No-Reference Perceptual Quality Assessment of Streamed Videos Using Optical Flow Features

Mohammed A. Aabed and Ghassan AlRegib
Center for Signal and Information Processing (CSIP)
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332, U.S.A.
{maabed, alregib}@gatech.edu

***Abstract*—This paper proposes a novel perceptual video quality assessment metric for streamed videos using optical flow statistical features. We analyze the impact of network losses on the decoded videos and the resulting error propagation. We show that the statistical features of the optical flow of the corrupted frames can be used to measure the distortion in the received video. We show that this approach is suitable for videos with complex motion patterns. Our technique does not make any assumptions on the coding conditions, network loss patterns or error concealment techniques. The proposed approach is pixel-based and relies only on the inconsistency of the optical flow of the corrupted frames. We validate our proposed quality metric by testing it on a variety of coded sequences subject to network losses from the recently proposed LIVE mobile database. Our results show that the proposed metric can estimate perceptual quality of channel-induced distortions at the frame and sequence levels. For the test videos, we report Pearson's and Spearman's correlation coefficients with the temporal mean opinion scores (MOSs) reported in the database. The results show average correlations of 0.91 and 0.92 for the test sequences, respectively.**

***Index Terms*—video quality monitoring, HEVC, H.264/MPEG-4 AVC, temporal distortion, video streaming, optical flow, network losses**

I. Introduction

With the rapid growth of Internet traffic and the increase of active mobile devices, the communication community's concern with bandwidth and quality of services/experience has consequently escalated. In 2013, global mobile data traffic grew 81% compared to 2012. Furthermore, video data accounted for 53% of total traffic by the end of 2013. In fact, it has been reported that cloud applications, such as Netflix, YouTube and Hulu, will grow 12-fold by 2018, which will make this type of traffic account for 90% of total global mobile data traffic [1]. Thus, the standardization bodies are adapting to this growth by motivating technologies that increase the efficiency of bandwidth utilization and data compression.

High efficiency video coding (HEVC) is the new video coding standard approved by the the Joint Collaborative Team on Video Coding (JCT-VC) [2]. It was also adopted by the the telecommunication standardization sector of the International Telecommunication Union (ITU-T) as its H.265 standard for video coding [3]. HEVC offers new coding features and tools to improve the compression gain of H.264/MPEG-4 AVC. In fact, it has double the coding efficiency of AVC [4], [3]. However, the complexity of the coding operations in HEVC is much higher compared to its predecessor. The coding unit tree (CTU) structure introduced in HEVC facilitates more efficiency for coding, transform and prediction. HEVC also applies an open group of picture (GOP) format which utilizes inter-coded frames more than AVC [4]. While this format facilitates higher compression gain, it imposes higher dependencies between the frames. This, in turn, makes an HEVC bitstream, and consequently decoded video, more sensitive to errors. All of these issues and others open new challenges in quality assessment, error concealment, etc. This paper analyzes the impact of network losses on the fidelity of the decoded video and proposes a new approach to measure the channel-induced distortion.

The problem of quality assessment for streamed video sequences has been recently addressed in several papers in the literature [5], [6], [7], [8], [9], [10]. De Simone *et al.* [5] report the performance of their subjective quality assessment campaign of the HEVC standard involving 494 test subjects. The authors in [6] test the performance of various full-reference (FR) quality metrics on 4k UHD videos. This work shows that PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM metrics were accurate in distinguishing different quality levels for the same content. In [8], the performance of MS-SSIM and the General VQM for high-definition videos is investigated. The authors show that MS-SSIM and VQM outperform PSNR on both HD and non-HD data and that MS-SSIM is slightly better on all databases. Wang *et al.* [7] combine subjective tests and objective analysis to propose a video quality assessment method. The paper uses a subject-objective mapping strategy with four indicators that affect video quality to set the relationship between the indicator and user experience. In [9], the authors utilize motion vectors, bit rate, and packet loss ratio to propose a video quality assessment algorithm. The authors in [10] have shown that video quality due to packet loss can be estimated at the decoder after concealment using a proposed reduced-reference (RR) approach that uses averages across superblocks of four macroblock parameters, along with the received motion vectors. The work herein addresses the objective no-reference (NR) perceptual quality assessment of streamed videos subject to network losses with access only to the decoded videos.

