



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

No Reference Video Quality Evaluation for Multimedia Applications

Hajera Siddiqua, and Z. M. Parvez Sazzad

Department of Applied Physics, Electronics and Communication Engineering,
University of Dhaka, Dhaka-1000, BANGLADESH.

Abstract: *Perceptual video quality evaluation plays important roles in multimedia applications, such as video teleconferencing, video on demand and Internet streaming media. Most of the traditionally objective video quality assessment attempted to quantify the predefined artifacts of the coded video sequences. In this paper, we present a new no reference perceptual video quality evaluation model based on spatial and temporal features irrespective of any predefined artifacts in Multimedia applications. The proposed spatial features are pixel distortion and edge information measurement of each frame and the temporal features measure are effective temporal information, effective frame and frame similarity measurement of the sequence. Subjective experiments results are used to verify the model, and show that the proposed model performances are reasonably close to the subjective scores.*

Keywords: *No reference, Multimedia, Video quality, H.264, WMV 9*

I. Introduction

Due to increasing transmission of multimedia (MM) contents over the Internet and 3G mobile networks, video quality evaluation has become an important issue. There is no doubt that subjective test is the most accurate method for the quality evaluation because it reflect true human perception. Moreover, human is the ultimate user in all cases. But it's time consuming and expensive. Furthermore, it cannot be done in real time. As a result, nowadays, developing objective video quality evaluation methods are getting more attraction. There are three types of method that are used for objective video quality evaluation, full-reference (FR), reduced-reference (RR) and no-reference (NR). In full-reference method, a reference / original video is required to assess the quality of the distorted video, think like that the loss of quality of the distorted video arises as a result of some processing on the reference video. Therefore it's highly desirable to develop quality assessment method that doesn't require full access to the reference video. In reduced-reference method, some extracted features of reference / original video are required to assess the quality. But in no-reference method, no reference / original video are required to assess the quality. This NR method has recently received a great attention because the reference video is not available in many practical applications or may be too expensive to provide. Designing objective No-Reference (NR) quality measurement algorithms is a very difficult task. This is mainly due to the limited understanding of HVS, and it is believed that effective NR quality assessment is feasible only when the prior knowledge about the video distortion types is available.

Several objective video quality assessment methods are recommended by the Video Quality Expert Group (VQEG) of ITU that are applicable to the evaluation of MPEG-2 coded video in conventional broadcasting applications, and the methods are standardized in ITU-T Rec. J.144 [1]. FR video quality measure based on peak signal-to-noise ratio is proposed in [2], but this type of model performance based on PSNR and SNR are poorly correlated with human quality prediction [3]. In [4] a FR objective video quality assessment method is proposed for MM based on weighted sum of average edge energy, spurious edge generation and motion distortion. But it consider only different bit rate with different coders. However video quality assessment for MM applications must be evaluated with different compression and transmission errors of different video coders. In [5], an objective method is proposed that takes into account of the artifacts introduced by spatial and temporal activities in the hybrid block based coding methods. It mainly calculates the blockiness, blur and jerkiness features. The work in [6], proposed an estimation of the pattern of lost macro blocks which produces an accurate estimate of the mean-square-error (MSE) distortion introduced by channel errors. In [7], the authors try to review the basic background knowledge necessary to design an efficient no-reference video or image quality evaluation method. In this research, we propose an NR video quality assessment in which the sequences are coded by two different coders with wide variety of bit rate and frame rate for MM applications.

