VIDEO QUALITY ENHANCEMENT VIA QP ADAPTATION BASED ON PERCEPTUAL CODING MAPS

*M. A. Papadopoulos, †Y. Rai, *A. V. Katsenou, *D. Agrafiotis, †P. Le Callet and *D. R. Bull

*Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB, UK
{ma.papadopoulos, angeliki.katsenou, d.agrafiotis, dave.bull}@bristol.ac.uk
† LUNAM University, University of Nantes, IRCCyN UMR CNRS 6597, Nantes Cedex 3, FR
{yashas.rai-kurlethimar, patrick.lecallet}@univ-nantes.fr

ABSTRACT

This paper introduces a method for adapting block quantisation parameter values in HEVC video compression based on perceptual coding maps. These maps are computed per block taking into account masking effects. Masking levels are calculated using spatial, temporal and foveation features that are extracted from the video and are stored in a perceptual coding map. The produced map drives a QP adaptation process that aims to redistribute coding bits in the frame so that the perceived quality is improved, especially at those bitrates where coding artifacts become visible (mid to high QP values). The subjective performance evaluation that was conducted showed that the proposed method can offer a measurable improvement in perceived quality relative to a constant QP approach, with Bjøntegaard mean opinion scores (MOS) gains reaching almost 9% for the test sequences used. The paper additionally highlights the need for further work in order to increase gains in perceived quality and optimise parameter selection.

Index Terms— Perceptual coding maps, QP variation, HEVC, subjective quality.

1. INTRODUCTION

Video compression performance has been improving with every new video coding standard, with the latest one, HEVC, offering significant improvements over its predecessor, AVC. Despite the fact that video coding standards already exploit certain characteristics of the human visual system (HVS) for achieving rate distortion performance gains, there is still room for further improvements in perceived quality by adapting the coding process to the characteristics of the content and how these relate to the HVS [1, 2]. In this paper, we investigate and propose a quantisation parameter (QP) adaptation process that is driven by a perceptual coding map which aims to exploit masking aspects of the HVS by assigning masking levels to blocks of pixels (Coding Tree Units - CTUs- in HEVC). In most scenarios, QP allocation within a coded frame is uniform, with any variation done mainly at the frame level or being the result of a rate control process that aims to restrict the output bitrate to within certain levels. Region of interest coding is one application where QP values are allocated on a perceptual basis, with the aim being to increase the perceived quality of the coded video, especially when the available bit budget is not sufficient for providing a good overall quality. This is usually done by applying a segmentation process, that typically separates foreground from background and redistributes the available bits from the latter to the former. Numerous such examples exist in the literature (e.g. [3, 4]). The requirement for segmentation makes this process difficult to automate as it relies solely on top level features such as object detection and semantic processing of each video frame (although some applications are easier than others, e.g. video conferencing).

Previous work [5, 6] has indicated that video coding standards could benefit in terms of perceived visual quality by focusing more on local features of the coded video, such as textures and motion. In this paper we extract such features at the CTU level with the aim of quantifying the levels of masking that characterise each CTU. We additionally employ foveation aspects to further capture masking activity in the video. Foveation (gaze location) is of course a top level feature, but it is one that can be determined offline using hardware that is becoming more commonly available (eye trackers). In certain scenarios, it has been argued that it can also be predicted [7].

The rest of the paper is organised as follows. Section 2 describes the masking factors used in building the perceptual coding map and how this map drives QP adaptation in coding with HEVC. Section 3 presents the subjective study conducted to evaluate the performance of the proposed coding approach, discusses the results obtained and highlights possible further work based on them. The paper finishes with a summary of the major points and possible future improvements in the conclusions section.

2. PERCEPTUAL CODING MAP AND QP ADAPTATION

2.1. Perceptual Coding Map

The proposed perceptual coding map aims to exploit visual masking characteristics of the Human Visual System (HVS) for assigning coding importance to blocks of pixels (CTUs in HEVC). In particular, the coding map employs spatial (texture) and temporal (motion) features for quantifying masking levels associated with a particular block. Content specific visual attention patterns and principles of foveation (eccentricity) are additionally used for determining a block's masking level. Blocks with higher masking levels are associated with higher tolerance to artefacts and can thus be ranked as of lower coding importance (coarser quantisation/ less bits). In the following paragraphs, we describe in detail the three perceptual factors used in building our coding map.

