Full Reference Video Quality Model for UHD HEVC Encoded Sequences

Franck Chi, Xavier Ducloux, Gérard Madec, John Puentes

FULL REFERENCE VIDEO QUALITY MODEL FOR UHD HEVC ENCODED SEQUENCES

Franck CHI, Xavier DUCLOUX, Gérard MADEC, John PUENTES

Hypermedia laboratory, B<>COM, 35510 Cesson-Sévigné, France
Image and Information Processing department, TELECOM Bretagne, 29238 Brest, France
Lab-STICC - UMR CNRS 6285, 29238 Brest, France

ABSTRACT

Despite its relative robustness, subjective video quality evaluation is a time-consuming and costly process. Alternatives are required therefore to simplify visual quality estimation, particularly in the case of new video formats. This paper presents an analysis of full reference quality metrics focused on Ultra High Definition sequences, encoded with H.265/High Efficiency Video Coding. After evaluating the individual performance of three objective video quality metrics - structural similarity, gradient difference, and motion distortion - an optimal combination is defined, to be weighted by three perceptibility criteria, considering luminance, motion, and texture masks, in uniform and selective perception contexts. Performances at each step are compared by correlation, to subjective scores of each sequence given by Subjective Assessment Methodology of Video Quality session. A close correlation to subjective quality measurements is measured applying three indicators.

Index Terms— Video quality evaluation, High Efficiency Video Coding, Ultra High Definition, structural similarity, gradient, motion vectors.

1. INTRODUCTION

Estimating consumer perceived video quality has been always a major challenge for content distribution and delivery professionals. Although, significant advances have been accomplished in video quality assessment research in the past twenty years, the recent emergence of a new Ultra High Definition (UHD) video format is much likely to require adapted quality measures. Namely, application of the novel H.265/HEVC compression standard [1], raises the question about how pertinent are existing quality measures.

Separate benchmarking of full reference objective video quality metrics, over UHD and/or HEVC encoded content have been already analyzed in [2]. Beyond this essential analysis, this paper studies the performance gain of a complete quality model, based mainly on local spatial noise weightings, according to coding defect perception resilience, and selective visual perception. Although more performant and refined video quality metrics exist, their application to UHD 4K 50Hz HEVC encoded sequences, requires a complex infrastructure to cope with high processing constraints. For this reason, simpler local quality metrics, associated to modeled perception criteria are proposed, as an alternative approach.

To present this contribution, Section 2 introduces the Subjective Assessment Methodology of Video Quality (SAMVIQ) that serves as reference to all the objective measurements of this work. The three selected video quality metrics for coding defect estimation are described in Section 3. A suitable combination of quality metrics to improve the known performance of individual measures is proposed in Section 4. The impact of distortion is described in Section 5, and the effect of selective perception, is described in Section 6. Main obtained results are discussed in Section 7.

2. SUBJECTIVE EVALUATION

The SAMVIQ methodology [3] has been defined in order to discriminate perceived quality of multimedia content. Initiated by the European Broadcasting Union (EBU), the video evaluation part is standardized in ITU-R-BT.1788. For each scene of 10s to 20s containing explicit and hidden references, several sequences under particular test conditions are proposed to viewers. They can play and rate sequences in any order on a continuous quality scale (0 to 100), making use of five quality items.

2.1. Test scenes

Three UHD 4K video sequences (in a YUV 420 8 bit format) are selected for subjective evaluation, because of their complex diversity in motion and texture. These sequences come from Sveriges Television AB (SVT) and EBU test sets [4, 5]. Table 1 lists the basic video characteristics of each scene.

Table 1: Scene description

Scene	Source	Frame rate	Duration
Crowdrun	SVT	50	10s
Park_Dancers	EBU	50	15s
Studio_Dancers	EBU	50	15s

2.2. Encoder configuration

Scenes are encoded using the HEVC test model, HM11 [6]. The chosen configuration is random access. In this configuration, some frames are periodically intra-coded (I-frame) while others called predicted frames (P-frame) or bi-predicted frames (B-frame) are coded using previously coded I, P or B frames as shown in Figure 1. The selected profile is "main" and the level is 6.2, because it allows up to 3840*2160p60 HEVC encoding. Video encoding is performed applying a fixed quantization parameter (QP).

