Research and Application of the Adaptive Model of the Human Visual System for Improving the Effectiveness of Objective Video Quality Metrics

Vladimir Mazin
Moscow Technical University of Communications and Informatics
Moscow, Russia
vladimir.a.mazin@gmail.com

Michael J. Cree
The University of Waikato
Hamilton, New Zealand
michael.cree@waikato.ac.nz

Lee Streeter
The University of Waikato
Hamilton, New Zealand
streeter@waikato.ac.nz

Ksenia Nezhivleva
Moscow Technical University of Communications and Informatics
Moscow, Russia
nezhivlevaksenia@gmail.ru

Anastasia Mozhaeva
The University of Waikato
Hamilton, New Zealand
Moscow Technical University of Communications and Informatics
Moscow, Russia
am476@students.waikato.ac.nz

***Abstract*—Video traffic from content delivery networks occupied 82% of all consumed bandwidth in 2022. Nevertheless, the available bandwidth is sometimes volatile and limited. Adaptive video streaming or, in other words, prediction of quality is the key to increasing throughput and reducing storage. Unfortunately, while developing video quality metrics, a problem exists in the algorithmic representation of the human visual system, such as the cognitive component, namely the delay of human reaction to artifacts, which is not represented in the current works. The presented new methodology of data collection of the delay of the human visual system response to video artifacts in modern terms of providing information in natural conditions is presented. New knowledge of the human visual system adaptation or other words time of reaction of perception of artefacts, including the response to motion perceptions necessary for correct work of video quality assessments, is presented and tested. The proposed work introduced that the use of new data on the human visual system adaptation gives an improvement in the performance of video quality assessment metrics.**

I. INTRODUCTION

Video quality metrics are a major feature of modern streaming video processing algorithms. Unfortunately, not many video quality metrics predict well the subjective perception of quality by the human visual system (HVS). However, the knowledge of HVS's working represents to developers the opportunity to increase throughput and reduce storage. While developing video quality metrics, a problem exists in the algorithmic representation of HVS. Algorithmic modelling of HVS is possible based on data sets obtained from research in the field of visual psychophysics. Collecting new data on the performance of HVS provides a better understanding of HVS by linking changes in the physical attributes of a visual stimulus to corresponding changes in psychological responses (visual perception and cognition). The studies typically involve carefully constructed human experiments using tightly controlled visual stimuli and viewing conditions.

Many of the most fundamental properties of visual perception are used to create predictors of video quality assessments (VQA). However, the main purpose of the vast majority of research in visual psychophysics is to gain knowledge about how HVS works, when the object of research is extracted from the natural environment, specifically the use of a head holder, and medical eye drops. Using the data, obtained in psychophysical studies, is difficult for the developers of video quality assessment algorithms [1]. In other words, in developing a more complete computational model of HVS, video quality scientists face the problem of visual neurons, often responding quite differently to natural stimuli than to simply controlled stimuli [2].

Another problem developers encounter, when incorporating psychophysical data into algorithms, is for evaluating distortions to be either compound or over-threshold during the experiment. The term compound is used to describe a visual target that stimulates more than one channel in multichannel HVS analysis. In other words, two stages of early vision [3]. A small number of modern video quality predictors using HVS models [4] [5] work quite well, and use HVS models for the initial part of the early vision [3]. Specifically, the filtering stage, which contains the interpretation. In other words, the attempt to reconstruct the response of the HVS to the spatial and temporal components of the video sequences, as well as the brightness and eccentricity (distance from the centre of the fovea in visual degrees). Unfortunately, the second part of the early vision, which contains the cognitive component, specifically, the response of HVS to artifacts is omitted in the current works.

The accurate time of reaction of human perception to artefacts in video, including the perception of motion, is necessary for the correct imitation of HVS. The primary function of the mediovisual complex has been precisely established to analyze the direction and velocity of the object's movement in the visual world [6]. In studies in psychophysics, the potentials evoked by images were proven to be significantly different from the potentials generated by videos: 60 to 320 ms after the stimulus onset and 120 to 400 ms [7]. Currently, the most commonly used objective quality metrics do not use data on the cognitive delay of users when evaluating video. However, when testing and evaluating the video quality metrics, user perception datasets are used. Such sets usually consist of video sequences with frame-by-frame user evaluation. The absence of information about the cognitive delay of users in the video quality metrics leads to an underestimation of the estimates and an underestimation of the correlation between the subjective data of users and the algorithmically obtained metrics. This is due to the fact that the objective video quality metrics give an instant response to the appearance of artefacts, while the user needs time to evaluate what is presented on the screen [8].