The work presented was supported by the "Marie Skłodowska-Curie Actions - The EU Framework Programme for Research" project PROVISION.

2.1.1. Texture Masking

According to *Gestalt Theory* [8], contour closure and continuity are important tools that aid perceptual organisation and help our brain paint a perceptual picture of a real world scene [9]. In recent times, this phenomenon has been attributed to the aggregation and pooling mechanisms that occurs in the primary visual cortex V1 [10]. Any compression algorithm can therefore handle such regular contours and edges very delicately while introducing artefacts more aggressively in smooth and textured regions [11]. In this vein, we first search for such regular and homo-directional edges in the scene using the directional edge entropy approach [12, 13]. First, a Sobel filter is applied to determine the horizontal and vertical gradients, i.e. $G = [G_x, G_y]$. The dominant direction d in a localized region is obtained by maximizing $d'G$. This is simply the eigenvector of the 2×2 structure tensor $S = [G, G']$. After determining the direction of edges in every local region, we calculate the entropy of the resulting histogram. This is a direct representation of the homogeneity of contours in a given area and forms the basis for our coding map (see the example in Fig. 1).

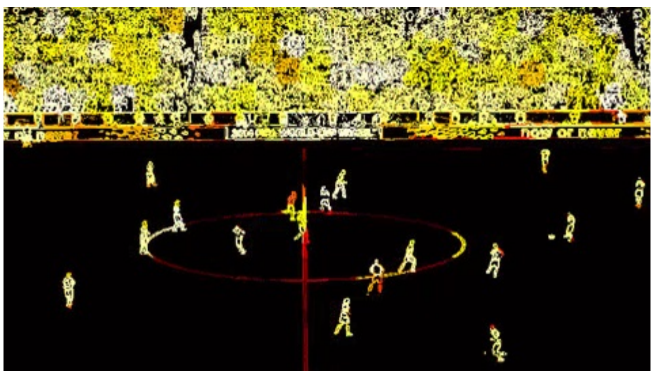

Fig. 1: Heatmap of edge entropy (high is shaded with yellow and low with red) in different areas for a random frame of a football match.

2.1.2. Temporal Masking

Sensitivity to visual artefacts is not only determined by an objects's spatial characteristics, but also by its temporal [14]. Especially, in scenarios where a subject is making a smooth pursuit eye movement to follow an object of interest, the sensitivity is determined by a stationary Contrast Sensitivity Function (CSF) computed at velocity zero, rather than a temporal one. Our sensitivity towards pursued targets is therefore much higher than in cases when the motion is more random and less likely to be tracked. This concept forms the basis of several objective quality algorithms [15, 16, 17, 18]. To determine the nature of movement of different objects in the scene, we first cluster the scene into superpixels using the SLIC approach [19]. Each of these individual super-pixels is then tracked over time using a simple super-pixel matching approach to obtain a time dependent trajectory [20]. In every case, we compensate for the global (camera) motion, that is determined using a dense optical flow algorithm [21]. The "peakiness" of these trajectories are then used to determine the smoothness of movement. Here, we imply that a smoothly moving object is likely to be tracked and must be compressed less aggressively than those cases where the motion is more irregular.

2.1.3. Foveation Masking

Both the temporal and spatial sensitivity of the human visual system decrease as we move away from the fixation point [22] and temporal [23]. Visual processing is said to switch to a coarser spatial scale due to the reduced density of ganglion cells, also known in literature as cortical magnification. Concepts of foveation have been used to design foveated coders [24] and transmission strategies [20]. The drop in resolution across the periphery has been quantified in a sound manner in the studies of Geisler et al. [25] with a setup known as the gaze contingent display. Further, Rai et al. [26] have observed that the cortical magnification factor highly correlates with the drop in Differential Mean Opinion Scores (DMOS) for the same distortion at different eccentricities. In this work, we determine a foveated coding map for each video frame by using pre-recorded gaze locations as the center of interest in every frame. The cortical magnification factor is important for deciding the fall-off characteristics of the assigned coding importance.

2.2. QP adaptation

The previously described perceptual coding map drives a QP adaptation process via a mapping function that translates the combined texture, motion and foveation (TMF) masking levels into QP values for each CTU. TMF levels range between 0 and 1. The mapping function employs a weighted average of the three masking factors as shown in Eq. (1) below:

$$\mathrm{TMF}(f_1, f_2, f_3) = \frac{w_1 f_1 + w_2 f_2 + w_3 f_3}{w_1 + w_2 + w_3}\ , \tag{1}$$

where f_1, f_2 and f_3 denote the three masking factors that build the perceptual coding map and w_1, w_2, w_3 are their corresponding weights. As a first approximation, equal weights have been used for the three masking factors. Follow-up work will investigate optimised weight values.

