This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Quantification of error perception for full-face digital video

PLEASE CITE THE PUBLISHED VERSION

<http://dx.doi.org/10.1049/el:20050898>

PUBLISHER

© Institute of Electrical and Electronics Engineers (IEEE)

VERSION

AM (Accepted Manuscript)

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Rimell, Andrew N., Hina Keval, Neil J. Mansfield, and D. Hands. 2019. "Quantification of Error Perception for Full-face Digital Video". figshare. <https://hdl.handle.net/2134/12720>.

This item was submitted to Loughborough's Institutional Repository (<https://dspace.lboro.ac.uk/>) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

Quantification of error perception for fullface digital video

A.N. Rimell, H. Keval, N.J. Mansfield and D. Hands

A.N. Rimell, H. Keval and N.J. Mansfield (Department of Human Sciences, Loughborough University, Loughborough LE11 3TU, United Kingdom)

D. Hands (BT Exact, Adastral Park, Ipswich IP5 3RE, United Kingdom)

A subjective quality test was undertaken to quantify the amount by which the perceived quality of full-face digital video (such as a newsreader) varies with the spatial location of degradation. It was found that perceived error was greatest for degradation occurring around the eyes and mouth. Traditional digital video assessment techniques, such as peak signal-to-noise ratio, rate all errors equally, irrespective of spatial location and thus do not give a meaningful measure of perceived quality.

Introduction:

To design effective digital electronic broadcast and communication systems, it is necessary to understand how the user interacts with the technology involved. Perceptible errors may occur in the visual data stream in practical systems. However, while the errors can be identified and measured objectively in engineering terms, characterisation of subjective quality opinion presents a challenge.

It is important that the subjective opinion of individual users is understood when assessing degraded video quality, so that a reliable measure of perceived quality of service can be achieved. Traditional objective methods of assessing picture quality include the peak signal-to-noise ratio (PSNR).

This Letter describes an experiment where areas of a talking-head digital video image (full-face, such as a newsreader) were degraded at different spatial locations and participants were asked for their subjective opinion of the error. The aim of the study was, first, to quantify the degree by which error perception varies with spatial location and, secondly, to demonstrate the weakness of methods such as PSNR.

Experimental method:

Participants were shown a series of 6 s talkinghead clips with different areas of the image degraded by 8_8 pixellation; in each case the number of pixels degraded remained constant. The participants were then asked to subjectively rate the annoyance of each clip using a standardised scale [1].

Thirty-two participants (50% male and 50% female) were recruited from Loughborough University. They were all screened for visual acuity and normal colour vision. All participants were aged 20 years or above. The number of participants was defined by the use of a balanced Graeco-Latin square experimental design with 32 conditions.

Participants were all non-experts, and were not employed in the field of television picture quality assessment. The independent variable used in the experiment was the spatial location of the video degradation. After running a pilot study, seven degradation locations were selected: mouth, eyes, background, chest, forehead, neck, shoulder and control condition (with no error). The dependent variable used in the experiment was a measure of the perceived annoyance of the degraded talking-head video clip. This was achieved using the Degradation Category Rating (DCR) quality scale proposed in the Single-Stimulus Continuous Quality Evaluation (SSCQE) test method defined in ITU-R BT.500 [1]: 'Imperceptible', 'Perceptible, but not annoying', 'Slightly annoying', 'Annoying' and 'Very annoying'. The five ratings were allocated numerical values to enable a Mean Opinion Score (MOS) to be calculated.

The experiment used 16 female and 16 male talking-head prerecorded video clips without audio. Each video clip comprised a 6 s talking-head speaking a pair of phonetically balanced sentences.

After an initial colour blind and visual acuity test, participants were seated in a acoustically isolated room with a low ambient lighting level. They were seated four screen heights from the television in accordance with the relevant ITU Recommendation [1].

Before starting the test, participants were shown the subjective rating scale and were given two example degraded video sequences. Each participant viewed a single video clip and gave their subjective response before moving on to the next clip. The video clips were played from the computer in a sequence predetermined by the Graeco-Latin square experimental design.

Results:

The results are summarised in Fig. 1, which shows the MOS for each spatial error location. From Fig. 1 it can be seen that errors around the eyes and mouth were considered to be more annoying than those in the chest or background. The Figure also shows the calculated PSNR scores. The PSNR values are expressed as a ratio in dB, and have been suitably scaled for presentation in the Figure. The calculated scores vary a little as the area degraded for each facial location varied slightly because of the different geometries of the different facial areas.

Two main categories of spatial error location were considered: facial and non-facial. Facial areas include the eyes and mouth and non-facial areas include all other error locations.

For each participant, the measured scores were collated into the facial and non-facial categories. The MOS values were calculated for each category and are shown in Fig. 2. It can be clearly seen that every one of the 31 participants analysed showed a greater sensitivity to error at the facial locations than at the non-facial locations.

A Wilcoxon matched-pairs signed-ranks test (a non-parametric statistical test) was conducted to test the experimental hypothesis presented above. This test was chosen to examine the significance of an annoyance effect on facially degraded talking-head video clips over non-facially degraded talking-head video clips. From the data analysis, it was found that there was a significant difference between the facial and non-facial conditions ($p < 0.05$), i.e. the probability of the observed difference occurring by chance between the facial and non-facial regions is less than 5%. Therefore,

the hypothesis that the perceived annoyance ratings are consistently higher in facial degradations compared to non-facial degradations was accepted.