The present work hypothesizes that by including in the work of current video quality metrics the results of the HVS time response to artefacts, the percentage of accuracy of the predictors will increase by at least 5%.

The purpose of the proposed work is to demonstrate the necessity of introducing new data about the work of HVS into the existing metrics of video quality assessment. The proposed work represents the creation of a module of time adaptation of HVS to artifacts, including the response to motion perceptions required to correctly simulate the work of HVS.

II. Related work

Numerous image and video evaluation methods based on HVS models have been developed [9] [10] [11] [12]. Images are usually processed, using a set of spatial filters to produce signal processing-oriented spatial-frequency decompositions of images designed to simulate initially linear neuronal responses. The quality of the distorted image is evaluated based on the degree to which the adjusted responses to the reference image differ from the adjusted responses to the distorted image. Many HVS-based methods were originally designed to work as predictors of visible image differences; in other words, were developed to determine whether changes are visible. Consequently, the methods work best, when distorted images contain artifacts closest to the detection threshold.

Currently, algorithmic metrics for evaluating video quality, based on the interaction of spatial and temporal perception of the user, are also relevant. Well-known examples of metrics that take into account temporal aspects are STRRED [13] and HDR-VQM [14]. STRRED - Spatiotemporal Entropy Reference Difference, a metric estimates quality degradation by calculating the entropy of the difference between the reference and distorted video sequences. Entropy is calculated from the distribution of wavelet coefficients. The difference in entropy is evaluated in non-overlapping blocks, separately for the spatial and temporal aspects. Spatial entropy differences are calculated using a spatial multi-scale multi-orientation decomposition of each frame in the video sequence. STRRED does not require large computing resources and uses much less random information to transmit. The disadvantages of STRRED are that the method does not show a linear relationship with subjective perception, does not capture granular effects, and does not integrate multi-scale information. HDR-VQM, which is one of the most popular metrics for HDR video, takes physical calibration into account. The metric splits the video into spatial bands, which are then split into spatio-temporal "streams".

The scientific community has repeatedly argued that the basic models of HVS need to be extended to better account for the properties of human vision [1]. The current understanding of the near-threshold vision for controlled stimuli is relatively mature in terms of modeling. However, much less is known about how HVS works, when distortions are more complex and in the suprathreshold mode (which may involve areas of the visual cortex). Nevertheless, recent HVS-based methods have begun to use enhanced and/or mid- to high-level visual models [15] and many of the presented methods have been shown to be extremely effective for evaluation. However, when creating VQAs, predictors based on HVS models are still in the initial stage of development. Several video estimation metrics based on psychophysical models exist, such as the contrast sensitivity function [5]. The predictors of video quality assessment work quite well, but the HVS models do not consider all the necessary variations in stimuli and aspects of the initial vision. Several visual HVS models have been developed, containing a comprehensive description of spatial and temporal contrast sensitivity, and a dependence of such sensitivity on retinal illumination [8]. However, for modeling the current HVS models underlying predictors of video estimation, as described earlier, most researchers use data presented in psychophysiological studies; only for the filtering stage, without consideration of the cognitive component.

In our previous work, we presented the PSNR-M+ video quality metric, which includes the HVS time response to artefacts [8]. PSNR-M+ considers the data for the first part of early vision, specifically the filtering stage, which determines, which spatial and temporal fluctuations in stimuli the HVS responds to. The advantage of the given metric is for the PSNR-M+ to be created under the current conditions of information provision. Our previous work demonstrates that the metric based on psychophysical HVS models including the HVS time response to artefacts explains the human perception of video quality, outperforming statistically based metrics.

III. HVS time adaptation method

The new methodology for collecting data on HVS adaptation is the use of delay analysis of user reactions to the appearance of artifacts in videos. This work created a low-pass filter block for human visual system functions based on new subjective data about the user's response rate to artifacts.

