid fatigue of the viewers' eyes we included a break after presentation of about half the sequences.



(1) (2) (3) (4)

(5) (6) (7) (8)

(9) (10) (11) (12)

(13) (14) (15) (16)

(17) (18) (19) (20)

Figure 2. Example frames for all video sequences, visualising both the salient (green/bright lines) and non-salient (red/dark lines) distortion regions and also the saliency information using heat maps.

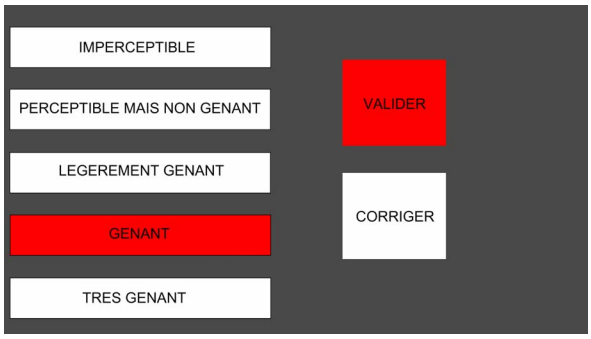


Figure 3. Five-grade impairment scale as utilised in the subjective experiment.[23]

4. ANALYSIS

The outcomes of the subjective experiment will be discussed in the following sections. In particular, the impact of the distortion location and the distortion length on the overall perceived annoyance will be analysed in detail.

4.1 Notation

The 30 subjective impairment ratings for each sequence are averaged into a single score, the mean opinion score (MOS). Corresponding to the subsets of sequences, SEQ_R, $\mathrm{SEQ}_{S,0.4}$, $\mathrm{SEQ}_{N,0.4}$, $\mathrm{SEQ}_{S,1.2}$, and $\mathrm{SEQ}_{N,1.2}$ (see Sec. 2.2), we define subsets of MOS as MOS_R, $\mathrm{MOS}_{S,0.4}$, $\mathrm{MOS}_{N,0.4}$, $\mathrm{MOS}_{S,1.2}$, and $\mathrm{MOS}_{N,1.2}$. We further define MOS differences, Δ_{MOS}, as follows

$$\Delta_{\mathrm{MOS},0.4} = \mathrm{MOS}_{N,0.4} - \mathrm{MOS}_{S,0.4} \quad (1)$$

$$\Delta_{\mathrm{MOS},1.2} = \mathrm{MOS}_{N,1.2} - \mathrm{MOS}_{S,1.2} \quad (2)$$

$$\Delta_{\mathrm{MOS},S} = \mathrm{MOS}_{S,0.4} - \mathrm{MOS}_{S,1.2} \quad (3)$$

$$\Delta_{\mathrm{MOS},N} = \mathrm{MOS}_{N,0.4} - \mathrm{MOS}_{N,1.2} \quad (4)$$

Here, for instance, $\Delta_{\mathrm{MOS},0.4}$ represents the MOS difference between the salient (S) and the non-salient (N) region in case of short distortion propagation of 0.4 sec. Similarly, $\Delta_{\mathrm{MOS},S}$ represents the MOS difference between short (0.4 sec) and long (1.2 sec) distortion propagation in case of distortions in the salient region. These subsets allow us to evaluate the impact of content saliency and distortion duration on the overall annoyance.

4.2 Distribution of annoyance scores

Given the 30 participants and the 100 video sequences, a total of 3000 annoyance scores were collected during the experiment. As such, 600 scores were given for each subset of sequences. The normalised distribution of the scores for the four subsets of distorted sequences is presented in Fig. 4 (the subset of reference sequences, SEQ_R, has been left out as almost exclusively all scores were equal to 5). Here, the number of ratings for each annoyance score have been normalised with respect to the total number of 600 scores within each subset.

Figure 4 shows a strong tendency that the salient region distorted sequences, $\mathrm{SEQ}_{S,0.4}$ and $\mathrm{SEQ}_{S,1.2}$, received in general lower ratings as compared the non-salient region distorted sequences, $\mathrm{SEQ}_{N,0.4}$ and $\mathrm{SEQ}_{N,1.2}$. It can also be observed that the ratings for $\mathrm{SEQ}_{S,0.4}$ and $\mathrm{SEQ}_{S,1.2}$ are generally more spread as compared to $\mathrm{SEQ}_{N,0.4}$ and $\mathrm{SEQ}_{N,1.2}$, which observe high peaks at an annoyance score of 4. These observations indicate, that the ratings are more similar between the sequence subsets that contain distortions in the same region (salient or non-salient) than between the sequence subsets that contain distortions of the same duration (long or short).