II. Subjective Experiments

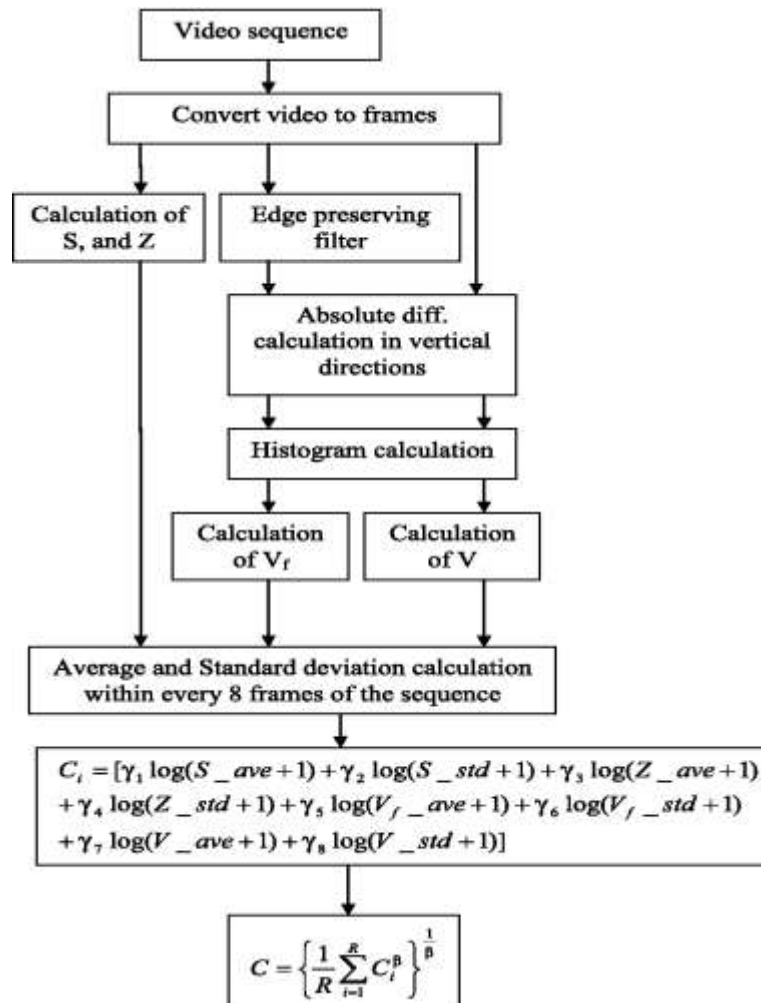
The subjective experiments were conducted on true color video to evaluate MM video quality. Twenty subjects were participated to conduct the subjective test, most of them were student. The subjects were asked to provide their perception of quality on a discrete quality score that was divided into five and marked with the numerical value of adjectives "Bad =1," "Poor=2," "Fair=3," "Good=4," and "Excellent=5" under the test conditions of ITU-R Rec. 500-10 and ITU-T P.910 [8], [9]. The Single Stimulus Absolute Category Rating method with

hidden reference removal (ACR-HRR) was selected to conduct the subjective test. We considered twenty video sequences of each length of eight seconds with two different coders and three different frame rate and bit rate. Therefore, we had total 420 video sequences (400 coded + 20 references) in the experiment. All subjective test conditions and coding parameters are summarized in Table 1. The twenty scores of each video were averaged to get a final Mean Opinion Score (MOS) of the video with subject reliability of 95% confidence interval.

Table 1: Subjective test conditions and parameters

Method	ACR-HRR
Format	VGA (640×480)
Color space	YUV (UYVY)
Coder	WMV 9, and H264 (MPEG-4 part 10)
Coding parameters	Bit rate and Frame rate
Bit rate	448Kbps, 1024Kbps, and 4096Kbps
Frame rate	5fps, 10fps, and 30fps
Display	LCD 19-inch EIZO
Viewing distance	4H (H: Picture Height)
Ambient illumination	About 200 lux

Figure 1: Spatial features estimation process.



III. Proposed Model

Many researchers have already been established that the main function of the human visual system is to extract structural or edge information from the viewing field and the human visual system is highly adapted for this purpose [10]. Under the assumption that human visual perception is very sensitive to edge information and natural image/ video signals are highly structured that is the samples of the signals have strong dependencies

between each other, especially when they are close in space. That is any kinds of spatial artifacts create pixel distortions from neighbor pixels. Therefore in this research, we want to develop a new NR video quality assessment model based on spatial features of pixel distortions and edge information and temporal features of effective temporal and frame information and also similarity measure within the frames of the sequence. In all calculation, first we convert the video sequences to frames and consider only luminance part of the frame for simplicity.

A. Spatial Features Measure

A.1. Pixel Distortion Measure

Pixel distortion is estimated by the standard deviation of a central pixel within 5×5 neighborhood pixels which is applied for all available pixels in the frame. And then average the standard deviation values within 5×5 partially overlapping block. Block diagram of spatial features evaluation process is shown in Fig. 1. Let X_{13} is the central pixel of the 5×5 block that is shown in Fig. 2, and let $\bar{X}$, S_{std} , and $\bar{S}_{std}$ be the mean of pixels within the block, the standard deviation of X_{13} pixel in the block, and the average standard deviation within 5×5 partially overlapping block. The statistical features can be estimated as follows:

$$\bar{X} = \frac{1}{L} \sum_{i=1}^L X_i \quad (1)$$

where L is the total number of pixels in the block.

$$S_{std} = \sqrt{\frac{1}{L-1} \sum_{i=1}^L (\bar{X} - X_i)^2} \quad (2), \quad \text{and} \quad \bar{S}_{std} = \frac{1}{L} \sum_{i=1}^L S_{std} \quad (3)$$

Figure 2: 5×5 Pixel block; Central pixel, X_{13} of the block.