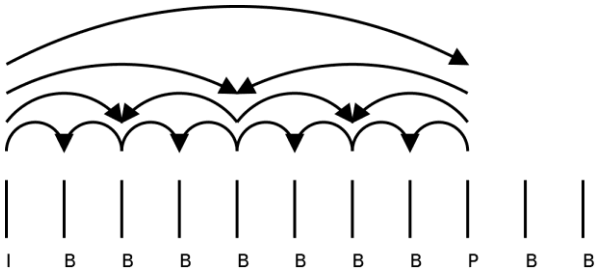

Figure 1: Example of the encoding structure.

QP sets are selected to provide a large subjective quality variation, from high/acceptable to impaired. Table 2 lists defined QP sets for each scene and Table 3 the associated output HEVC bit rate.

Table 2: QP selection per scene

Scene	QP 1	QP 2	QP 3	QP 4
Crowdrun	32	35	38	41
Park_Dancers	29	32	35	38
Studio_Dancers	29	32	35	38

Table 3: Bitrates in Mbps per scene and per QP

Scene	QP 1	QP 2	QP 3	QP 4
Crowdrun	22.7	15.1	10.3	6.9
Park_Dancers	10	6.5	4.3	2.7
Studio_Dancers	7.6	5.1	3.5	2.4

2.3. Methodology

The SAMVIQ session is performed by Orange Labs in Rennes, France, respecting the viewing room illumination recommended by ITU-R BT. 500. A JVC ProVerite is used for testing display. Table 4 lists the main display characteristics.

Table 4: Display specifications

Screen technology	LCD-LED
Screen size	84''
Screen definition	3840*2160
Video Input	SDI*4

Based on a HP Z840 with a Matrox 4K video monitoring card, the player hardware achieves real-time playback of 3840x2160p60 raw video samples. It uses the Subjective Evaluation of Video Quality player software, developed by Orange. To carry out the tests, 24 non-expert observers are selected. Nevertheless, after the rejection method is applied, only 21 observer scores are kept. All observers are seated at 1.5H during the tests. As illustrated in Figure 2 by three separated lines, observers easily distinguish the different levels of quality. Note also that the Crowdrun scene shows a consistent lower perceived video quality, despite its higher bit rate.

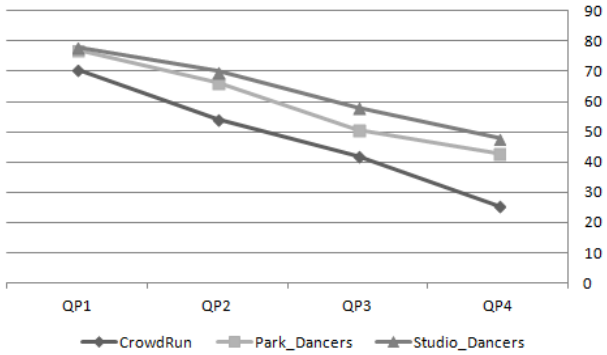

Figure 2: SAMVIQ perceived video quality per QP and scene.

3. INDIVIDUAL OBJECTIVE VIDEO QUALITY METRIC ANALYSIS

Several quality measurements could be applied in our case. Among those possibilities, Structural SIMilarity (SSIM), gradient magnitude difference, and motion distortion provide complementary and suitable local measures of video distortion. SSIM is widely used in the industry and academic research, for video quality estimation, while spatial gradient magnitude difference and motion distortion are used to complete SSIM analysis. These metrics are evaluated separately, to compare their correlation to the described subjective evaluations.