To further illustrate the above observations we have conducted a curve fitting of the score distributions using a Gaussian fitting function as follows

$$y(x) = p_1 \cdot e^{-\left(\frac{x-p_2}{p_3}\right)^2}. \quad (5)$$

The fitting function parameters as well as the goodness of fit measures are summarised in Table 1 for all four distorted sequence subsets. Here, the parameter p_1 determines the height of the distribution maximum, the

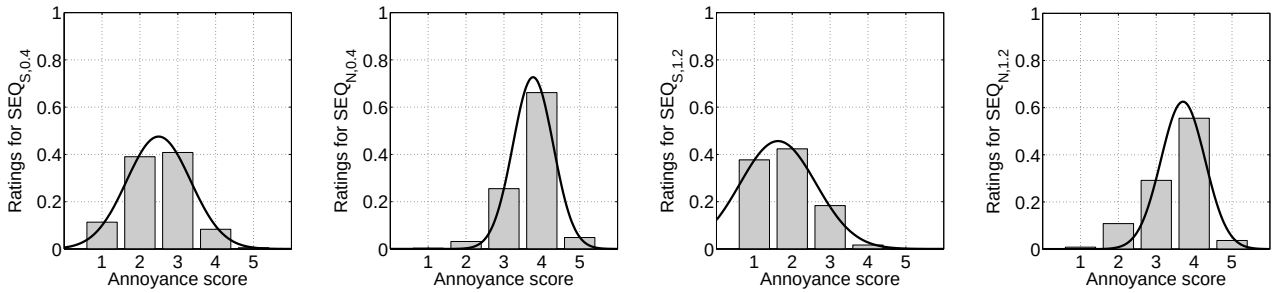


Figure 4. Normalised distributions of the total number of ratings for the 4 distorted sequence subsets: (a) $\text{SEQ}_{S,0.4}$, (b) $\text{SEQ}_{N,0.4}$, (c) $\text{SEQ}_{S,1.2}$, (d) $\text{SEQ}_{N,1.2}$.

Table 1. Gaussian curve fitting for the total number of ratings within all distorted sequence subsets.

Subset	Fitting parameters			Goodness of fit	
	p_1	p_2	p_3	R^2	RMSE
$\text{SEQ}_{S,0.4}$	0.476	2.495	1.195	0.996	0.016
$\text{SEQ}_{N,0.4}$	0.727	3.769	0.753	0.997	0.021
$\text{SEQ}_{S,1.2}$	0.456	1.625	1.418	0.998	0.009
$\text{SEQ}_{N,1.2}$	0.625	3.706	0.824	0.95	0.072

Table 2. MOS averaged over all sequences within the five subsets of sequences.

MOS_R	$\text{MOS}_{N,0.4}$	$\text{MOS}_{N,1.2}$	$\text{MOS}_{S,0.4}$	$\text{MOS}_{S,1.2}$
4.97	3.72	3.5	2.48	1.84

parameter p_2 represents the corresponding value on the annoyance score scale, and the parameter p_3 is related to the width of the Gaussian fitting curve. These parameters provide quantitative evidence that the ratings of $\text{SEQ}_{S,0.4}$ and $\text{SEQ}_{S,1.2}$ are similarly distributed and so are the ratings of $\text{SEQ}_{N,0.4}$ and $\text{SEQ}_{N,1.2}$. The corresponding goodness of fit measures, the root mean squared error (RMSE) and the squared correlation coefficient R^2, further show that the ratings of all four subsets can be accurately fitted using a Gaussian distribution.

4.3 Dependency on distortion classes

The results from the previous section indicate that the sequences with distortions in the salient region generally receive lower ratings as compared to the sequences with distortions in the non-salient region. Further evidence of this observation is given by the MOS computed for each of the five subsets and averaged over all 20 different contents, which is presented in Table 2. It can be seen that, naturally, SEQ_R received the highest MOS, followed by $\text{SEQ}_{N,0.4}$, $\text{SEQ}_{N,1.2}$, $\text{SEQ}_{S,0.4}$, and $\text{SEQ}_{S,1.2}$. Thus, as an average over a large number of different contents, the distortions in the non-salient region were perceived by far less annoying as compared to the salient region. It is particularly worth noting that $\text{MOS}_{N,1.2}$ received an average MOS that is 1.02 higher than $\text{MOS}_{S,0.4}$, even though the distortion in the non-salient region is three times longer than the distortion in the salient region.