In this paper, we propose a NR optical flow based video

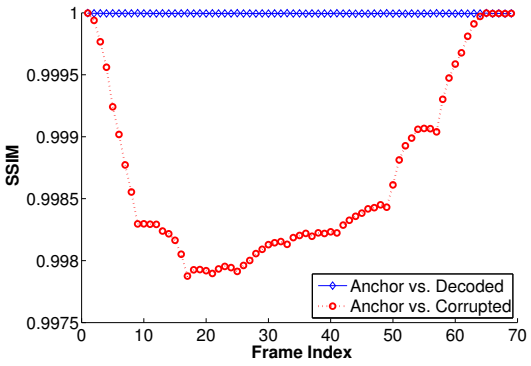

Fig. 1. The impact of loosing frame 8 on the SSIM values of the GOP for `Students Looming Across Street` sequence; frame rate is 60 fps.

quality assessment approach for streamed videos. First, we consider the coding conditions in AVC and HEVC and the impact of network losses on the decoded video under these conditions. We shows that the error propagation in HEVC videos, for instance, is more significant than the propagation in AVC videos. Then, we introduce a perceptual quality metric based on analyzing and understanding the statistical features of the optical flow in the received video. This approach does not make any assumptions on the concealment technique, network conditions or coding standard or parameters. It blindly operates on the decoded video after the decoder. We argue that the change in the optical flow due to channel-induced losses can be used to capture the distortion in the frames.

The rest of this paper is organized as follows. In Section II, we illustrate the significant impact of network losses on a GOP structure in terms or temporal error propagation. We then explain our approach to measure the channel induced distortions, which utilizes the statistical features of the optical flow. Section III details the simulations setup and test sequences used in the experiments herein, followed by the results and analysis of the model validation experiments. Finally, Section IV concludes the paper and outlines future directions of this work.

II. Optical Flow Based Video Quality Assessment

In this section, we first discuss the coding structure of AVC and HEVC. We study the impact of network errors or losses under these coding conditions which introduces the motivation behind this work. We then explain our proposed perceptual video quality metric and the intuition behind it. This approach operates only on the decoded video without making any assumptions about the encoding configurations, error concealment strategy or network conditions.

A. Error Propagation in A GOP Configuration

Both AVC and HEVC employ a group of picture (GOP) operation where an intra-coded frame (I-frame) indicates the beginning of a GOP. Afterwards, several inter-coded frames (P- and B-frames) follow and the video stream consists of a set of successive GOPs. In contrast to AVC, HEVC employs an *open* group of picture (GOP) operation. In this coding

Fig. 2. The optical flow maps of frame 240 of `Students Looming Across Street` sequence. The top corresponds to the correctly decoded frame and the bottom corresponds to the distorted version.

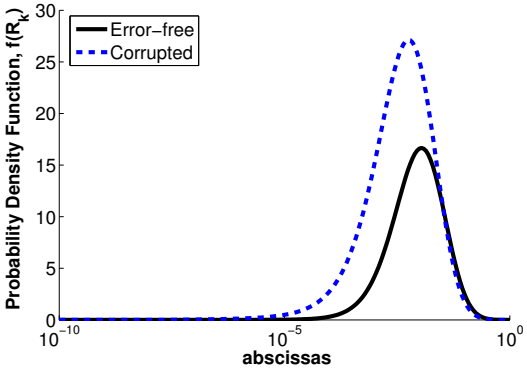

Fig. 3. The empirical probability density functions of $\mathbf{R}_{240}$ of the corrupted and error-free `Students Looming Across Street` sequences. The SSIM value for the distorted frame is 0.998.

configuration, a new clean random access (CRA) picture syntax is used, wherein an intra-coded picture is used at the location of random access point (RAP) to facilitate efficient temporal coding [4]. The intra period varies depending on the frame rate to introduce higher compression gain [12]. This format has been also implemented in some AVC encoders. Hence, HEVC utilizes inter-coding more than AVC. Thus, the data dependency between the frames is higher in HEVC. Consequently, the impact of channel-induced errors on certain frames that potentially propagate to the end of the GOP is more significant in HEVC than in AVC. Fig. 1 shows an example of the impact of loosing the Network Abstraction Layer (NAL) unit corresponding to frame 8 and replacing it with the temporally closest available frame at the decoder, which is frame 0 in this example. In our simulations and tests, we abide by the recommended encoding format wherein every

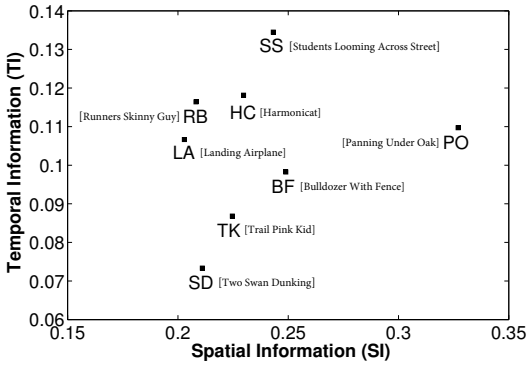

Fig. 4. Spatial information (SI) versus temporal information (TI) indices for the selected sequences [11].

frame is taken as a single slice which is encapsulated in a separate NAL unit [13]. Fig. 1 shows that the channel loss under these coding conditions propagates until a new I-frame is encountered, which is frame 64 in this example.