Through the Moscow Technical University of Communications and Informatics, 30 observers between the ages of 18 and 36 with normal vision were recruited. In the current work, normal vision is defined by the typical participant who does not use glasses, lenses or other medical devices to correct vision in their normal daily activities. Most participants have no experience with human perception of visual information. Informed consent was obtained from all participants. The

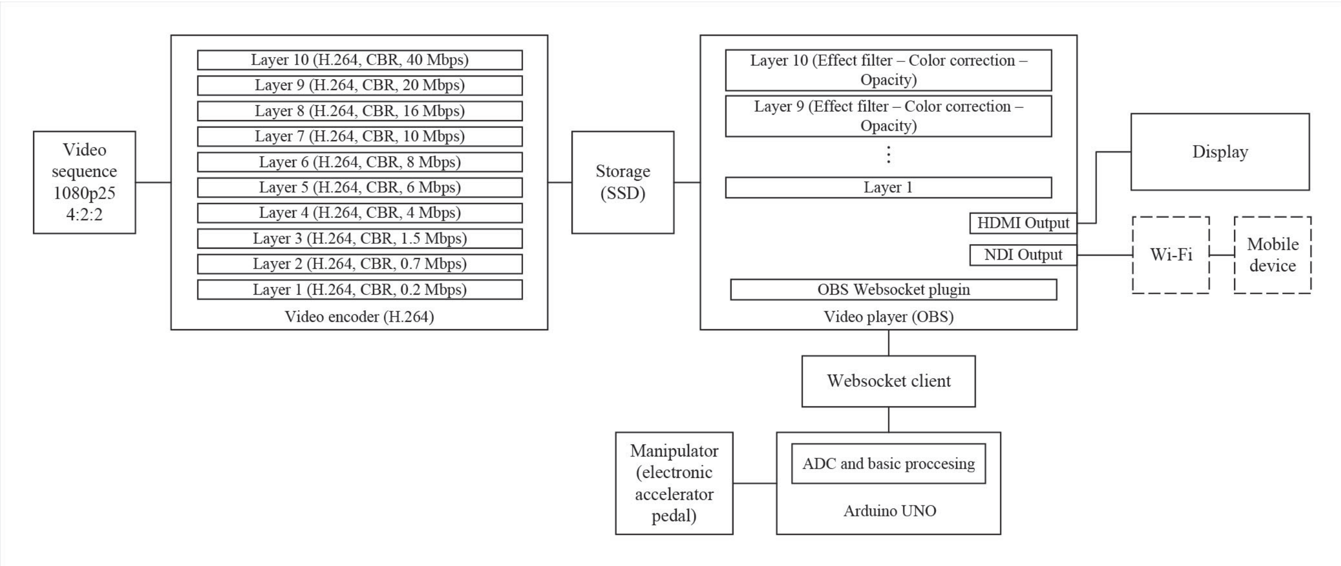

Fig. 1. Scheme of the method for finding an acceptable minimum perceptual threshold

subjective testing was performed on the equipment introduced in [16]. 15 uncompressed video clips have been used as the reference video. For one reference video, was been created 9 versions with compression artifacts (H.264) with different qualities. Participants viewed a video sequence of 15 video clips of 15 seconds each, separated by a grey background of 3 seconds [17]. Participants measured the quality of the encoded video by finding an acceptable minimum perceptual threshold. When evaluating the video, participants used a manipulator to set a minimum acceptable threshold of video perception. Where the minimum acceptable threshold indicates the threshold below which the video quality does not satisfy the users and therefore is not acceptable for transmission. Scheme of the method for finding an acceptable minimum perceptual threshold is shown in Fig. 1. In order to avoid a step change in quality with a limited number of levels (10), the manipulator set intermediate levels of quality by the programming of mixing neighbouring levels of quality in a proportion determined by the pressing force. All videos are presented in YUV422 progressive format with a fixed resolution. 1920×1080 is currently the most popular. The frame rate is 25 fps. Video clips for the database were professionally recorded in uncompressed digital format, which makes the distortions in the video possible. To ensure a constant quality of perception frame by frame, a two-pass coding scheme was used for subjective real estimation.

The display is a 22-inch flat screen. The display provides proper brightness, color adjustment, and calibration with a professional exposure meter. A curved display can also be used in the experiment since the curvature of the display surface can be ignored because the "working" area of the stimulus is small. We measure the stimulus in the middle, the edges are used so that there is no stimulus-background transient. The brightness is 200 cd/m^2 and the white color temperature is calibrated to standard D65. According to the knowledge of the HVS and the field of clarity, visual acuity is 1/60th of a degree [18]. The minimum allowed distance from this monitor to the participant to find the stimulus in the point of fixation of vision is 0.872 m, and the maximum is 1.149 m [19]. The color of the inactive screen should be light grey.