On the other hand, the distortion duration seems to play only a minor role as compared to the saliency of the location. This is particularly true for distortions in the non-salient region, where the small difference of 0.22 between $\text{MOS}_{N,0.4}$ and $\text{MOS}_{N,1.2}$ indicates only little impact of distortion duration on perceived annoyance. The larger difference of 0.64 between $\text{MOS}_{S,0.4}$ and $\text{MOS}_{S,1.2}$ indicates that the duration plays a more prominent role in case of distortions appearing in the salient region.

4.4 Content dependency

To identify whether the hierarchy of MOS presented in Table 2 is valid for different content, the MOS for all 20 sequence contents in the 5 subsets are presented in Fig. 5. It can be observed that the hierarchy of MOS between the subsets is almost exclusively the same as in Table 2 for all video sequences. This is a strong indication that

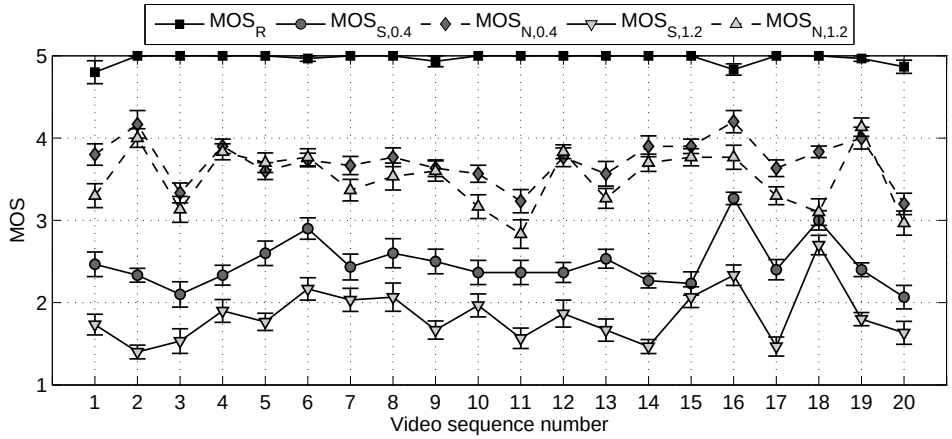


Figure 5. MOS and standard errors for all 20 contents of all sequence subsets (SEQ_R, $\mathrm{SEQ}_{S,0.4}$, $\mathrm{SEQ}_{N,0.4}$, $\mathrm{SEQ}_{S,1.2}$, $\mathrm{SEQ}_{N,1.2}$).

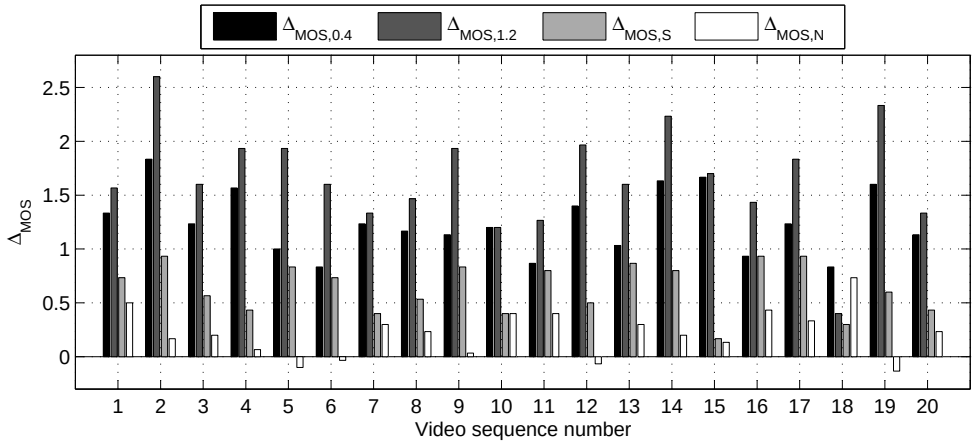


Figure 6. MOS differences for all 20 contents of the distorted sequence subsets ($\mathrm{SEQ}_{S,0.4}$, $\mathrm{SEQ}_{N,0.4}$, $\mathrm{SEQ}_{S,1.2}$, $\mathrm{SEQ}_{N,1.2}$).

the higher annoyance of distortions in the salient region as compared to the lower annoyance of distortions in the non-salient region is valid for a broad range of different video contents with strongly varying spatial and temporal characteristics (see Fig. 1).