X_1	X_2	X_3	X_4	X_5
X_6	X_7	X_8	X_9	X_{10}
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}

Figure 3: Masking block of the edge preserving filter.

	I	
K	X	L
	J	

If the total row and column of the frame are respectively M and N, then finally, the 5×5 partially overlapping standard deviation feature, S is estimated by the following equation.

$$S = \frac{1}{\left(\frac{M-3}{4}\right)\left(\frac{N-3}{4}\right)} \sum_{i=1}^{\left(\frac{M-3}{4}\right)\left(\frac{N-3}{4}\right)} S_{std_i} \quad (4)$$

A.2. Edge Information Measure

Edge information is estimated using two features. First, zero-crossing (ZC) rate is estimated both in horizontal and vertical direction of the image. We denote the test frame signal as $x(m, n)$ for $m \in [1, M]$ and $n \in [1, N]$ and calculate a differencing signal along each horizontal line:

$$d_h(m, n) = x(m, n+1) - x(m, n); \quad n \in [1, N-1] \quad (5)$$

For horizontal ZC:

$$d_{h-sign}(m, n) = \begin{cases} 1 & \text{if } d_h(m, n) > 0 \\ -1 & \text{if } d_h(m, n) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$d_{h-mul}(m, n) = d_{h-sign}(m, n) \times d_{h-sign}(m, n+1) \quad (7)$$

$$z_h(m, n) = \begin{cases} 1 & \text{if } d_{h-mul}(m, n) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Then the horizontal zero-crossing rate, Z_{bh} is calculated within 5×5 partially overlapping block and then the average horizontal zero-crossing rate, Z_h is estimated as follows:

$$Z_{bh} = \frac{1}{25} \sum_{i=1}^5 \sum_{j=1}^5 z_h(i, j) \quad (9)$$

$$Z_h = \frac{1}{\left(\frac{M-1}{4}\right)\left(\frac{N-3}{4}\right)} \sum_{i=1}^{\left(\frac{M-1}{4}\right)\left(\frac{N-3}{4}\right)} Z_{bh_i} \quad (10)$$

Similarly we can calculate the average vertical zero-crossing rate, Z_v and finally the overall feature of zero-crossing rate is given by

$$Z = \frac{Z_h + Z_v}{2} \quad (11)$$

The second edge information measure is the histogram measure with and without edge preserving filter (Figure 3). The edge preserving filtering algorithm is calculated by the following concept.

$$H_d = K - 2X + L \quad (12)$$

$$V_d = I - 2X + J \quad (13)$$

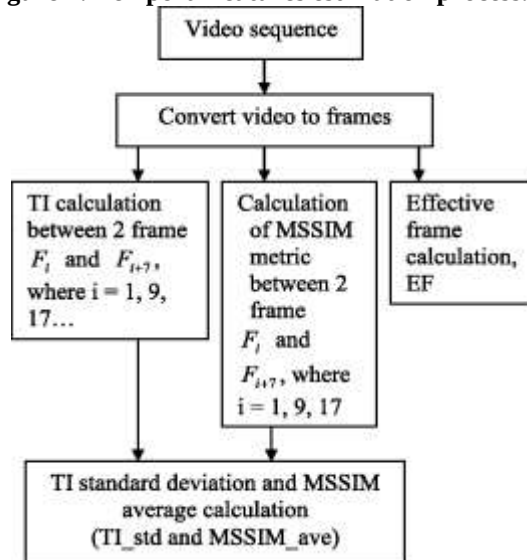
$$\begin{aligned} \text{if } (H_d < V_d), \quad X &= (K + 2X + L) / 4 \\ \text{esle } X &= (I + 2X + J) / 4 \end{aligned} \quad (14)$$