Using this weighted average of CTU masking levels, a QP offset value $\mathrm{QP_{off}}$ (difference from the frame QP) is generated for each CTU, as shown in Eq. (2).

$$\mathrm{QP_{off}(TMF)} = \mathrm{D(TMF)}\ \mathrm{QP_{off,max}}\ \ , \tag{2}$$

where $\mathrm{QP_{off,max}}$ is the maximum allowed CTU $\mathrm{QP_{off}}$ (maximum deviation from the frame QP) and D(TMF) is the mapping function. The mapping function can have a significant impact on perceived quality as it affects the spatial and temporal quality variation introduced in each frame through the QP adaptation process. We have investigated the performance of the following three mapping functions:

$$\mathrm{D_1(TMF)} = \mathrm{TMF}; \tag{3}$$

$$\mathrm{D_2(TMF)} = \mathrm{TMF}^2; \tag{4}$$

$$\mathrm{D_3(TMF)} = \sqrt{\mathrm{TMF}}; \tag{5}$$

by means of an informal subjective study that involved four experts as subjects. Based on this study, we selected the linear mapping of Eq. (3). Figure 2 shows an example frame with the TMF values and the assigned $\mathrm{QP_{off}}$ values for four sample CTUs from one of the employed test sequences.

3. SUBJECTIVE PERFORMANCE EVALUATION

In order to evaluate the performance of the proposed TMF based QP adaptation (TMF-QP), we performed a subjective quality evaluation

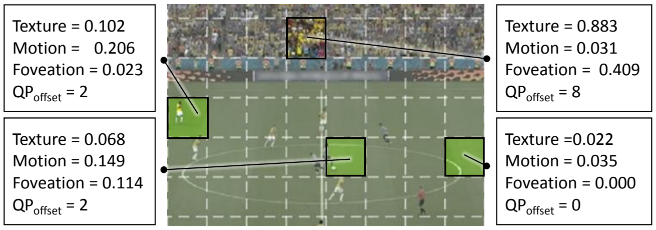

Fig. 2: Example TMF values (texture, motion foveation) and assigned QP_{offset} values for four sample CTUs in one of the test sequences.

study involving 22 subjects and two test sequences. The duration of each test sequence was 7 seconds. A total of 56 pairs of sequences were shown to each viewer with each session lasting 30 minutes. Subjects were sitting at a distance of three times the height of the screen. The displayed sequences were coded at Full HD resolution (1920×1080) at a frame rate of 30 frames per second using HEVC HM 16.2. QP adaptation was facilitated through the Adaptive QP (ADQP) functionality offered by HM 16.2. The BT500 protocol [27] was followed for displaying the test sequences and collecting quality scores for each sequence. The participants were asked to answer to the following question: "How annoying are the compression artefacts in each sequence?" Scores were given on a continuous scale from 0 to 100 with evenly spaced five ticks annotating the "Very annoying", "Annoying", "Slightly Annoying", "Perceptible, but not annoying" and "Imperceptible" levels.

Utilising foveation masking assumes knowledge of the viewers' gaze pattern and strong intra-viewer gaze location consistency. Even though the former is becoming increasingly possible with advances in eye tracking technology, real time coding using foveation remains problematic due to latency issues. Intra viewer gaze location consistency has been shown to be high for generic content (e.g. movies [28]) and even higher for sports type content (e.g football) ([7]), making fixation location prediction a strong possibility. The scenario we envisaged in this study is one of sequences being coded offline at various bitrates for streaming purposes. The two test sequences that we used were both broadcast football sequences for which eye tracking data were collected and used in calculating foveation masking.