3.1. SSIM

SSIM [7] is a method to measure local spatial similarities. For two signals *x* and *y*, it is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (1)$$

With:

- μ_x and μ_y *x* and *y* averages respectively.
- σ_x^2 and σ_y^2, *x* and *y* variances respectively.
- σ_{xy}, *x* and *y* covariance.
- $c_1$and c_2, constants.

3.2. Gradient Magnitude Difference

The gradient magnitude difference highlights defects on edges. This measurement is applied due to SSIM deficiencies in detecting block artifacts on test sequences. Gradient maps of the reference and test sequences are generated with Sobel horizontal and vertical operators in respectively Equation (2) and (3).

$$\begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (3)$$

The gradient magnitude difference is processed as depicted in Figure 3.

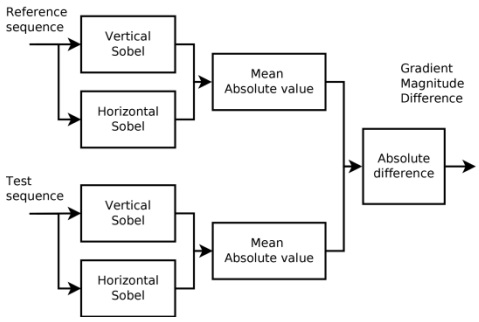

Figure 3: Processing of gradient magnitude difference.

3.3. Motion distortion

SSIM and gradient magnitude do not consider motion variation. Motion distortion is expressed as the distortion between motion vectors of the reference and test sequences. Motion magnitude difference (D_M) and the cosine of motion orientation difference (D_{angle}) are processed according to Equations (4) and (5). Besides, *Motion_Difference* (Equation (6)) provides an estimation of local motion quality:

$$D_M = \frac{2Mag_{Ref}Mag_{test}}{Mag_{Ref}^2 + Mag_{test}^2} \quad (4)$$

$$D_{Angle} = \frac{1 + cos(\theta_{Ref} - \theta_{Test})}{2} \quad (5)$$

$$Motion\_Difference = 1 - D_M * D_{Angle} \quad (6)$$

3.4. Performance comparison

Estimated qualities of video sequences are calculated by averaging local quality scores spatially and temporally, over the whole sequences. Three commonly applied correlation indicators are selected for performance evaluation [8]:

- The Spearman rank correlation coefficient (SRCC).
- The Kendall rank correlation coefficient (KRCC).
- The Pearson linear correlation coefficient (PLCC).

These correlation coefficients denote high correlation for values close to 1. Correlation values are computed using the subjective and objective quality estimations, over the three scenes and their respective QP sets. Figure 4 presents the individual performance of the three previously described objective quality metrics, according to the correlation indicators. The best performance corresponds to SSIM, while motion vector difference is the worst. Such difference can be explained by the high variability of the motion vector difference measurement, due to the importance given to orientation difference.

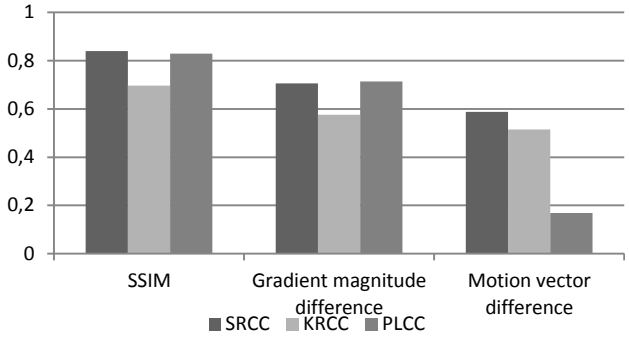

Figure 4: Comparison of individual objective video quality metric performances.

Given these dissimilarities and considering the importance of taking into account the quality metrics, an optimized combination was further investigated, looking for an improved correlation with subjective evaluations.

4. QUALITY METRIC COMBINATION

The proposed approach to combine the previously analyzed metrics is to build a weighted sum. As seen in Figure 5, a weight is defined for each metric.

