# PERCEPTUAL QUALITY ASSESSMENT FOR H.264/AVC COMPRESSION

*Piotr Romaniak, Lucjan Janowski, Mikołaj Leszczuk, Zdzisław Papir*

AGH University of Science and Technology, Department of Telecommunications, Kraków, Poland

## ABSTRACT

The paper proposes a No-Reference (NR) metric to objectively assess the H.264/AVC video quality. The proposed model takes into account the typical artefacts introduced by hybrid block-based motion compensated predictive video codecs as the one related to the H.264/AVC standard. More specifically, these artefacts are the blockiness introduced at the boundaries of each coded block and the temporal flickering due to different coding modes used for the same macroblock along the video sequence. Furthermore, a flickering metric for intra coded frames is also derived. The quality prediction accuracy of the proposed NR quality metric is validated over subjective data collected during a video subjective evaluation experiments. Moreover, the quality prediction accuracy is also compared with the one provided by the well known state-of-the-art Structural SIMilarity (SSIM) metric which works in a full-reference mode. The proposed metric achieves a higher Pearson's correlation coefficient with subjective scores than the one achieved by the SSIM metric.

## 1. INTRODUCTION

The video streaming market is growing rapidly, partly because it is being used by a growing number services requiring different levels of network resources allocation. Therefore, video streaming bit-rate has to be limited by compression what results in a degradation of quality perceived by an end-user.

For this reason video service providers are looking for reliable solutions for a constant quality monitoring in an in-service mode. The implementation of such solutions is beneficial on only for the providers but also for the end-users. A key factor here is to ensure the perfect balance between cost and quality of the provided services.

The question arises whether it is possible to derive a reliable quality monitoring system for in service applications? Two main challenges that have to be met are: 1) high performance expressed in terms of a correlation with MOS, and 2) usability understood as the ability to utilize a metric in real life applications.

A reliable approach towards perceptual assessment of video compression schemes requires addressing artefacts related to intra- and inter-frame compression. Furthermore, any metric suited for in-service applications should represent the no-reference approach because in a real system it is very likely that the original sequence is not available [1].

This paper proposes *no-reference* metrics for compression artefacts measurement based on decompressed *image analysis*. In contrast to the parametric approach, a metric based on image analysis detects *a cumulative effect* of compression, i.e. all the artefacts actually seen by the user. We verify it using the H.264/AVC compression scheme. Image analysis approach is not aware of transmission scheme, bit-stream standard and coding parameters, therefore it can be generalized over different DCT-based video compression schemes involving both intra- and inter-frame compression. Intra-frame compression is addressed by a typical metric measuring the blockiness artefact $B$. In order to account for inter-frame compression artefacts we used an *improved flickering metric* operating on a macro-block level $F$. The overall video quality assessment is supplemented by a custom metric dedicated to *I-frames flickering* $IF$.

We also present three MOS models for H.264/AVC compression: 1) a model based on the blockiness metric MOS($B$), 2) a model based on the flickering metric MOS($F$), and 3) an *integrated model* including all three metrics MOS($B$, $F$, $IF$). The models were derived from results obtained in an extensive *subjective experiment* conforming to ITU-T P.910 [2] and VQEG [3] methodology.

Two video *content characteristics* (spatial activity $SA$ and temporal activity $TA$) were considered in order to improve model performance in terms of correlation with Mean Opinion Scores (MOS). In order to demonstrate *high performance* of the integrated model, it has been compared with SSIM, a well known full-reference metric.

This paper is organized as follows. Section 2 describes related work. In section 3 no-reference video quality metrics are discussed. Subjective experiment and analysis of results are detailed in section 4 and 5 respectively. Section 6 concludes the paper.

## 2. RELATED WORK

The blockiness artefact has been analyzed by many researchers because it is one of the crucial compression effects. An interesting overview and comparison of blockiness metrics is presented in [4]. Blockiness artefact measurement utilizes a property of a block-based coding schemes. The

978-1-4244-8790-5/12/$26.00 ©2012 IEEE 

higher compression the more visible boundaries between the neighboring coding blocks. Previous works utilizing this fact are described in [5]. According to the work presented by Pandel in [6], flickering is the most annoying temporal artefact inherent for predictive inter-coded video sequences (in particular for H.264/AVC encoded sequences). Video sequences with slow camera movements or zoom are especially exposed to flickering artefact. In temporal predictive coding macro-blocks are not updated (encoded) until the difference between corresponding macro-block of successive frames is above a specific threshold. The stronger the compression, the higher the threshold. This suggests that each macro-block remains in one of two possible states: 1) no-update – differences between successive frames are below the threshold, and 2) update – the opposite case [6]. Frequent changes between these two states denote a severe flickering artefact. The metric is calculated as a normalized number of transitions between states. The two state model including a hysteresis is detailed in [6]. All the presented compression artefact metrics were analyzed separately and no integrated compression metric was proposed.