When viewing a grey background between videos, participants set the manipulator to a position of satisfaction according to instructions. We investigated users' reaction times to the appearance of artifacts, and scene changes; in other words, the adaptation time of the human visual system. The experiment finished when the experimental uncertainty, as measured by the confidence interval, became less than 5% of the current value for all tests performed during the experiment. Among all the tests, the maximum variation of delay values is achieved in the 11655th frame: from 662 to 724 milliseconds (this happens due to camera movement and the large number of objects in the frame immediately after the scene change), the minimum - at 7777-th frame: from 720.2 to 720.3 milliseconds. On average, the scatter of values is 0.0058 milliseconds. The example of the spread of results for scene changing is shown in Fig. 2.

More than 700 HVS adaptation thresholds were obtained. Analyses of the HVS adaptation thresholds resulted in an average response delay of 766.7 ms. Based on this delay, a Gaussian filter was created, which, due to its properties, is best suited for modelling the HVS delay [20]. The filter coefficients were obtained using formula (1).

$$w(n) = e^{-\frac{1}{2}\left(\alpha \frac{n}{(L-1)/2}\right)^2} = e^{-\frac{n^2}{2\sigma^2}}, \quad (1)$$

where

$$-\frac{L-1}{2} \leq n \leq \frac{L-1}{2}, \quad (2)$$

where L is window length, n is window function argument, α is inversely proportional to the standard deviation, σ, of the Gaussian random variable, $\sigma = (L-1)/(2\alpha)$. After processing with the filter described above, the vector of video quality metric values goes to the adaptation filtering block. The block scheme is shown in Fig. 3. VQAs require 1 second of the

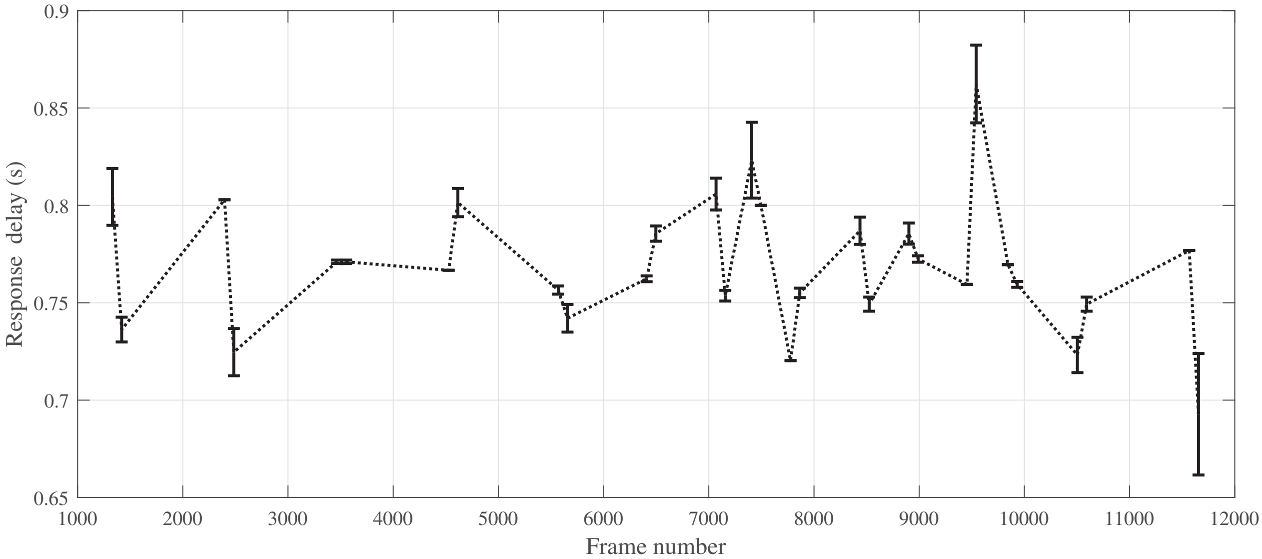

Fig. 2. The deviation of results of the subjective evaluation for scene changing.