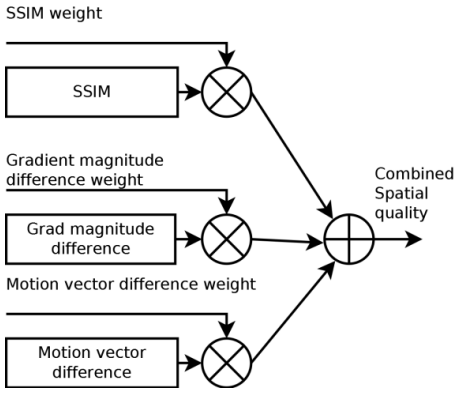

Figure 5: Combination of spatial quality metrics.

For our purpose, following experimental sampling tests, weights are assigned using three values - 0, 1, and 5 - permitting to cope with processing time constraints,

adapted to multiple UHD 4K 50Hz sequences. A spatial quality weighted sum is calculated for each evaluated block of a sequence. Thereafter, obtained values are integrated over the sequence by using an average function. All the test video sequences are processed in the same manner, before calculating the relative performance results.

Making use of the different associations of values {0-1-5; 0-1-5; 0-1-5}, the quality metrics are combined and then compared applying the three correlation coefficients. Figure 6 represents the obtained performance for each weights set of local spatial quality metrics. Each orientation denotes a weight set identified by its label. The digits represent from left to right, the SSIM weight, the gradient magnitude difference weight, and the motion vector difference weight. The closer a point is to the outer edge of the circle, the higher the respective correlation coefficient is. As observed in the first analysis (section 3.4), values sets giving the highest weight to SSIM, display the highest correlation with the subjective quality. Similarly, weight sets giving the highest weight to motion vector difference, display the lowest correlation.

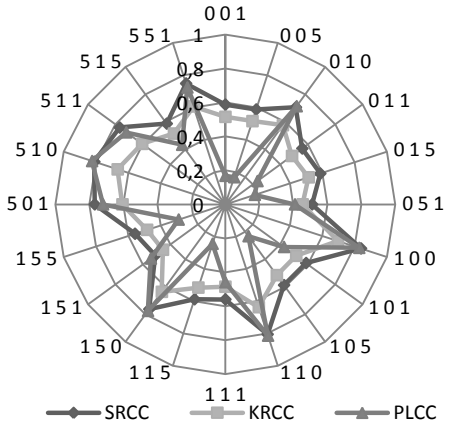

Figure 6: Correlation coefficient values per spatial quality metric weight set.

5. INFLUENCE OF DISTORTION PERCEPTIBILITY

Section 4 analyses the video quality estimation variation brought by combining several local quality metrics. However, it does not take into account the spatio-temporal content of the sequence which can highly influence the locally perceived video quality. In this section, the distortion perceptibility map represents the local content resilience to coding artifacts. Such map is generated through a sequence of pre-analysis. Three maps based on human visual system properties are calculated and summed into a perceptibility mask.

5.1. Processed masks

- *Luminance masking*: It assumes that this masking effect considers darkest and brightest areas, as being more resilient to distortions [9]. The masking effect integrates local and neighboring luminance.
- *Motion masking*: It considers that motion characteristics can reduce perceptual impacts of coding artifacts. Motion magnitude and ego-motion [10] are therefore estimated for the mask.
- *Texture masking*: Content texture is analyzed using three components decomposition [11]. As a consequence, textured zones are more perceptively resilient to coding artifacts, than smooth and edge areas.

5.2. Processing of spatial perceptible quality

A perceptibility mask is defined as the weighted sum of the luminance, motion vector magnitude, and texture masks. The mask is applied then to the previously defined combined local spatial quality. Tested values for perceptibility weights are also {0; 1; 5}. Methodology for sequence quality estimation and performance indicators is identical to the one previously described. Figure 7 displays how spatial perceptible quality is processed.