An interesting work presenting an integrated no-reference quality metric for degraded and enhanced video is presented in [7]. Several image and video artefacts were combined into an integrated formula for the overall quality assessment. Considered artefacts were related to the source quality and video compression. Subjective experiments were carried out in order to train and verify the proposed model. Obtained correlation was satisfactory and significantly higher than for PSNR but only a limited number of tests sequences was used. Additionally, no discussion on the sequences complexity and diversity was provided. The results have been elaborated for previous coding standards (when compared to H.264/AVC) where except blockiness other compression artefatcs like ringing or clipping were essential [7]. For the current hybrid block-based motion compensated predictive coding schemes (H.264/AVC) it is more important to account for temporal artefacts like flickering [6], what is reflected in this paper.

Another group of video quality metrics suited for compression represents a full-reference approach implying serious limitations, especially when real video services are considered [1]. Applications of such metrics are restricted to laboratory scenarios, for example comparison of compression schemes. An example of metrics suited for H.264/AVC is given in [8].

In parametric approaches, quality estimation is based on quantizer information from the H.264 bit-stream (examples can be found in [9] and [10]). Such an approach is restricted to a given bit-stream model, video codec, and even codec profile; however, it cannot be easily adopted to other compression schemes. Moreover, it is common for video content to be compressed prior to transmission and transcoded then. In such a scenario quality estimation based on quantizer information will reflect only artefacts that result from the final transcoding. In contrast, a metric based on image analysis detects a cumulative effect, i.e. all the artefacts actually seen by the user.

## 3. METRICS FOR VIDEO COMPRESSION ARTEFACT MEASUREMENT

This section presents metrics designed to measure specific video compression artefacts. Intra-frame compression is addressed by a blockiness metric; an improved flickering metric operating on a macro-block level is used for inter-frame compression artefacts; and finally a custom metric dedicated to I-frames flickering is proposed.

### 3.1. No-Reference Blockiness Metric

We used a common approach for calculating the blockiness artefact. It is calculated locally for each coding block. Absolute differences in pixel luminance were calculated separately for intra-pairs, represented by neighbouring pixels from a single coding block, and inter-pairs, represented by pixels from neighbouring blocks. A ratio between the total values of intra- and inter-differences is calculated over the entire video frame. For a real time application the metric should be calculated over a time window (the number of video frames). Mean value for the window represents a blockiness level $B$. For the purposes of the experiment the window size was equal to the sequence length (10 seconds). It was verified that the level of the blockiness artifact does not change significantly over time within the same video scene. Thus, any other window size or different method for temporal pooling would yield similar results.

### 3.2. No-Reference Flickering Metric

Our flickering metric was inspired by the work presented by Pandel in [6]. The task in our implementation was threefold. The first aim was to define the threshold used to decide whether a given macro-block remains in state of no-update. In [6] the threshold was defined as the mean squared difference between the pixels of the current and corresponding macro-blocks, although the exact value was not revealed. We calculated the threshold as an average of absolute differences in pixel luminance for each $16 \times 16$ macro-block. Second, we propose a different method for spatial pooling, i.e. calculate the frame-level flickering measure as a mean value over a small number of macro-blocks with the highest values (number of transitions between states). Third, we adjust two previous parameters in order to optimize prediction performance defined as a correlation with subjective scores. Similar to the blockiness metric, averaging over a time window is required, and for the purpose of the experiment the window size was equal to the sequence length (for the same reasons as in case of blockiness).

In order to maximize the correlation of the flickering metric $F$ with MOS we considered several threshold values (between 0.5% and 2% of luminance change) and several numbers of macro-blocks with the highest number of transitions between states (between 0.5% and 10% of macro-blocks). The highest correlation with MOS was achieved for the threshold equal to 1% and frame-level flickering averaging over 3% of the total number of macro-blocks.