The Δ_{MOS} presented in Fig. 6 reflect the difference in annoyance both with respect to distortion location and distortion duration. It can be seen, that for almost all sequences the difference in MOS is significantly larger for $\Delta_{\mathrm{MOS},0.4}$ and $\Delta_{\mathrm{MOS},1.2}$ as compared to $\Delta_{\mathrm{MOS},S}$ and $\Delta_{\mathrm{MOS},N}$. These results support the above observations that the observers distinguished annoyance levels more pronounced with respect to the content saliency of the distorted region (salient or non-salient) as compared to distortion duration (long or short) and give further evidence that this is true for a large variety of different content. Figure 6 also shows that the difference between salient region and non-salient region is usually more pronounced in case of long distortions, $\Delta_{\mathrm{MOS},1.2}$, as compared to short distortions, $\Delta_{\mathrm{MOS},0.4}$. Similarly, the distinction between long and short distortions is observed to be more pronounced in the salient region, $\Delta_{\mathrm{MOS},S}$, as compared to the non-salient region, $\Delta_{\mathrm{MOS},N}$. In particular the small values of $\Delta_{\mathrm{MOS},N}$ indicate that the annoyance of distortions in the non-salient region varies only very little with respect to the duration. Similar observations were made on the MOS averaged over the whole subset and indeed, this appears to be true for a broad range of sequence contents.

The above observations are true for all sequences but one, sequence 18. This sequence contains a close up of a rugby game with extremely high motion over large parts of the frames, which is also apparent in the highest TI indicator out of all sequences (see also Fig. 1). Here the distinction between long and short distortions was in fact stronger in the non-salient region. This is thought to be due to stronger masking effects caused by the extremely high motion in the salient region as compared to the lower motion in the background. As such, the distortions in

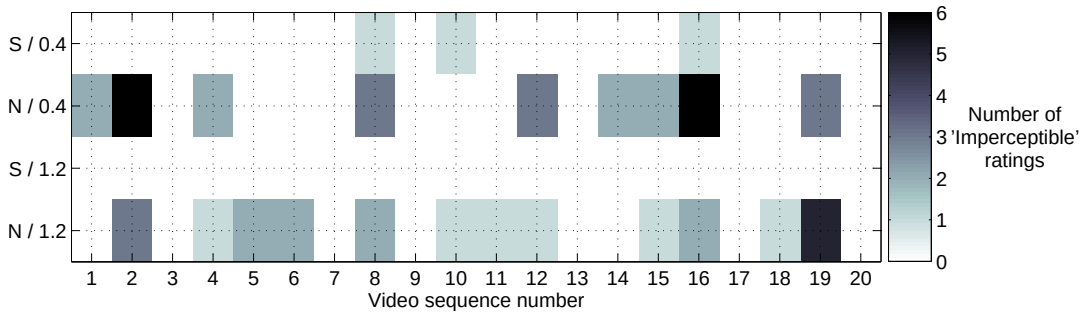


Figure 7. Number of 'Imperceptible' ratings for all 20 contents of the distorted sequence subsets ($SEQ_{S,0.4}$, $SEQ_{N,0.4}$, $SEQ_{S,1.2}$, $SEQ_{N,1.2}$).

the salient region were not perceived as severe, which also explains the fairly high MOS scores. This may also be the reason for the distinction between salient region and non-salient region being more pronounced for the short distortion duration, unlike for all other sequences, where it is more pronounced for the long distortion duration.

4.5 Detection of distortions

The 'Imperceptible' rating given in the impairment scale provides valuable information whether distortions have actually been detected by the observer or not. In this respect it is of interest to evaluate the degree of distortion detection in relation to the saliency of the distortion region, to the distortion duration, and to the video content. For this purpose, the number of 'Imperceptible' ratings (annoyance scores equal to 5) are visualised in Fig. 7 for all 20 video contents of the distorted sequence subsets.

It can be seen that many 'Imperceptible' ratings have been given for $SEQ_{N,0.4}$ (22 ratings) and $SEQ_{N,1.2}$ (29 ratings), whereas only few have been given for $SEQ_{S,0.4}$ (3 ratings) and in fact none for $SEQ_{S,1.2}$. This is thought to be due to mainly two reasons. Firstly, as the attention is usually on the salient region, the observer is more likely to miss distortions in non-salient regions. Secondly, salient regions typically exhibit features that facilitate stronger visualisation of distortions, such as high local contrast, and distinguished shapes and colours. Non-salient regions often are composed of image parts that are more uniform, such as a sky or a water surface.

It can be further observed from Fig. 7 that for sequences 2, 16, and 19 there was a particularly high number of 'Imperceptible' ratings in the non-salient distorted sequences. These three sequences exhibit fairly uniform non-salient regions and in addition, the attention of the observers is strongly focused on the salient regions in all three sequences, as indicated by the heat maps in Fig. 2.