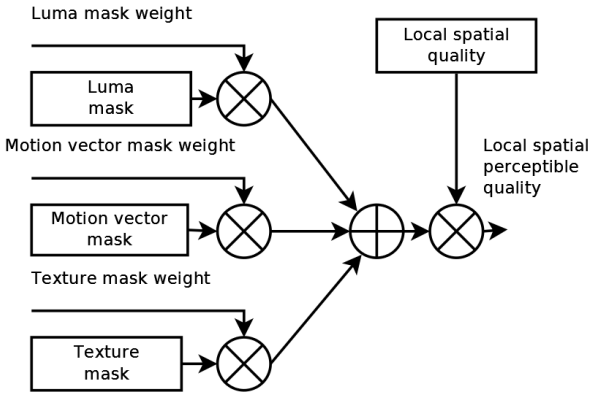

Figure 7: Spatial perceptible quality calculation.

Our analysis focused on one particular spatial weight set, {5; 1; 0}, which is 5 for SSIM, 1 for gradient magnitude difference, and 0 for motion vector difference. From an experimental point of view, it appears to be the best configuration using two spatial quality metrics.

Figure 8 displays the obtained performance for the different weight sets applied to calculate the perceptibility mask. Each line of the figure represents a perceptibility weight set, identified as in Figure 6. From left to right, digits describe respectively the weights of luminance perceptibility, motion perceptibility, and texture perceptibility. The lowest line represents the {0; 0; 0} weight set, which means that no perceptibility mask is applied. As a result, this perceptibility configuration shows one of the lowest correlations with subjective quality. While it is difficult to highlight one perceptibility configuration, it is clear that configurations considering only one perceptibility mask do not generate the best performances. Conversely, different masks appear to

complete each other, since configurations including the three masks perform well.

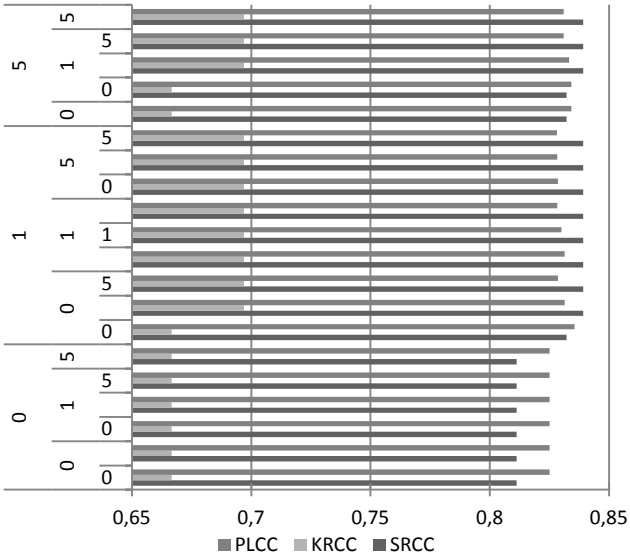

Figure 8: Perceptibility weight set performance.

6. SELECTIVE PERCEPTION

In Section 5, a content distortion perceptibility mask was added to the local quality estimation. Nevertheless, it was not considered that local distortion perception also depends on its visibility relatively to other distortions. This section assumes that distortion perception is not uniform, depending on the visual context and spatio-temporal distortion distribution. Firstly, this paper considers thus that the perception of distorted zones depend on their size [8]. A mask for spatial visibility of distortions is processed for each local quality metric, taking into account spatially neighboring quality. Secondly, as documented in [12], highly distorted zones inhibit perception of lesser distorted ones. This section analyzes therefore the quality estimation variation brought at the sequence level by using percentile frame distortion estimation, instead of a simple average of the local values. Figure 9 displays the overall process applied to the spatial perceptible quality map.

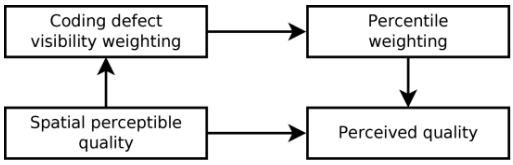

Figure 9: Perceived quality estimation.