The two sequences used vary in the amount of textured areas and the types and amount of motion present. Test sequence *S1* was shot with a still camera, while test sequence *S2* was captured with a panning camera. The sequences were encoded in Random Access configuration for base QP (QP_{base}) values of 15-51 and a QP variation (QP_{var}) of 0-35 in increments of 5. Sequences coded with with a zero QP_{var} are referred to as "anchor points", and represent the performance achieved with standard HM. The rest of the sequences are referred to as "test points" and represent the performance of our method. Out of all available anchor points, sequences with QP_{base} values of 26, 30, 34 and 38 were selected as being representative of high, medium-high, medium-low and low quality respectively. These sequences where then paired with test points of the same bitrate (within an error margin of up to 6.5%). Table 1 shows the 56 pairs of anchor and test points used in the subjective study. Sequences coded with TMF-QP adaptation have lower QP_{base} values than those generated using standard HM for the same bitrate. Increasing QP_{var} values further lowers QP_{base}.

Table 1: Pairs of anchor (column 1, QP_{var}=0) and test points (columns with QP_{var}=[5 to 35, step 5]) resulting in similar bitrate.

Sequence		S1							S2						
QP_{var}	0	5	10	15	20	25	30	35	5	10	15	20	25	30	35
QP_{base}	26	24	23	21	20	19	18	17	24	22	21	20	19	17	17
	30	28	27	25	24	23	22	21	28	26	25	24	22	21	20
	34	32	31	29	28	27	26	25	32	30	29	28	26	25	24
	38	36	35	33	32	31	30	29	36	34	33	31	30	29	28

3.1. Results and Discussion

The collected opinion scores were normalized and filtered for outliers, as described in recommendation BT.500 [29]. This resulted in the scores of one subject being rejected. A second outlier rejection pass was performed at the individual score level as suggested in [30]. The filtered scores where averaged to form a Mean Opinion Score (MOS) for each anchor and test point.

Table 2 lists the number of times (as a percentage of the total number of votes) that subjects expressed preference towards one method over the other. The frequency of the no-preference case is also given (denoted as "Equal"). HM stands for sequences coded with no QP adaptation and TMF-QP denotes sequences coded with the proposed approach. Results are averaged over all QP_{base} values and presented separately for each QP_{var} step. It can be seen that the TMF-QP approach is on average preferred over standard HM coding when QP_{var} values range between 10 and 30, with a maximum average preference of 54% recorded for a QP_{var} value of 20. Subjects on average preferred the HM coded sequences when the QP variation was either too small (QP_{var} of 5) or too large (QP_{var} of 35).

Table 2: Percentage of preference for stock HM versus our method averaged for all QP_{base} values of both sequences.

QP_{var}	5	10	15	20	25	30	35
TMF-QP	38.7%	**43.0%**	**44.3%**	**54.5%**	**44.5%**	**48.5%**	40.4%
Equal	17.8%	20.0%	19.8%	15.2%	12.8%	13.9%	10.8%
HM	**43.6%**	37.0%	35.9%	30.3%	42.7%	37.6%	**48.8%**

Table 3 offers more insight on the previous observations by listing preference results for each QP_{base} and QP_{var} combination. Each row of Table 3 shows results at test points of similar bitrate to the anchor point corresponding to the QP_{base} of that particular row. For each bitrate (each row) we have highlighted the cell corresponding to the combination of QP_{base} and QP_{var} for which subjects showed the strongest preference for the proposed method. It can be seen that for each tested bitrate there is potential for gains in the perceived quality (stronger preference) when using the proposed method.

Figures 3 and 4 depict the MOS-Rate (MR) curves for all anchor and test points for the sequences S1 and S2 respectively. The solid blue line represents the MR curve of the anchor points. The test points are located around the anchor points and depicted using different markers. The upper (gray) and lower (brown) dashed lines represent the best and worst MR curves, respectively, that our method can achieve. The shaded area between these two dashed lines encloses all the MR points generated by TMF-QP and represents the potential perceptual gains or lossess in quality that can be achieved when coding with the proposed method relative to HM. The dotted lines represent the standard error of the mean for the best and worst MR curves. From these two figures it is clear that the TMF-QP method can offer a measurable improvement in perceived quality relative to

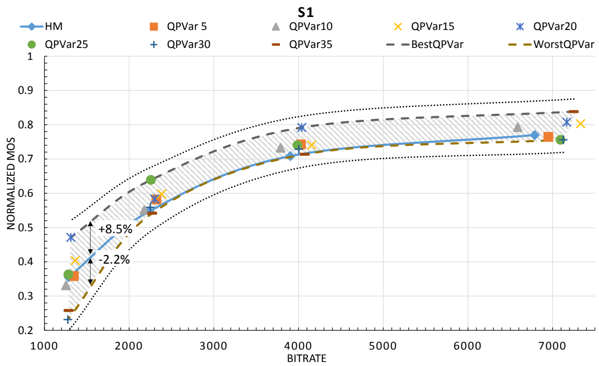

Fig. 3: MOS versus bitrate for anchor (solid blue line) and test points for sequence S1.