where X is the central pixel and the I, J, K and L are the four closest pixels of its that are shown in Fig. 3. With and without applying the edge preserving filter, the absolute difference calculation is estimated between two neighbor pixels value in vertical direction. Then we calculate the histogram features and observe that most of the histogram values are in the lowest pixels amplitude. Here we consider only three lowest absolute difference pixel amplitude of 0, 1 and 2 to get the major information of the frame. Let v_{f0}, v_{f1}, v_{f2} and v_0, v_1, v_2 respectively be the number of absolute difference amplitude pixels with and without edge preserving filter that have been lied on the position of 0, 1 and 2 on the histogram and also let V_f and V are respectively the vertical histogram features of the frame of size $M \times N$ with and without filter, then the vertical histogram features can be estimated as follows:

$$V_f = \frac{(v_{f0} + v_{f1} + v_{f2})}{(M - 2) \times (N - 2)} \quad (15)$$

$$V = \frac{(v_0 + v_1 + v_2)}{M \times N} \quad (16)$$

Figure 4: Temporal features estimation process.



B. Temporal features measure

Statistics based a set of features can be defined for temporal changes to the frame pixels. Theses temporal statistics are indicated the amount of temporal change or motion in the video sequences. Block diagram of temporal estimation process is shown in Fig. 4.

B.1. Temporal features measure

To compute the effective temporal information at time t_n for pixel $x(m, n, t_n)$, consider frame $x(t_n)$ and another frame that is eight time earlier in time. The absolute difference between those frame, pixel by pixel and also the statistical features of mean and standard deviation are calculated. Let TI , $\overline{TI(m, n, t_n)}$ and $TI_{frame-std}$ be the absolute temporal difference between the frames, average absolute temporal difference and standard deviation of the temporal difference of the frame that are estimated by the following equations:

$$TI(m, n, t_n) = |x(m, n, t_n) - x(m, n, t_{n-8})| \quad (17)$$

where $n = 1, 9, 17, \dots$

$$\overline{TI(m, n, t_n)} = \frac{1}{M \times N} \sum_{m=1}^M \sum_{n=1}^N TI(m, n, t_n) \quad (18)$$

$$TI_{frame-std} = \sqrt{\frac{1}{L-1} \sum_{i=1}^L (\overline{TI} - TI_i)^2} \quad (19)$$

where $L = M \times N$ is the total number of pixels in the frame. The effective TI measure, TI_{std} is calculated of the video sequence by the following equation.

$$TI_{std} = \sqrt{\frac{1}{P-1} \sum_{i=1}^P (\overline{TI_{frame-std}} - TI_{frame-std_i})^2} \quad (20)$$

where P is the number of difference point in the time interval of the video sequence.

B.2. Number of effective frame measure

Number of effective frame of the sequence is defined as the number of individual frame compare to the next consecutive frame of the sequence. That is every time we calculate pixel-by-pixel absolute difference between one frame and the next consecutive frame and then calculate the sum of the differences. If the sum is non zero the frame is affective otherwise not. The estimation is calculated by the followings:

$$TI_{eff}(m, n, t_n) = \sum_{i=1}^{M \times N} |x(m, n, t_n) - x(m, n, t_{n+1})| \quad (21)$$

$$\left. \begin{array}{l} EF = 0 \\ \text{if } (TI_{eff} \neq 0) \\ EF = EF + 1 \\ \text{else } EF = EF \end{array} \right\} \quad (22)$$

Here EF is the number of effective frames per sequence. Initially, it's set to zero and after execution of the above algorithm, the number of effective frames, EF of the sequence is calculated.

B.3. Similarity measure

Structural similarity or distortion measure of a frame compare to the eight time later frame in time of the sequence is defined how similar a frame to the next frame that are calculated by Mean Structural Similarity Metric (MSSIM) [11]. And then the average similarity measure, $MSSIM_{ave}$ is calculated of the video sequence by the following equation.

$$MSSIM_{ave} = \frac{1}{P} \sum_{i=1}^P MSSIM_i \quad (23)$$

where P is the number of point in the time interval of the video sequence.