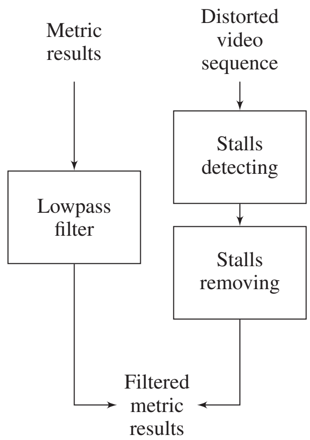

Fig. 3. The HVS adaptation model.

video sequence to stabilize weight values, which is why the video quality metric adaptation filtering block was created [8]. In a given block, the quality measure values at the time of adaptation are replaced by the mean value of the next two video frames after adaptation(3):

$$P(n) = \frac{P(T+1) + P(T+2)}{2}, n \leq T \tag{3}$$

$P(n)$ is the video quality metric value in the n-th frame, T is the delay in the video quality metric adaptation.

Also, the delay in HVS response to sudden scene change (hard cut) during the viewing of content is present. VQAs, such as PSNR, due to the principle of operation, react to a scene change by an abrupt change in the metric value. We named it scene's stalls because the metric's score changes instantly, but a person needs time to change the quality score after a scene change [18]. In practice, we get stalls in human perception. The scene cuts detecting (stalls detecting) and scene cuts removing blocks (stalls removing) were created to process the scene change. If several scene changes are presented at intervals shorter than the HVS delay, such changes will be smoothed out, because the HVS will not have time to recognize the change in quality. Another aspect, which is not considered in the VQA, is the compression errors leading to fading frames. In experiments, was found the user smoothly lowers the score while watching the content. Then the subjective score evens out, as the HVS receives no new information. VQAs, in contrast, are arranged in such a way in the occurrence of such artefacts is accompanied by a sharp drop in the score. The value of the VQA score changing is much more strongly concerning the change in subjective score. Stalls detecting and removing blocks were created to simulate the response of HVS. The Stalls detecting block implements recognition using the Histogram Differences method. The adaptive threshold for this method is set to 80% of the maximum histogram difference. Recognition of smooth scene changes is a part of our future work. In the scene cuts detecting (stalls detecting) block, the shift of metric values at scene change is implemented according to formulas (4), (5).

$$P(C+T+n) = P(C+T+n) + (P(C+T-1) - P(C+T)), \tag{4}$$

$$P(n) = P(n-T).n \in [C; C+T), \tag{5}$$

where C is the scene change frame number, T is a delay in HVS adaptation, $P(n)$ is metric value in the n-th frame.

IV. Results and Discussion

The most popular methods of video quality assessment were chosen for analysis. The peak signal-to-noise ratio (PSNR) is calculated on a logarithmic scale by amplitude

TABLE I. THE NUMERICAL DATA WITH THE AVERAGE INCREASE IN THE PERCENTAGE OF THE METRICS ON THE LIVE-NFLX DATABASE

VQA	PSNR	FVVDP	VMAF	HDR-VQM
Without adaptation module	0.282	0.240	0.258	0.245
Using adaptation module	0.304	0.297	0.318	0.275

TABLE II. THE NUMERICAL DATA OF THE METRICS ON THE LIVE-NFLX DATABASE WITHOUT ADAPTATION MODULE

Video/VQA	PSNR	FVVDP	VMAF	HDR-VQM
Content 4, Sequence 0	0,175	-0,277	0,070	0,119
Content 4, Sequence 2	0,117	0,112	0,169	0,090
Content 4, Sequence 4	0,509	0,464	0,562	0,505
Content 4, Sequence 7	0,580	0,560	0,707	0,557
Content 5, Sequence 0	0,099	0,253	0,155	-0,073
Content 5, Sequence 2	-0,500	-0,092	-0,131	-0,504
Content 5, Sequence 4	-0,470	0,130	0,124	-0,231
Content 5, Sequence 7	-0,266	0,230	0,270	-0,030
Content 6, Sequence 0	0,282	0,333	-0,027	0,445
Content 6, Sequence 2	-0,284	-0,095	-0,164	-0,178
Content 6, Sequence 4	0,077	0,224	0,427	0,016
Content 6, Sequence 7	0,020	0,114	0,285	-0,187