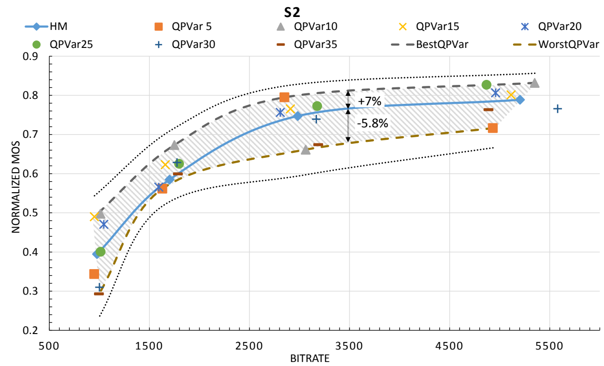

Fig. 4: MOS versus bitrate for anchor (solid blue line) and test points of sequence S2.

Table 3: Percentage of preference for stock HM versus our method averaged per QP_{base} value of both sequences.

QP_{base}	QP_{var}	5	10	15	20	25	30	35
26	TMF-QP	23.7%	40.5%	35.9%	52.4%	42.1%	**55.0%**	36.8%
	Equal	26.3%	16.7%	23.1%	16.7%	18.4%	**17.5%**	23.7%
	HM	50.0%	42.9%	41.0%	31.0%	39.5%	**27.5%**	39.5%
30	TMF-QP	57.1%	37.5%	35.7%	**70.0%**	42.9%	53.7	43.9%
	Equal	16.7%	30.0%	31.0%	**12.5%**	14.3%	17.1	7.3%
	HM	26.2%	32.5%	33.3%	**17.5%**	42.9%	29.3%	48.8%
34	TMF-QP	27.5%	48.8%	**54.8%**	52.4%	53.7%	45.2%	45.2%
	Equal	22.5%	22.0%	**11.9%**	16.7%	19.5%	16.7%	11.9%
	HM	50.0%	29.3%	**33.3%**	31.0%	26.8%	38.1%	42.9%
38	TMF-QP	40.0%	43.9%	**50.0%**	43.9%	42.9%	42.9%	35.7%
	Equal	7.5%	12.2%	**14.3%**	14.6%	0.0%	4.8%	2.4%
	HM	52.5%	43.9%	**35.7%**	41.5%	57.1%	52.4%	61.9%

a constant QP approach, with the potential gain reaching 7 to 9% in terms of Bjøntegaard MOS [31] for the two sequences tested.

Clearly the performance achieved with TMF-QP depends on appropriate selection of values for parameters QP_{base} and QP_{var}. This in turn depends on the characteristics of the content and the targeted bitrate. In addition, and as mentioned in Section 2.2, the resulting perceived quality also depends on the coding map driving the QP adaptation. Encouraged by the results shown in Fig. 3 and Fig. 4, we are currently investigating the above aspects with the aim of maximising the potential gain in perceived quality and automating the QP adaptation process using content features and bit rate targets.

4. CONCLUSIONS

In this paper, we presented TMF-QP, a QP adaptation method driven by a perceptual coding map. The subjective study performed showed that our method provides perceptually better results compared to the HM HEVC reference software, for the examined content (sports sequences) and bitrate levels (compression artifact visibility range). With careful selection of values for parameters QP_{base} and QP_{var}, the improvement in perceived quality can get close to 9%. Our work has highlighted the importance of developing a model that accurately assigns values to parameter QP_{var} given a certain QP_{base} and perceptual coding map. This is the topic of our current work.

5. REFERENCES

[1] J. Kim, S. Bae, and M. Kim, "An HEVC-compliant perceptual video coding scheme based on JND models for variable block-sized transform kernels," *Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1786–1800, IEEE,2015.

[2] H. Zeng, K. N. Ngan, and M. Wang, "Perceptual adaptive lagrangian multiplier for high efficiency video coding," in *Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 69–72.

[3] D. Chai, K. N. Ngan, and A. Bouzerdoum, "Foreground/background bit allocation for region-of-interest coding," in *International Conference on Image Processing*. IEEE, 2000, vol. 2, pp. 923–926.