Under the assumption that we do not have access to the decoder and we only have access to the decoded sequences as explained in Section I, we do not have any knowledge of how losses have propagated to other frames. Hence, in order to estimate these distortions without any FR information, we can only rely on the spatial and temporal features of the decoded video.

B. Optical Flow Based Perceptual Quality Assessment

In this section, we explain our proposed methodology for video quality assessment using the optical flow. The proposed approach is based on the fact that any channel induced distortion will cause a temporal inconsistency in the optical flow. The Human Visual System (HVS) observes the distortion in the form of visual discomfort due to inconsistency in the pixels or objects in the distorted frame. Let f_k be the frame of interest. Furthermore, let $\mathbf{U}_k$ and $\mathbf{V}_k$ denote the matrices of its horizontal and vertical optical flow velocities, respectively. Furthermore, let $\mathbf{R}_k$ denote the matrix of magnitudes of the flow velocities [14]:

$$\mathbf{R}_k = \sqrt{\mathbf{U}_k^2 + \mathbf{V}_k^2} \tag{1}$$

where k is the temporal index of the frame in the received video. Fig. 2 shows the visualizations of the optical flow maps of a correctly decoded frame and a distorted one, respectively. Furthermore, Fig. 3 shows the probability density functions (PDFs) of these optical flows, $\mathbf{R}_k$, for the error-free and corrupted frames. The figures show that there is a discrepancy in the optical flow due to distortion in the frame. It should be noted that the SSIM values of the corrupted frame in this case is 0.998. Our goal is to capture these inconsistencies in the optical flows throughout the video due to the channel-induced losses. All the results and experiments in this paper were obtained using the Horn-Schunck optical flow estimation method [14]. This approach, nevertheless, is valid for any optical flow estimation algorithm.

Next, in order to calculate the inconsistencies in the optical flow map for a certain frame, we use the diffusion distance as a dissimilarity metric [15]. We utilize the increment in the diffusion distance of the optical flow when a frame is corrupted to measure the distortion. We calculate the diffusion distance, $\delta(\cdot)$, of the optical flow matrix, $\mathbf{R}_k$, iteratively:

$$T(\mathbf{R}_k, t) = \text{Downsample}_{\frac{1}{2}}\left[T(\mathbf{R}_k, t-1) * \phi(\alpha)\right] \tag{2a}$$

$$\delta(\mathbf{R}_k, t) = \delta(\mathbf{R}_k, t-1) + \Sigma_{\forall y}\Sigma_{\forall x}|T(\mathbf{R}_k, t)| \tag{2b}$$

where t is the iteration index, $T(\mathbf{R}_k, t)$ is the downsampled version of $T(\mathbf{R}_k, t-1)$ after filtering, and $\phi(\alpha)$ is a Gaussian kernel with standard deviation α. It should be noted that $T(\mathbf{R}_k, 0) = \mathbf{R}_k$ and $\delta(\mathbf{R}_k, 0)$ is initialized to $\Sigma_{\forall y}\Sigma_{\forall x}|\mathbf{R}_k|$. For simplicity, we refer to the final $\delta(\mathbf{R}_k, t)$ as $\delta(\mathbf{R}_k)$. In our implementation, we begin with a downsampled version the original $\mathbf{R}_k$ to facilitate computational efficiency for 720p sequences and higher resolutions. The number of iterations for every frame is relatively low (1-3 iterations) which makes this approach feasible for real-time implementations. Then, the estimated perceived quality of frame k, D_k, is given by:

$$D_k = k \cdot \exp\left[-\left(\frac{k}{\delta(\mathbf{R}_k)} + \frac{k-1}{\delta(\mathbf{R}_{k-1})}\right)\right] \tag{3}$$

Close inspection of the expression in (3) shows that $\delta(\mathbf{R}_k)$ is inversely proportional to the perceptual quality in the logarithmic domain. This translates into that the higher the diffusion distance of the optical flow matrix, the higher the distortion. Furthermore, the exponent does not rely only on frame k, it rather takes into accounts frame $k-1$. We introduce this term to take into consideration the previous frame's perceptual impact since the HVS does not evaluate the quality of each frame independently.