### 3.3. No-Reference I Frame Flickering Metric

During a visual inspection of video sequences encoded with the H.264 codec another temporal artefact associated with H.264/AVC compression was identified. It can be defined as a flickering of the entire video frame whenever an I frame is decoded . In our case this means one flicker per second (FPS = 30 and GoP = 30). Sequences with a slow global motion and high spatial activity are especially vulnerable to this artefact [11]. For such sequences, strong compression imposes that the majority of coding blocks remains in the no-update state during the entire GoP structure. This results in a significant flickering whenever I frame arrives (suddenly all coding block are updated). For lower compression most of the coding blocks are updated (even several times) during the GoP structure and the effect disappears. It should be noted that this effect is inherent for GoP structures starting with one I frame, and should not be generalized over different schemes.

By analogy with the two-states model for flickering metrics described in Section 3.2, each decoded I frame activates the *update state* and causes visible global flickering. In contrast to P and B frames, all macro blocks are updated (intra-coded) on I frames even when strong compression is applied. We propose the following formula to calculate the I-frame flickering $IF$ effect:

$$IF = \text{mean}\left[\frac{SA_I(n)}{SA(n-1)}\right] \quad (1)$$

where $SA_I(n)$ and $SA(n-1)$ are spatial activities calculated for the I frame and the preceding one respectively (see Section 4 for details). $IF$ for the entire video sequence is calculated as a mean value over all pairs (I frame and the preceding one). As with both previous metrics, averaging over a time window is required, and for the purpose of the experiment the window size was equal to the sequence length. It was verified that this artifact is periodic and does not change significantly over time within the same video scene.

## 4. SUBJECTIVE EXPERIMENT

In order to analyze the influence of H.264/AVC compression on QoE we carried out a subjective experiment. The first step was to select a pool of test sequences. The key parameters describing any video sequence characteristics are spatial activity $SA$ and temporal activity $TA$, i.e. the number of details and the movement dynamics respectively. In order to make the selection task easier we used a *scene complexity* $o$ measure, which is a combination of $SA$ and $TA$ [12].

The scene complexity analysis resulted in choosing 13 source sequences from standard VQEG content, namely SRC 2, 3, 5, 7, 9, 10, 13, 14, 16, 18, 19, 20, and 21, with a scene complexity $o$ varying from 5.96 (SRC 21) to 8.45 (SRC 10).

We used the ACR-HR methodology, i.e. a no-reference subjective test. Original sequences were 10 seconds long at SD resolution and 30 FPS rate. We considered six different bit-rates (100, 200, 300, 500, 1000, and 4000 kbit/s) with a constant Group of Pictures (GoP) length equal to 30. We used x.264 coded, main profile, and the constant bit-rate mode for the rate-control.

Different bit-rate values were removed for different sequences in order to decrease the overall number of test sequences. Each sequence was scored by 25 subjects.

The sequences were displayed in the centre of a 17" LCD monitor with a native resolution of 1280x1024 pixels. The subjects started with a color blindness test. The sequences were then played out in random order. After the subject watched a sequence, he or she scored it using an eleven point discrete scale (see ITU-T P.910).

The experiment started with a training phase in order to familiarize subjects with the specificity of the test. The phase consisted of 8 sequences (selected from the main pool) covering the entire range of considered distortions. The answers obtained were not considered in the further work.

A post-experiment inspection of the subjective results was necessary in order to discard viewers who were suspected to give random answers. The rejection criteria (correlation coefficient $R^2$ lower than 0.75) verified the level of correlation of the scores of one viewer according to the mean score of all the subjects over the entire experiment.

## 5. ANALYSIS OF RESULTS

The goal of the result analysis phase is to propose a model mapping our metrics based on H.264/AVC compression artefacts assessment onto an eleven-point MOS scale. Statistically valid methodology for model derivation consists of several steps that are followed in the remaining part of the paper. For details please refer to [13].

Since we used an eleven point quality scale, the assumption that residuals of the subjects' scores have a Gaussian distribution is plausible, what allows us to use a linear regression for the modelling. This is an advantage when compared to five point MOS scales where the GLZ (Generalized Linear Model) should be used instead [13].

### 5.1. Data Sets

As a result of the performed subjective experiment we obtained 63 different MOS. The sequence pool was divided into

a *training set* (SRC 2, 5, 7, 10, 13, 16, 18, and 20) and a *verification set* (SRC 3, 9, 14, 19, and 21). The separation was carried out prior to any analysis of MOS scores. Both of these sets contain sequences covering a similar range of $o$ values, although the training set is larger than the verification set. Furthermore, sequences from the *verification set* cover the entire $o$ range evenly.