5. CONCLUSIONS

In this paper, the results of a subjective experiment were presented that we conducted to identify the impact of visual saliency on the annoyance of packet loss distortions in H.264/AVC coded videos. For this purpose, the saliency in a number of different sequences was identified using eye tracking data. Loss patterns of different lengths were then induced in salient regions and in non-salient regions. A second experiment was then conducted in which human observers rated the annoyance of the distortions. The quantitative results show strong indications that distortions in salient regions are significantly more annoying as compared to distortions in the non-salient regions. These findings were consistent over a broad range of different video content.

The findings from this work strongly indicate that visual saliency should not be neglected in objective video quality assessment in the context of transmission errors. Thus, we will continue in this line of work by incorporating models of visual saliency into video quality metrics with the goal to improve the agreement of their quality prediction with subjectively perceived visual quality.

ACKNOWLEDGMENTS

This work has been supported by strategic funding for internationalisation from the Blekinge Institute of Technology, Sweden, and by the National French project FUI SVC4QoE.

REFERENCES

[1] International Telecommunication Union, "Advanced video coding for generic audiovisual services," Rec. H.264, ITU-T (Nov. 2007).

[2] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A., "Overview of the H.264/AVC video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology* **13**, 560–576 (July 2003).

[3] Kanumuri, S., Cosman, P. C., Reibman, A. R., and Vaishampayan, V. A., "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. on Multimedia* **8**, 341–355 (Apr. 2006).

[4] Liu, T., Wang, Y., Boyce, J. M., Yang, H., and Wu, Z., "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *IEEE Journal of Selected Topics in Signal Processing* **3**, 280–293 (Apr. 2009).

[5] Yang, K., Guest, C. C., El-Maleh, K., , and Das, P. K., "Perceptual temporal quality metric for compressed video," *IEEE Trans. on Multimedia* **9**, 15281535 (Nov. 2007).

[6] Pastrana-Vidal, R. R., Gicquel, J. C., Colomes, C., and Cherifi, H., "Sporadic frame dropping impact on quality perception," in [*Proc. of IS&T/SPIE Human Vision and Electronic Imaging*], **5292**, 182–193 (Jan. 2004).

[7] Claypool, M. and Tanner, J., "The effects of jitter on the perceptual quality of video," in [*Proc. of ACM Int. Conf. on Multimedia*], 115–118 (Oct. 1999).

[8] Wandell, B. A., [*Foundations of Vision*], Sinauer Associates, Inc. (1995).

[9] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20**, 1254–1259 (Nov. 1998).

[10] Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D., "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **28**, 802–817 (May 2006).

[11] Engelke, U. and Zepernick, H.-J., "A framework for optimal region-of-interest based quality assessment in wireless imaging," *Journal of Electronic Imaging, Special Section on Image Quality* **19** (Jan. 2010).

[12] Liu, H. and Heynderickx, I., "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).

[13] Vu, C. T., Larson, E. C., and Chandler, D. M., "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in [*Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation*], 73–76 (Mar. 2008).

[14] Ninassi, A., Meur, O. L., Le Callet, P., and Barba, D., "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in [*Proc. of IEEE Int. Conf. on Image Processing*], **2**, 169–172 (Oct. 2007).

[15] You, J., Perkis, A., Hannuksela, M., and Gabbouj, M., "Perceptual quality assessment based on visual attention analysis," in [*Proc. of ACM Int. Conference on Multimedia*], 561–564 (Oct. 2009).

[16] Feng, X., Liu, T., Yang, D., and Wang, Y., "Saliency based objective quality assessment of decoded video affected by packet losses," in [*Proc. of IEEE Int. Conf. on Image Processing*], 2560–2563 (Oct. 2008).

[17] Video Quality Experts Group, "VQEG FTP file server." ftp://vqeg.its.bldrdoc.gov/ (2009).

[18] Boulos, F., Chen, W., Parrein, B., and Le Callet, P., "A new H.264/AVC error resilience model based on regions of interest," in [*Proc. of Int. Packet Video Workshop*], (May 2009).

[19] Cambridge Research Systems, "Tools for vision science." http://www.crsltd.com/catalog/eye-trackers/index.html (2009).

[20] International Telecommunication Union, "Subjective video quality assessment methods for multimedia applications," Rec. P.910, ITU-T (Sept. 1999).

[21] Heinrich Hertz Institute Berlin, "H.264/AVC reference software JM 16.1." http://iphome.hhi.de/suehring/tml/ (2009).

[22] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, "SVC/AVC Loss Simulator." http://wftp3.itu.int/av-arch/jvt-site/2005_10_Nice/ (2005).

[23] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures," Rec. BT.500-11, ITU-R (2002).