III. Experiments and Results

All the experiments and tests in this paper were done on the recently proposed LIVE Mobile Video Quality Assessment database [16], [17], [18]. We used eight different video sequences in our experiments. All the video sequences were coded using the H.264 scalable video codec (SVC) standard. All the tested sequences have 450 frames and the same resolution of 1280×720. The corrupted sequences are taken from the wireless channel packet-loss set in the database. Next we detail the coding parameters and the obtained results.

All the sequences were coded using the JM software implementation of the H.264 scalable video coding (SVC) [19], [20], [21]. The videos were encoded with a GOP period of 16 and packet-loss rates set to 3% for the reference decoded sequences. As per the recommendation for wireless transmission [21], the packet size was 200 bytes. The coded bitstreams were transmitted over a simulated wireless channel in order to induce losses which would impact the perceptual quality. For the details and configurations of these simulation, we refer the reader to [18], [21].

We tested the proposed metric on 8 video sequences from the LIVE mobile database. We diversify the selection of test sequences to span different temporal and spatial complexities.

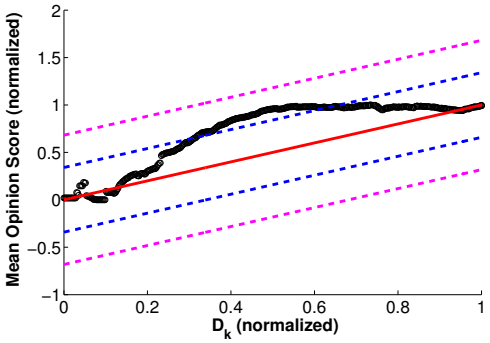

Fig. 5. The proposed no-reference perceptual quality metric scatter plot versus the reported MOS for `Runners Skinny Guy` sequence. The blue and pink lines are $D_k \pm \sigma$ and $D_k \pm 2\sigma$, respectively, where σ is the data standard deviation.

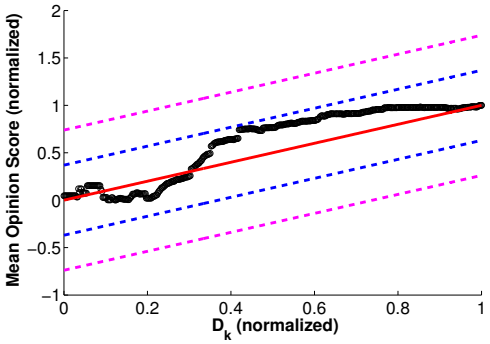

Fig. 6. The proposed no-reference perceptual quality metric scatter plot versus the reported MOS for `Panning Under Oak` sequence. The blue and pink lines are $D_k \pm \sigma$ and $D_k \pm 2\sigma$, respectively, where σ is the data standard deviation.

Fig. 4 shows the spatial information (SI) and temporal information (TI) indices on the luminance channel for the selected sequences, as per the recommendation in [11]. The higher the score on the SI and the TI scale, the more complex the spatial and temporal features of the test sequence.

A. Results and Analysis

Fig. 5-6 show the scatter plots for `Runners Skinny Guy` and `Panning Under Oak` sequences, respectively. From the two plots, we notice that the value of D_k is highly correlated with the temporal mean opinion scores (MOSs) reported in the database [18]. Clearly, the relationship is not ideally linear. Nevertheless, the black points on the scatter plot form a smooth curve, which indicates that the proposed metric captures the MOS temporal pattern. Notably, *these results were obtained by operating blindly on the decoded pixels without any auxiliary information about the sequence or bitstream*. Furthermore, we did not perform any curve fitting or apply any accuracy enhancement techniques.

In order to validate the proposed distortion model, we calculate the Pearson's and Spearman's correlation coefficients between the estimated perceptual quality based on the proposed approach and the measured temporal MOS of the corrupted sequences. Tables I-II summarize the results for all the tested sequences for Pearson's correlation coefficients (PCC) and Spearman's correlation coefficients (SCC), respectively. Furthermore, the authors in [18] reported that the highest correlation coefficients for FR approaches with MOSs were obtained using the Noise Quality Measure (NQM) [22] and Visual Information Fidelity (VIF) metrics [23]. Hence, we report the correlation coefficient for these two full-reference metrics at the sequences level for comparative analysis. We note that the proposed model correlates well with the temporal MOS values. Pearson's correlation coefficients for all test sequences range between 0.86 and 0.95 with an average of 0.91. In particular, the proposed approach works well for the sequences with high temporal complexity such as the `Students Looming Across Street` and `Bulldozer With Fence` video sequences. Furthermore, our quality metric works well for sequences with medium or low temporal complexity, such as `Two Swan Dunking` and `Harmonicat`. These results show the effective utility of relying on the optical flow in analyzing the distortion in videos with complex motion patterns where the temporal feature across frames are highly variant.