### 5.2. Preliminary Discussion of Results

MOS versus metric correlation obtained for the blockiness and flickering artefacts is presented in Fig. 1. It seems that the blockiness metric fails in cross content assessment. It is appealing that some points are (too) distant from the others (see Fig. 1(a)). Two sequences that stand out are SRC 18 Waterfall (green dots) and SRC 19 Football (red circles). Furthermore, the sequences stand out in opposite directions. In case of the flickering metric significantly higher correlated results were obtained (see Fig. 1(b)). This may suggest that an impact of content characteristics does not influence the results to the same degree as for the blockiness metric.

The SRC 18 sequence was rated lower by the testers than by the objective blockiness metric (Fig. 1(a)). In contrast, another artefact commonly referred to as flickering is dominant here. As described in the previous section, sequences with a slow camera movement encoded with H.264 are the most exposed to the flickering artefact. The sequence was rated slightly higher by the viewers than by the objective flickering metric (see Fig. 1(b)).

The second sequence that stands out from the mean is SRC 19. This time it was rated higher the by testers than by the objective blockiness metric (Fig. 1(a)). The answer to such discrepancy can be found in the analysis of spatial (amount of detail) and temporal (motion level) characteristics combined with the well-known masking theory. In our case, high spatial and very high temporal activities of the Football sequence are maskers to the blockiness artefact. It suggests perceptual weighting of the objective blockiness metric with regard to the spatial and temporal activity.

The presented analysis of results suggests that the perceptual impact of H.264 compression for diverse video content cannot be estimated properly using only one metric. The combination of the two metrics presented seems to be a much better solution in terms of correlation with MOS.

We decided to derive two models for single metrics (blockiness and flickering) and one integrated model for H.264/AVC compression. Because the I-frame flickering effect is restricted to low bit-rates only (strong compression) we consider it as an additional parameter of the integrated model rather than a stand-alone compression metric.

### 5.3. Model Based on the Blockiness Metric

Based on the preliminary discussion of results we assume that for the blockiness metric spatial and temporal activity will improve performance in terms of correlation with MOS. In order to verify this assumption we derive two models: 1) a basic model denoted as $MOS(B)$ based only on a single explanatory variable – the blockiness metric $B$, and 2) a complex model denoted as $MOS(B, SA, TA)$ including two additional explanatory variables, i.e. spatial activity $SA$ and temporal activity $TA$. The derived models are given by the following equations:

$$\text{MOS}(B) = -10.38 + 17.86B \tag{2}$$

$$\text{MOS}(B, SA, TA) = -10.88 + 14.68B + 0.02SA + 0.08TA \tag{3}$$

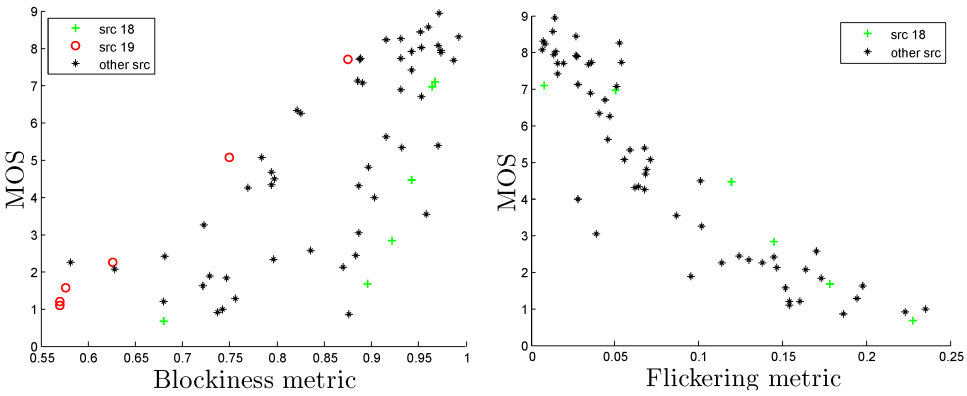
The following correlation coefficients were obtained for the $MOS(B)$ model: 1) $R_t^2 = 0.50$ for the training set, 2) $R_v^2 = 0.65$ for the verification set, and 3) $R_{t+v}^2 = 0.55$ for both sets at the same time. The complex model $MOS(B, SA, TA)$ achieved higher correlation, with statistical analysis revealing that both $SA$ and $TA$ are significant: 1) $R_t^2 = 0.66$, 2) $R_v^2 = 0.61$, and 3) $R_{t+v}^2 = 0.62$. The most meaningful correlation results are obtained using both sets at the same time while the verification set allows to determine whether a model is a general one. In case of a significant correlation drop for the verification set, a model should be considered as over-fitted to the training set. This is not the case for the blockiness metric, in particular if the most outstanding points were included in the verification set. Nevertheless, even correlation obtained for the complex model is not high enough to represent the overall quality for H.264/AVC compression.