As shown in Figure 10, weighted coding defect visibility has no impact on the correlation with subjective scores.

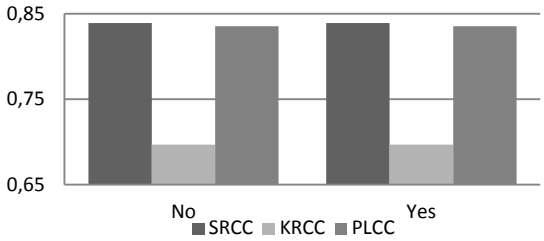

Figure 10: Influence of spatial visibility of distortions

On the other hand, Figure 11 shows the performance gain when the average integration (letter M prefix) is replaced by percentile integration (letter P prefix), using the same spatial configuration as defined in section 5. This diagram shows that the perceptible configurations perform better with percentile application. It is interesting to note as well, that similar correlation values are obtained in this case, regardless of the perceptibility weight set.

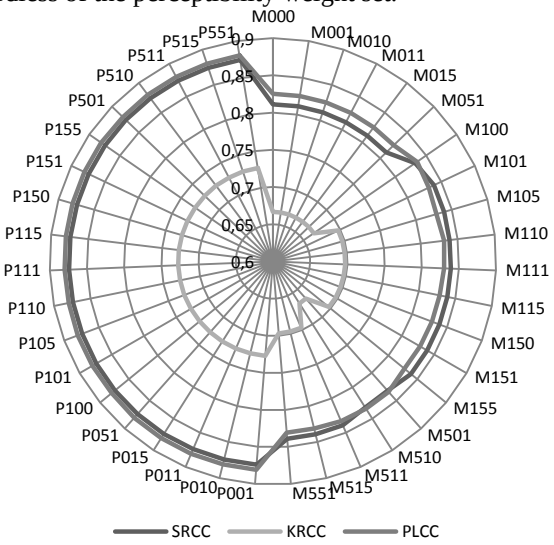

Figure 11: Sequence percentile/mean quality evaluation comparison.

Given the strong impact of percentile application, it is important to check whether the spatial quality metrics combination selected in section 4 remains the optimum one. An interesting finding is that a significant improvement is brought by selecting the configuration {1; 5; 0} instead of {5; 1; 0}, i.e. giving a stronger weight to gradient magnitude distortion in combination with SSIM. The application of percentile is particularly efficient in this configuration, with a 0.12 gain in correlation values. Such configuration results in the following correlation scores: PLCC = 0.897, SRCC = 0.881 and KRCC = 0.758.

Figure 12 compares the previously described configuration result to subjective quality evaluation scores.

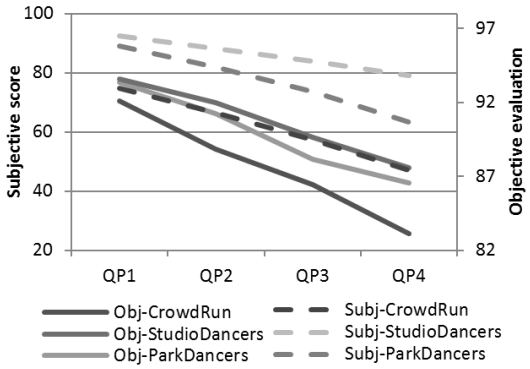

Figure 12: Subjective and perception objective quality estimation.

A step by step full reference video quality model can be built assembling all the presented processing stages. Figure 13 represents the integration of the previously described functions, in order to estimate the perceived distortion in each block of the test sequences.