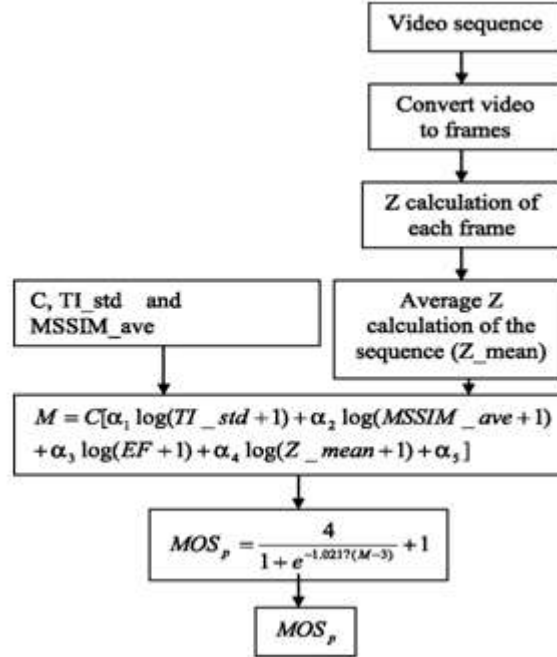
IV. Features Combination of the Proposed Model

Let S_{ave} , S_{std} , Z_{ave} , Z_{std} , V_{fave} , V_{fst} , V_{ave} and V_{std} are the respectively average and standard deviation of pixel distortion, zero crossing rate, and vertical histogram features with and without filter within every eight frames of the sequence. The features are combined by the following equation and block diagram of the proposed model is shown in Fig. 5.

$$C_i = [\gamma_1 \log(S_{ave} + 1) + \gamma_2 \log(S_{std} + 1) + \gamma_3 \log(Z_{ave} + 1) + \gamma_4 \log(Z_{std} + 1) + \gamma_5 \log(V_{ave} + 1) + \gamma_6 \log(V_{std} + 1) + \gamma_7 \log(V_{ave} + 1) + \gamma_8 \log(Z_{std} + 1)] \quad (24)$$

In the combine equation, to avoid “log(1)”, we introduce additional one to all features. The Minkowski weighting function is used to calculate the total spatial assessment of the sequence by the following equation:

Figure 5: Proposed Video Quality Evaluation Model.



$$C = \left[\frac{1}{R} \sum_{i=1}^R C_i^\beta \right]^{1/\beta} \quad (25)$$

Here γ_1 to γ_8 and β are the spatial model parameters that must be estimated with the subjective test data. The estimation is performed by Particle Swarm Optimization (PSO) algorithm [12]. Finally video sequence temporal features TI_{std} , $MSSIM_{ave}$, and EF and spatial feature and assessment, Z_{mean} (average Z calculation of the sequence) and C are combined by the following equation:

$$M = C[\alpha_1 \log(TI_{std} + 1) + \alpha_2 \log(MSSIM_{ave} + 1) + \alpha_3 \log(EF_{ave} + 1) + \alpha_4 \log(Z_{mean} + 1) + \alpha_5] \quad (26)$$

where α_1 to α_5 are the model parameters. The parameters estimation are performed by PSO algorithm. We consider a logistic function as the non linearity property between the human perception and the physical features. Finally, obtained the MOS prediction score, MOS_p is derived by the following equation [13].

$$MOS_p = \frac{4}{1 + \exp[-1.0217(M - 3)]} + 1 \quad (27)$$

V. Results

In order to verify the performance of my proposed model extensively I considered the reference video sequences which are most used in VQEG MM test plane [14]. Consequently, the video database was created according to the test condition and parameters which is shown in Table 1. The subjective experiment details are explained in Section 2. The database consists of four hundred and twenty video sequences. Out of four hundred coded and twenty reference sequences. In order to use the database, I randomly divided the database into two halves one training and another for testing. However, there is no overlapping between training and testing. The model's parameters are estimated by the PSO optimization algorithm with the random division of the database into two halves in every pass [12]. The model's parameters are shown in Table 2. In order to provide quantitative measures on the performance of the proposed model, we will follow the standard performance evaluation procedures employed in the video quality experts group (VQEG) FR-TV Phase II test [15], where mainly Pearson linear correlation coefficient (CC), Root mean square prediction error (RMSE), Rank order correlation coefficient (ROC) and Outlier Ratio (OR) between objective (mean opinion score prediction, MOS_p and subjective scores (mean opinion score prediction, MOS) were used for evaluation. The evaluation results are summarized in Table 3. The Table shows that the proposed model performances are sufficient for every one of

the evaluation metrics. It has also been observed from the Tables 3, that the proposed model provides sufficient prediction accuracy (higher CC, and lower RMSE), sufficient monotonicity (higher ROC) and sufficient prediction consistency (lower OR). The MOS versus MOS_p predictions (MOS_p) of the proposed model is shown in Fig. 6 for all video sequences. The Fig. 6 shows that the proposed model prediction performance is quite sufficient.