### 5.4. Model Based on the Flickering Metric

The preliminary analysis of results did not significantly help in our understanding of the relation between content characteristics and the flickering metric. As a result, we decided to repeat the approach from the blockiness metric case. We derived two models: 1) a basic model denoted as $MOS(F)$ based only on a single explanatory variable – the flickering metric $F$, and 2) a complex model denoted as $MOS(F, SA, TA)$. Statistical analysis revealed that for the complex model neither was the correlation coefficient $R^2$ improved nor were $SA$ and $TA$ statistically significant. Therefore the final model consists of the flickering metric only. It suggests that either the flickering artefact in H.264/AVC compression or the flickering metric calculation methodology are content insensitive. The model is given by the linear equation:

$$\text{MOS}(F) = 7.68 - 33.61F \tag{4}$$

The following correlation coefficients were obtained for the $MOS(F)$ model: 1) $R_t^2 = 0.78$ for training set, 2) $R_v^2 = 0.94$ for verification set, and 3) $R_{t+v}^2 = 0.83$ for both sets.




(a) MOS as a function of the blockiness metric (b) MOS as a function of the flickering metric

**Fig. 1**. Preliminary correlation results.

As presented in Fig. 1(b), outstanding sequences were not included in the verification set. This explains why the correlation obtained for the verification set is so high. Correlation obtained for all sets is significantly higher than for the blockiness models and has a stronger potential in representing the overall quality for H.264/AVC compression.

### 5.5. Integrated Model

In order to verify the assumption that H.264/AVC compression yields blockiness artefact, flickering artefact, and the I-frame flickering effect at the same time, we derived an integrated model $MOS(B, F, IF)$, including all three parameters:

$$\text{MOS}(B, F, IF) = -14.55 + 6.33B - 26.22F + 16.72IF \quad (5)$$

The results are presented in Fig. 2, and for all sets they demonstrate that the model is very accurate ($R_t^2 = 0.89$, $R_v^2 = 0.90$, and $R_{t+v}^2 = 0.89$). All three model parameters are statistically significant and constitute the general model of perceptual evaluation of H.264/AVC compression.

We compared its performance with a well-known quality metric operating in a full-reference mode, choosing the Structural Similarity Index Metric (SSIM) [14]. The motivation for our choice was SSIM's availability, simplicity and good correlation with human perception. As presented by Wang [15], the human visual system (HVS) is very sensitive to the structural information provided on an image in a viewing field. Based on this assumption, SSIM can have good correlation with the perceptual quality in our case, since artefacts caused by H.264/AVC do destroy structural information.

Comparison of results between SSIM and the proposed model is presented in Fig. 2. The integrated model outperforms the SSIM metric in terms of correlation with MOS, for all sets. Correlation obtained for SSIM model was $R_t^2 = 0.81$, $R_v^2 = 0.88$, and $R_{t+v}^2 = 0.80$. Another advantage of the proposed model is no-reference approach, which significantly improves the application potential [1].