TABLE III. THE NUMERICAL DATA OF THE METRICS ON THE LIVE-NFLX DATABASE WITH ADAPTATION MODULE

Video/VQA	PSNR	FVVDP	VMAF	HDR-VQM
Content 4, Sequence 0	0,227	-0,36	0,117	0,174
Content 4, Sequence 2	0,149	0,086	0,225	0,124
Content 4, Sequence 4	0,625	0,571	0,686	0,65
Content 4, Sequence 7	0,644	0,635	0,779	0,655
Content 5, Sequence 0	0,059	0,282	0,214	-0,123
Content 5, Sequence 2	-0,488	0,033	-0,061	-0,45
Content 5, Sequence 4	-0,42	0,247	0,242	-0,164
Content 5, Sequence 7	-0,208	0,353	0,395	0,047
Content 6, Sequence 0	0,298	0,376	-0,021	0,47
Content 6, Sequence 2	-0,286	-0,095	-0,172	-0,19
Content 6, Sequence 4	0,163	0,34	0,537	0,086
Content 6, Sequence 7	0,081	0,188	0,366	-0,165

(in decibels), which is an advantage. However, PSNR correlates poorly with visual quality assessment and does not consider psychovisual patterns [21]. The structural similarity metric [22] for images with fragments of large or small mean brightness values gives unstable results and correlates poorly with human perception. Video multimethod assessment fusion (VMAF) [23] is based on machine learning and uses databases, which use the quality scores of real users of training videos. HDR-VQM considers the temporal aspects of VMAF [14]. Evaluations, simulating the performance of the HVS, are also included for analysis. FovVideoVDP is built on the HVS model, which considers the peripheral domain. The predictor, however, does not consider modern video screens.

Below shows the results of the performance of the above metrics on the LIVE-NFLX database [24] [25]. LIVE-NFLX consists of 112 distorted videos; the distorted videos were generated with compression errors. The LIVE-NFLX database was chosen because it represents very realistic content with Quality of Experience responses to various design parameters.

In this work, we used the contents of the LIVE-NFLX data set, which are presented in Fig. 4 with the compression artefacts : Sequences with the number 0 consist of a constant encoding bitrate of 500 kbps. Sequences with the number 2 consist of an encoding bitrate of 500 kbps and include a single video segment of 160 kbps. Sequences with the number 3 consist of one video segment encoded at 250 kbps followed by a 66 kbps segment, followed by another 250 kbps segment. Sequences with the number 4 consist of one video segment at 250 kbps followed by a segment at 100 kbps and then another segment encoded at 250 kbps. This pattern may be the least practical among all the considered playout patterns. However, it is of interest to be able to study the subjective data resulting from such an “ideal” client reaction.

The numerical data with the average increase in the percentage of the metrics are shown in Tables I, II and III for each video and the average for the database.

The poor of the Gaussian window in the first frames were represented by some ”very ability” of the filtered values, leading to an understatement estimation of the results for a number of videos. The analysis revealed that after the first frames, this effect is not observed in the rest of the video clips.

The research of human psychophysical reaction represented information on the response of HVS to photographs of objects implying motion (e.g., an athlete running or a cup falling from a table) gives a slower physiological response than to photographs without implying motion (e.g., a sitting person or a cup on a table) [26]. The neural response to implied movement in areas of human motion is slower compared to the response to actual movement. Based on the above information, our future work will include the adaptation module should be integrated with the motion detection module. In addition, the reaction detection capabilities for static frames should be extended.

V. CONCLUSION

Models of the human visual system are used to create predictors of video quality assessments, which have proven superior to algorithmic methods correlating poorly with subjective human assessments. The proposed work demonstrates the need for metrics containing a cognitive component, specifically

Fig. 4. The video content of LIVE-NFLX database

accurate human perception of video, including motion perception, to correctly simulate performance. The use of the new data on the adaptation of human visual systems has been shown to give an improvement in the performance of video quality assessment metrics by more than 10%. We predict that such a model could reduce the amount of information transmitted in video streaming by 70%. However, the presented data of the human visual system do not consider the neural response to statics, and implied motion in human motion areas is slower compared to the response to real motion. These extensions are planned to be considered in our future work.

REFERENCES

[1] D. Chandler, T. Phan, M. Alam, "Seven challenges for image quality research, in: Examining the limitations of current research into image quality assessment opens doors for further studies". *SPIE*, 2015.

[2] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity", *Human Vision*, 1993.

[3] The Nation. Academ. Press Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A., "The analysis of visual motion: A comparison of neuronal and psychophysical performance". *Journal of Neuroscience*, vol. 12, pp. 4745–4765, 1992.

[4] A. Mozhaeva, I. Vlasuyk, A. Potashnikov, L. Streeter, "Full reference objective metric for assessing video quality compatible with psnr taking into account frequency and peripheral characteristics of human vision", *Digital Signal Processing Application Considerations* , vol. 2, pp. 44-54, 2021.