Table 2: Model's parameters.

$\gamma_1 = -0.3417$	$\gamma_2 = 1.3821$	$\gamma_3 = 9.9039$
$\gamma_4 = 4.9160$	$\gamma_5 = 4.2360$	$\gamma_6 = 1.1154$
$\gamma_7 = 0.6018$	$\gamma_8 = -5.0325$	$\beta = -4$
$\alpha_1 = -0.0235$	$\alpha_2 = 0.4512$	$\alpha_3 = 0.1224$
$\alpha_4 = -0.1944$	$\alpha_5 = 0.4498$	

Table 3: Proposed model's performance.

CC	RMSE	ROC	OR
0.81	0.47	0.80	0.10

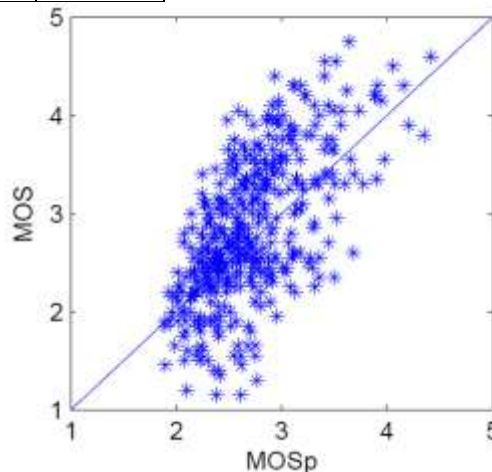


Figure 6: MOS vs MOS_p results.

VI. Conclusions

A no-reference video quality assessment model is proposed based on different spatial and temporal features for multimedia applications. The features are calculated irrespective of any predefined artifacts of the video coders. Pixel distortion and edge information measure of each frame are considered as spatial features. Effective temporal information, effective frame and frame similarity measure of the sequences are considered as temporal features. The subjective database which contains large varieties of scene content and motion indicates that the performance of the proposed model is sufficient. In the future, the research can be extended to incorporate color component and human visual characteristics for features calculation to get more sufficient performance.

References

- [1] ITU-T Rec. J. 144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," May, 2004.
- [2] ANSI standard T1.74-2001, Objective video quality measurement using a Peak Signal-to-Noise Ratio (PSNR) Full Reference Technique.
- [3] S. Winkler, "Digital video quality: Vision models and metrics," John Wiley and Sons, 2005.
- [4] J. Okamoto, T. Hayashi, A. Takahashi, and T. Kurita, "Verification of objective quality assessment method for arbitrary video streams," in Proc IMQA 2005, Sept. 2005.
- [5] R. Karthikeyan, G. Sainarayanan and S. N. Deepa, "Perceptual video quality assessment in H.264 video coding standard using objective modeling," SpringerPlus 2014, doi:10.1186/2193-1801-3-174.
- [6] G. Valenzise, S. Magni, M. Tagliasacchi, S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," IEEE transactions on circuits and systems for video technology, vol. 22, no. 4, pp. 605-618, 2012.
- [7] S. Hemami, and A. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," Signal Processing: Image Communication, vol. 25, no. 7, pp. 469-481, 2010.
- [8] ITU-R Recommendation BT.500-10, Methodology for the subjective assessment of the quality of television pictures.
- [9] ITU-T P.910, Subjective video quality assessment methods for multimedia applications, Geneva, Sept. 1999.
- [10] Z. Wang, "Rate scalable foveated image and video communications," PhD thesis, Dept. of ECE, The University of Texas at Austin, Dec. 2003.
- [11] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on Structural distortion measurement," IEEE Trans. vol. 19, no. 2, pp. 121-132, Feb. 2004.
- [12] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in Proc. IEEE ICNN, Perth, Australia, pp. 1942-1948, Nov. 1995.
- [13] Youkuu Horita, Makoto Miyahara, and Tadakuni Murai, "Estimation improvement in picture quality scale of monochrome still picture," IEICE Trans. vol. j80, B-I, pp. 505-514, June 1997.
- [14] VQEG Multimedia (MM) Report Final 2.6, Sept. 2008.
- [15] VQEG, "Final Report from the video quality experts group on the validation of objective models of video quality assessment, FR-TV Phase II (August 2003)," <http://www.vqeg.org/>.

Acknowledgement

Author would like to thank MICT lab, Japan for providing their database.