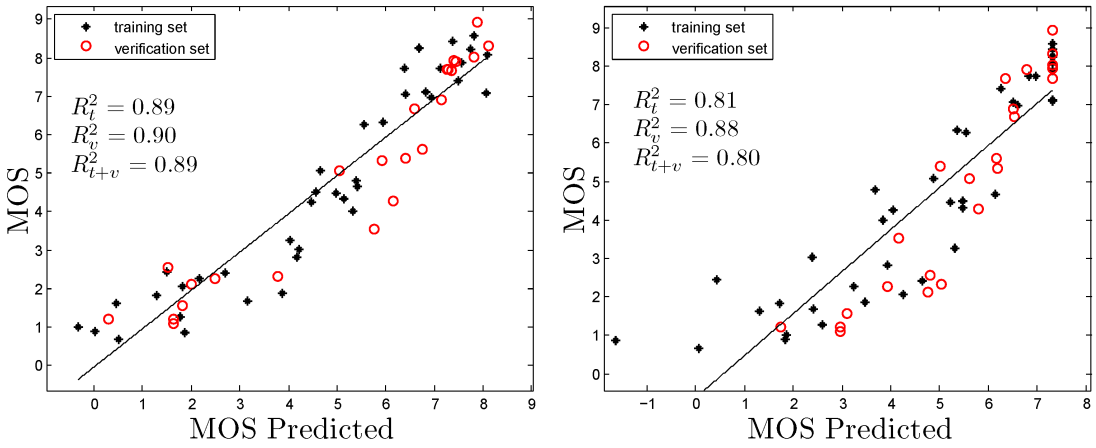
## 6. CONCLUSIONS

This paper describes a problem of video artefacts measurement related to intra- and inter-compression, using the H.264/AVC compression scheme as an example. Single no-reference metrics and integrated models were proposed to measure perceived video quality for the H.264/AVC compression. The models were derived and verified using subjective data. The high performance of the proposed models was demonstrated not only using correlation with MOS, but also by comparing it with a well-known full-reference metric. The metrics represent an image analysis approach and have relatively low computational requirements, which was evaluated experimentally. An important case study involving QoE assessment for live video streams was presented to realize a potential for implementation in a real environment.

We demonstrated that not only is the proposed integrated model well correlated with MOS, but also that our metrics are computationally light enough to fulfil real-time requirements. All metrics and content characteristics ($SA$ and $TA$) are calculated simultaneously using the same CPU core. Performance tests were carried out using 25 FPS videos in SD resolution, captured live from web cam or streamed from files. Metrics were implemented in a C/C++ environment using open video libraries available for Linux. The graphical interface was implemented using QT libraries.

## 7. ACKNOWLEDGMENT

The presented work was partly supported by the European Commission under the Grant INDECT No. FP7- 218086 and by the Polish Ministry of Science and Higher Education under the European Regional Development Fund, Grant No. POIG.01.01.02-00-045/09-00 Future Internet Engineering.




(a) $MOS(B, F, IF)$ (b) $MOS(SSIM)$

**Fig. 2**. Correlation with MOS of the integrated vs. SSIM models

## 8. REFERENCES


[1] S. Winkler, "Video Quality and Beyond," *in Proc. European Signal Processing Conference, Poznan, Poland*, September 3-7 2007.

[2] ITU-T, *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T, 1999.

[3] VQEG, *The Video Quality Experts Group*, `http://www.vqeg.org/`.

[4] Athanasios Leontaris and Amy R. Reibman, "Comparison of Blocking and Blurring Metrics for Video Compression," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 585 – 588, March 18-23 2005.

[5] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," *IEEE International Conference on Image Processing, ICIP 2005*, vol. 3, pp. III – 141–4, September 2005.

[6] Juergen Pandel, "Measuring of flickering artifacts in predictive coded video sequences," in *Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, Washington, DC, USA, 2008, pp. 231–234, IEEE Computer Society.

[7] J.E. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *Digital Video Image Quality and Perceptual Coding*, H.R. Wu and K.R. Rao, Eds. 2006, pp. 305–324, CRC Press.

[8] E. P. Ong, W. Lin, Zhongkang Lu, S. Yao, and M. H. Loke, "Perceptual quality metric for h.264 low bit rate videos," *IEEE International Conference on Multimedia and Expo*, pp. 677 – 680, July 2006.

[9] T. Brandao and M. P. Queluz, "No-reference quality assessment of h.264/avc encoded video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437 – 1447, November 2010.

[10] O. Sugimoto, S. Naito, S. Sakazawa, and A. Koike, "Objective perceptual video quality measurement method based on hybrid no reference framework," *IEEE International Conference on Image Processing (ICIP)*, pp. 2237 – 2240, Nov. 2009.

[11] Jie Xiang Yang and Hong Ren Wu, "Robust filtering technique for reduction of temporal fluctuation in h.264 video sequences," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 3, pp. 458 –462, march 2010.

[12] Charles Fenimore, John Libert, and Stephen Wolf, "Perceptual effects of noise in digital video compression," in *140th SMPTE Technical Conference*, Pasadena, CA, Oct. 1998, pp. 28–31.

[13] Lucjan Janowski and Zdzislaw Papir, "Modeling subjective tests of quality of experience with a generalized linear model," in *QoMEX 2009, First International Workshop on Quality of Multimedia Experience*, California, San Diego, July 2009.

[14] Z. Wang, L. Lu, and A. C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–13, 2004.

[15] Z. Wang, *Rate Scalable Foveated Image and Video Communications*, Ph.D. thesis, Dept. Elect. Comput. Eng. Univ. Texas at Austin, Austin, TX, December 2001.