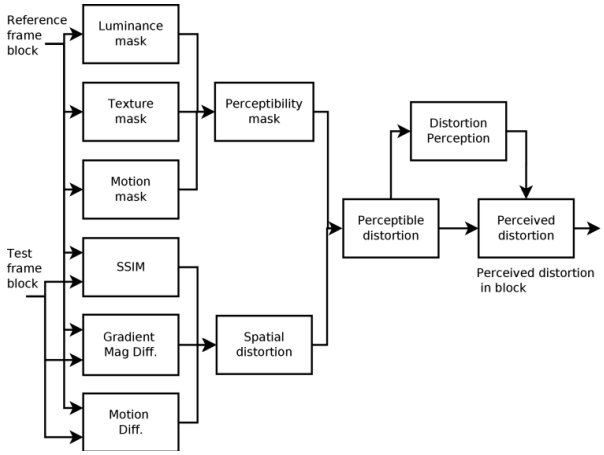

Figure 13: Overall architecture of the model.

7. DISCUSSION

The analysis of full reference objective quality models over UHD HEVC must take into account processing time constraints, given the pixel by pixel nature of the calculation. An alternative to simplify objective visual quality estimation, without requiring a complex processing infrastructure has been examined. The objective quality score highlights the most perceptible block distortions in each frame of the sequence. Perceptible block distortions are computed as a weighted sum of two local objective quality metrics, filtered by local perceptibility of coding defects. Combined proposed improvements display interesting results, especially on the PLCC coefficient with an increase of 10% compared to SSIM alone.

Obtained curves (Figure 12) have similar shapes but values differ and objective measures seem smoother than subjective scores. Results indicate that although calculated quality estimations follow the same qualitative order, from low to high visual quality (Crowd Run < Park Dancers < Studio Dancers) as subjective evaluations, the scales are not the same. While subjective measures vary from around 25 to 70, the proposed objective measure varies from 86 to 98. This suggests that the proposed model produces a similar evaluation as subjective scores, at a more compact scale, raising the question of how this behavior may replicate in other test sequences.

Several future works have already been identified. Models will be extended in the temporal dimension for quality estimation and perception weighting. Model performance will also be confirmed on a wider dataset. Finally, the model will also be evaluated, optimized, and eventually modified using other observation at the local level.

8. REFERENCES

[1] Rec. ITU-T H.265, High Efficicency Video Coding, http://www.itu.int/rec/T-REC-H.265-201304-I.

[2] Hanhart, P, Korshunov, P and Ebrahimi, T. "Benchmarking of quality metrics on ultra-high definition video sequences," *IEEE Trans. IP*, vol. 20, pp. 1185-1198, 2011.

[3] Rec. ITU-R Methodology for the subjective assessment of video quality in multimedia applications http://www.itu.int/rec/R-REC-BT.1788-0-200701-I/en

[4] Haglund, L. The SVT High Definition Multi Format Test Set, 2006. http://media.xiph.org/svt/SVT_MultiFormat_v10.pdf

[5] EBU UHD-1 Test Sequences, http://tech.ebu.ch/testsequences/uhd-1

[6] HM, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[7] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE T. Img. Processing*, vol. 13, no. 4, pp. 1-14, 2004.

[8] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," IEEE Trans. IP, vol. 20, pp. 1185-1198, 2011.

[9] Y.Zhao, L.Yu, Z.Chen and C. Zhu "Image Video Quality Assessment Based on Measuring Perceptual Noise from Spatial and Temporal Perspectives," *IEEE T. Circuits and Systems for Video Technology*, vol. 21, no. 12, pp.1890–1902, 2011.

[10] J. Park, K. Seshadrinathan, S. Lee and A.C. Bovik, "Video Quality Pooling Adaptive to Perceptual Distortion Severity," *IEEE T. Img. Processing*, vol. 22, no. 2 pp. 610–620, 2013.

[11] C.Li and A.C. Bovik, "Content-weighted video quality assessment using a three-component image model," *J. Elect. Imaging 19(1)*, doi:10.1117/1.3267087, 2010.

[12] Moorthy, A. K. and Bovik, A. C., "Perceptually significant spatial pooling techniques for image quality assessment," *Human Vision and Electronic Imaging XIV*. Proceedings of the SPIE 7240 (January 2009).