Discussion:

The results from the subjective video quality assessments showed that facial spatial location of video coding error on talkinghead sequences does produce greater levels of perceived annoyance than non-facial spatial location of video coding error. The difference between the perceived non-facial and facial error degradations was 2.32 MOS points. The results from this study support the proposed hypothesis.

Previous work in psychophysics supports the findings presented here. Zusne [2] used fixation patterns from users to understand where the user focuses when viewing various face and object pictures. Zusne discovered from these patterns that the eye receives the most attention. Laughery et al. [3] ascertained that users ranked facial features in the following order of importance: eyes; nose; mouth; lips=chin; hair; and finally the ears. The majority of facial perception studies support the theory that the face is the most viewed area of a head-shoulder image; furthermore, past studies have concluded that individuals are most likely to focus their attention on the areas around the eyes and the mouth. Kahneman [4] linked these findings to human memory, suggesting that spontaneous looking is guided by stored knowledge and when viewing faces, the eyes, lips and nose are given the most attention because the observer is aware that the human eyes and lips are the most mobile and expressive elements of the face.

Research by Calvert et al. [5], Haxby et al. [6] and Perrett et al. [7] found that the superior temporal sulcus is stimulated when individuals view faces but not when viewing objects such as cars and butterflies. It has been propounded that this area of the brain is concerned with the details of facial structure, such as eye gaze, expression and lip movement.

This study accounted for participant bias, as participants were all screened for visual acuity and colour-blindness and it was ensured that all those taking part in the study were non-experts of audio-visual systems. The use of a standardised scale and recommended procedures for this study ensured that the experiments were conducted reliably and could be compared with results from other experimenters in the field.

In this study, participants were instructed to focus their attention solely on the video image. Experiments conducted by Rimell and Owen [8] showed that, when attention is focused on one modality, participants will be more critical of errors in that modality than when in a relaxed environment with other sources of stimulation (such as an audio stream). Therefore, it is suggested that the results presented here form a worst-case scenario and if participants were engaged in a task other than assessing the video quality, they would be slightly less sensitive to errors. It should be noted that, in modelling subjective quality assessment, task and attention are important issues to be taken into account (Hollier et al. [9]) and this is an area of research which has yet to be exhausted.

Conclusions:

This study proposed that facial spatial location (eyes and mouth) of video coding error in talking-head sequences would produce greater levels of perceived annoyance than for non-facial spatial location of video coding error. The experiment showed that participants found that errors located on the eyes or mouth were the most annoying, and errors located on the shoulder, chest or background

least annoying. Further anecdotal evidence from participants' comments also supported the measured results. The majority of participants who expressed an opinion stated that degradation in the areas of the eyes and mouth were the most annoying because it was difficult to maintain eye contact and lip-read.

These findings have practical applications in a number of areas, including the design of automated multimedia Perceived Quality of Service (PQoS) measuring equipment and video coder=decoder (codec) design. A video codec algorithm could be designed to compress different parts of the image at different bit rates while keeping the overall bit rate constant, thus enabling the eyes and mouth to be coded at a higher resolution than, for example, the chest and shoulders, thereby improving the overall perceived quality of the image.

References

- 1 International Telecommunications Union (ITU), Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500, 2002.
- 2 Zusne, L.: 'Visual perception of form' (Academic Press, New York, 1970)
- 3 Laughery, K.R., Alexander, J.F., and Lane, A.B.: 'Recognition of human faces: effects of target exposure time, target position, pose position, and type of photograph', *J. Appl. Psychol.*, 1971, 55, pp. 477–483
- 4 Kahneman, D.: 'Attention and effort' (Prentice-Hall, New Jersey, 1973)
- 5 Calvert, G.A., et al.: 'Activation of auditory cortex during silent lipreading', *Science*, 1997, 276, (5312), pp. 593–596
- 6 Haxby, J.V., Hoffman, E.A., and Gobbini, M.I.: 'The distributed human neural system for face perception', *Trends Cogn. Sci.*, 2000, 4, (6), pp. 223–233
- 7 Perrett, D.I., et al.: 'Visual cells in the temporal cortex sensitive to face view and gaze direction', *Proc. R. Soc. Lond. B, Biol. Sci.*, 1985, 223, (1232), pp. 293–317
- 8 Rimell, A., and Owen, A.: 'The effect of focused attention on audiovisual quality perception with applications in multi-modal codec design'. *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP 2000)*, 2000, pp. 2377–2380
- 9 Hollier, M.P., et al.: 'Multi-modal perception', *BT Technol. J.*, 1999, 17, (1), pp. 35–46

Figure 1: The PSNR and MOS results from the experiment described in this paper. The experimental results show that errors occurring at and around the eyes are considered to be the most annoying. The PSNR results are an objective measurement of the quality of each degradation location. The calculated values of PSNR vary slightly as areas of degradation were not exactly the same size - for example, the area of the eyes was slightly different to the area of the mouth. The PSNR measurements ranged from 61.987 dB (eyes) to 63.181 dB (neck).

Figure 2: Comparison of the facial and non-facial degraded MOS areas for each participant. It can be clearly seen that degradation in the facial area (the eyes and mouth) was consistently considered more annoying than non-facial areas. In an actual television or video conferencing system, errors in the non-facial regions may be tolerated, but it is highly unlikely that the perceived poor quality within the facial regions would be considered acceptable by the users.