[4] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," in *Photonics West'98 Electronic Imaging*. International Society for Optics and Photonics, 1998, pp. 294–305.

[5] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. R. Bull, "A video texture database for perceptual compression and quality assessment," in *International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 2781–2785.

[6] M. A. Papadopoulos, D. Agrafiotis, and D. R. Bull, "On the performance of modern video coding standards with textured sequences," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2015, pp. 137–140.

[7] Q. Cheng, D. Agrafiotis, A. M. Achim, and D. R. Bull, "Gaze location prediction for broadcast football video," *Transactions on Image Processing*, vol. 22, no. 12, 2013.

[8] K. Koffka, *Principles of Gestalt psychology*, vol. 44, Routledge, 2013.

[9] P. J. Bex, "(In) Sensitivity to Spatial Distortion in Natural Scenes," *Journal of Vision*, vol. 10, no. 2, pp. 23–23, 2010.

[10] G. Loffler, "Perception of contours and shapes: Low and intermediate stage mechanisms," *Vision Research*, vol. 48, no. 20, pp. 2106–2127, 2008.

[11] Md M. Alam, K. P. Vilankar, D. J. Field, and D. M. Chandler, "Local masking in natural images: A database and analysis," *Journal of Vision*, vol. 14, no. 8, pp. 22–22, 2014.

[12] G.N. Srinivasan and G. Shobha, "Statistical texture analysis," in *Proceedings of world academy of science, engineering and technology*, 2008, vol. 36, pp. 1264–1269.

[13] D. Agrafiotis, D. R. Bull, and N. Canagarajah, "Enhanced spatial error concealment with directional entropy based interpolation switching," in *International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2006.

[14] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly, "Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli," in *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006, p. 605705.

[15] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *Transactions on Image Processing*, vol. 23, no. 1, pp. 200–213, 2014.

[16] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1653–1658, Nov 2010.

[17] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[18] S. Péchard, D. Barba, and P. Le Callet, "Video quality model based on a spatiotemporal features extraction for H.264-coded HDTV sequences," in *Picture Coding Symposium*, 2007, p. 1087.

[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[20] Y. Rai, P. Le Callet, and G. Cheung, "Role of HEVC coding artifacts on gaze prediction in interactive video streaming systems," in *International Conference on Image Processing (ICIP)*. IEEE, 2016, vol. 2, pp. II–169.

[21] M. Tao, J. Bai, P. Kohli, and S. Paris, "SimpleFlow: A non-iterative, sublinear optical flow algorithm," in *Computer Graphics Forum*. Wiley Online Library, 2012, vol. 31.

[22] R.P. Scobey and J.M. Horowitz, "Detection of image displacement by phasic cells in peripheral visual fields of the monkey," *Vision Research*, vol. 16, no. 1, pp. 15 – 24, 1976.

[23] V. Virsu, J. Rovamo, P. Laurinen, and R. Näsänen, "Temporal contrast sensitivity and cortical magnification," *Vision Research*, vol. 22, no. 9, pp. 1211–1217, 1982.

[24] Z. Wang and A. C. Bovik, "Foveated image and video coding," *Digital Video, Image Quality and Perceptual Coding*, pp. 431–457, 2006.

[25] W. S. Geisler, J. S. Perry, and J. Najemnik, "Visual search: The role of peripheral information measured using gaze-contingent displays," *Journal of Vision*, vol. 6, no. 9, pp. 1–1, 2006.

[26] Y. Rai, A. Aldahdooh, S. Ling, M. Barkowsky, and P. Le Callet, "Effect of content features on short-term video quality in the visual periphery," in *18th International Workshop on Multimedia Signal Processing (MMSP)*. Sept 2016, pp. 1–6, IEEE.

[27] ITU-RBT Recommendation, "500-11, methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union, Geneva, Switzerland*, vol. 4, pp. 2, 2002.

[28] Robert B Goldstein, Russell L Woods, and Eli Peli, "Where people look when watching movies: Do all viewers look at the same place?," *Computers in Biology and Medicine*, vol. 37, no. 7, pp. 957–964, 2007.

[29] ITU Radiocommunication Assembly, *Methodology for the subjective assessment of the quality of television pictures*, International Telecommunication Union, 2003.

[30] J. W. Tukey, "Exploratory data analysis," 1977.

[31] G. Bjontegaard, "Calcuation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April 2001*, 2001.