Sequences	Proposed (NR)	NQM [22] (FR)	VIF [23] (FR)
`Two Swan Dunking`	0.8961	0.9922	0.9922
`Runners Skinny Guy`	0.8684	1.0000	1.0000
`Students Looming Across St.`	0.9495	0.9983	0.9983
`Bulldozer With Fence`	0.9448	0.9955	0.9988
`Panning Under Oak`	0.9350	0.9999	0.9999
`Landing Airplane`	0.9178	1.0000	1.0000
`Trail Pink Kid`	0.9135	0.9738	0.9738
`Harmonicat`	0.8565	0.9979	0.9979

TABLE I
PEARSON'S CORRELATION COEFFICIENTS (PCC)

Sequences	Proposed (NR)	NQM [22] (FR)	VIF [23] (FR)
`Two Swan Dunking`	0.9274	0.9964	0.9964
`Runners Skinny Guy`	0.8384	1.0000	1.0000
`Students Looming Across St.`	0.9921	0.9944	0.9944
`Bulldozer With Fence`	0.9457	0.9955	0.9955
`Panning Under Oak`	0.9749	0.9999	0.9999
`Landing Airplane`	0.9317	1.0000	1.0000
`Trail Pink Kid`	0.9451	0.9715	0.9715
`Harmonicat`	0.8135	0.9941	0.9941

TABLE II
SPEARMAN'S CORRELATION COEFFICIENTS (SCC)

IV. CONCLUSION

We propose in this paper a no-reference perceptual video quality metric to estimate the channel-induced distortion due to network losses. The proposed technique does not make any assumption about the coding conditions or video sequence. It rather explores the temporal changes between the frames by analyzing the variations in the statistical properties of the optical flow. We validate our approach by testing it on various sequences and compare our estimated quality metric with the temporal MOS values reported in the LIVE mobile database for sequences subject to network errors or losses. Our experiments show that the proposed technique captures the perceptual quality very well.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 20132018," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html, accessed: 2014-06-01.

[2] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, Y.-K. Wang, and T. Wiegand, "High efficiency video coding (hevc) text specification draft 10 (for fdis & final call)," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-L1003.v34*, jan 2013.

[3] ITU-T, "H.265: High efficiency video coding," ITU Telecommunication Standardization Sector, Tech. Rep., april 2013.

[4] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[5] F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *J. Vis. Comun. Image Represent.*, vol. 22, no. 8, pp. 734–748, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2011.01.008

[6] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *2013 18th International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–8.

[7] W. Wang, Y. Jin, T. Yang, and Y. Cui, "A video quality assessment method using subjective and objective mapping stategy," in *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on*, vol. 02, Oct 2012, pp. 514–518.

[8] S. Wulf and U. Zolzer, "Full-reference video quality assessment on high-definition video content," in *Signal Processing and Communication Systems (ICSPCS), 2012 6th International Conference on*, Dec 2012, pp. 1–10.

[9] B. Konuk, E. Zerman, G. Nur, and G. B. Akar, "A spatiotemporal no-reference video quality assessment model," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 54–58.

[10] L. McLaughlin and S. S. Hemami, "Reduced-reference quality assessment with scalable overhead for video with packet loss," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 1622–1626.

[11] ITU-T, "P.910: Subjective video quality assessment methods for multimedia applications," ITU Telecommunication Standardization Sector, Tech. Rep., 2008.

[12] F. Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-K1100*, Shanghai, China, 11th meeting, oct 2012.

[13] F. Bossen, D. Flynn, and K. Sühring, *HM 12.1 Software Manual*, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-Software Manual, may 2013.

[14] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–204, 1981.

[15] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 246–253.

[16] A. Moorthy, L. K. Choi, G. de Veciana, and A. Bovik, "Subjective analysis of video quality on mobile devices," in *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, jan 2012.

[17] ——, "Mobile video quality assessment database," in *IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks*, vol. 6, no. 6, Oct 2012, pp. 652–671.

[18] A. Moorthy, L. K. Choi, A. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, Oct 2012.

[19] "Svc reference software (jsvm software), joint video team (jvt)," http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm, accessed: 2014-06-01.

[20] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, Sept 2007.

[21] T. Stockhammer, M. Hannuksela, and T. Wiegand, "H.264/avc in wireless environments," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 657–673, July 2003.

[22] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *Image Processing, IEEE Transactions on*, vol. 9, no. 4, pp. 636–650, Apr 2000.

[23] H. Sheikh and A. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, Feb 2006.