[5] R. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: a visible difference predictor for wide field-of-view video", *ACM Transactions on Graph* , vol.40, pp.1-19, 2021.

[6] S. Sunaert, P. Van Hecke, G., Marchal, & G. A. , Orban . "Motion-responsive regions of the human brain". *Experimental Brain Research*, vol.127, pp. 355–370, 1999.

[7] J.A. Lorteije, J. L. Kenemans, T. Jellema, R. H. van der Lubbe, F. de Heer, R. J. van Wezel, "Delayed response to animate implied motion in human motion processing areas". *J Cogn Neurosci*, vol. 18 (2), pp. 158 - 168, 2006. doi: 10.1162/089892906775783732. PMID: 16494678.

[8] A. Mozhaeva, V. Mazin, M. J. Cree, L. Streeter. (2023). "Video Quality Assessment Considering the Features of the Human Visual System." In: Yan, W.Q., Nguyen, M., Stommel, M. (eds) Image and Vision Computing. IVCNZ 2022. Lecture Notes in Computer Science, vol 13836. Springer, Cham.

[9] J. Lubin, "A visual discrimination model for imaging system design and evaluation"*Vision Models for Target Detection and Recognition*, E. Peli, Ed., pp. 245–283,World Scientific, 1995.

[10] S. Winkler, "Visual quality assessment using a contrast gain control model", *Proceedings of the IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 527–532, September 1999.

[11] D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images", *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[12] P. G. J. Barten, "Evaluation of subjective image quality with the squareroot integral method",*Journal of the Optical Society of America*, A, vol. 7, no. 102024, 2031 pages, 1990.

[13] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 4, pp. 684–694, 2012.

[14] M. Narwaria, M. P. Da Silva, P. Le Callet, "An objective quality measure for high dynamic range video.", *Sign. Proc.: Im. Comm.*, pp 46-60, 2015.

[15] R. Mantiuk, M. Ashraf, A. Chapiro, "stelaCSF - A Unified Model of Contrast Sensitivity as the Function of Spatio-Temporal Frequency, Eccentricity, Luminance and Area", *ACM Transactions on Graphics* (Proc. of SIGGRAPH 2022), vol: 41(4), article no. 145, 2022.

[16] A. Mozhaeva, A. Potashnikov, I. Vlasuyk and L. Streeter, "Constant Subjective Quality Database: The Research and Device of Generating Video Sequences of Constant Quality," 2021 International Conference on Engineering Management of Communication and Technology (EMCTECH), pp. 1-5, 2021.

[17] BT-500-11: "Methodology for the Subjective Assessment of the Quality of Television Pictures", 2012.

[18] K. Poynton, 'Digital Video and HD'. *Algorithms and Interfaces, Morgan Kaufmann Publishers*, 2012.

[19] A. I. Mozhaeva, I. V. Vlasuyk, A. M. Potashnikov, M. J. Cree and L. Streeter, "The Method and Devices for Research the Parameters of the Human Visual System to Video Quality Assessment," 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, pp. 1-5, 2021.

[20] A. Potashnikov, I. Vlasuyk, S. Romanov, "Method of measuring temporary and spatial characteristics of the human visual systems", *X International industrial scientific and technical conference: collection of works.* , pp. 102-103, 2016.

[21] K. Seshadrinathan, A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos", *IEEE transactions on image processing* , vol.19(2), pp.335-350, 2009.

[22] Z. Wang, A. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures", *IEEE Signal Processing Magazine*, vol.26(1), pp.98-117, 2009.

[23] L. Tsung-Jung, L. Yu-Chieh, L. Weisi and C.-C. Jay Kuo, "Visual quality assessment: recent developments, coding applications and future trends", *APSIPA Transactions on Signal and Information Processing* vol. 2, no. 1, pp. 20, 2013.

[24] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of Temporal Effects on Subjective Video Quality of Experience", *IEEE Trans. Image Process.*, vol. 26(11), pp. 5217– 5231, 2017.

[25] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron and A. C.Bovik, "LIVE Netflix Video Quality of Experience Database," [Online].Available://live.ece.utexas.edu/research/LIVE/NFLXStudy/index.htm

[26] Kourtzi, Z., & Kanwisher, N. (2000). "Activation in human MT/MST by static images with implied motion". *Journal of Cognitive Neuroscience*, vol. 12, pp. 48–55.