A Frame Rate Dependent Video Quality Metric based on Temporal Wavelet Decomposition and Spatiotemporal Pooling

Fan Zhang, Alex Mackin and David R. Bull

Abstract

This paper presents an objective quality metric (FRQM), which characterises the relationship between variations in frame rate and perceptual video quality. The proposed method estimates the relative quality of a low frame rate video with respect to its higher frame rate counterpart, through temporal wavelet decomposition, subband combination and spatiotemporal pooling. FRQM was tested alongside six commonly used quality metrics (two of which explicitly relate frame rate variation to perceptual quality), on the publicly available BVI-HFR video database, that spans a diverse range of scenes and frame rates, up to 120fps. Results show that FRQM offers significant improvement over all other tested quality assessment methods with relatively low complexity.

Index Terms

Frame rate, perceptual quality, video quality assessment, frame rate dependent quality metric, FRQM

I. Introduction

Visual experiences are key drivers, not just for the entertainment sector but also for business, security and communication technologies. Following the limited success of 3D content in consumer applications, there is a desire among consumers, content producers and broadcasters, for new and more immersive video content. Efforts in this respect have focused on extending the video parameter space with greater dynamic range, wider colour gamut, higher spatial resolution and increased frame rates [1, 2]. In this paper, we focus specifically on the influence of frame rate on video quality.

Increased frame rates reduce the visibility of temporal artefacts, such as motion blur and aliasing, (flicker, judder, etc.) leading to a more realistic portrayal of the scene [3]. The frame rates required to eliminate temporal artefacts can be calculated using the characteristics of the human spatiotemporal contrast sensitivity function [4–6], and are related to the speed of the stimulus. Various subjective experiments, conducted on a variety of source sequences and testing methodologies, have also attempted to quantify the relationship between frame rate and perceptual quality [7–11]. However the test material used in these experiments is either generally not publicly available, or the range of frame rates tested are restricted. Mackin *et al.* [12] recently published an HD video database, which spans a range of frame rates, from 15 to 120 frames per second (fps). This publicly available database provides a valuable tool for modelling the relationship between frame rate and perceived quality in a content dependent way.

Existing generic full-reference image and video quality metrics can be modified to provide objective video quality assessment at various frame rates, by temporally upsampling a lower frame rate version of a sequence to the same temporal resolution as the reference. A number of bespoke quality metrics [9, 10, 13–17] have also been developed to detect the temporal artefacts associated with variations in frame rate. However they have not been properly evaluated on high frame rate (60fps+) content.

In this paper, a new Frame Rate dependent Quality Metric (FRQM) is presented. FRQM predicts, with relatively low complexity, the difference in perceptual quality between different frame rates through temporal wavelet decomposition, subband combination and spatiotemporal pooling. FRQM provides excellent correlation performance with subjective scores compared to four popular generic quality metrics and two state-of-the-art frame rate quality

The authors acknowledge funding and support from UK EPSRC platform grant EP/M000885/1, and BBC Research and Development. The authors are with the Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: {fan.zhang, a.mackin, dave.bull}@bristol.ac.uk).

models. As a result, FRQM promises to be an efficient and useful tool for the purpose of adaptive frame rate selection, video quality assessment and high frame rate video compression.

The remainder of the paper is organised as follows: Section II outlines the proposed quality metric in detail. Section III benchmarks the performance and complexity of FRQM against the other tested quality metrics. Conclusions are provided in Section IV.

II. Proposed algorithm

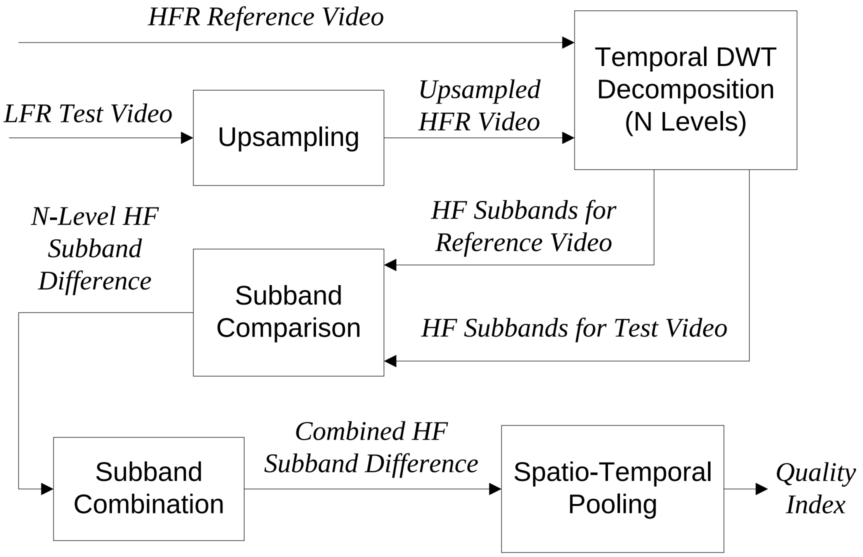

HFR Reference Video
LFR Test Video
Upsampling
Upsampled HFR Video
Temporal DWT Decomposition (N Levels)
HF Subbands for Reference Video
N-Level HF Subband Difference
Subband Comparison
HF Subbands for Test Video
Subband Combination
Combined HF Subband Difference
Spatio-Temporal Pooling
Quality Index

Fig. 1: Diagrammatic illustration of the proposed quality metric.

The architecture of the proposed quality metric is shown in Fig. 1. This metric requires both a reference video and a lower frame rate test version. In order to obtain a quality index, four steps are undertaken: i) temporal upsampling, (ii) temporal DWT decomposition, (iii) subband comparison and combination, and (iv) spatiotemporal pooling. These are described in detail as below.

A. Temporal upsampling

The first step of the model is to temporally upsample the lower frame rate video to the same frame rate as the reference. This is to ensure the reference and the test sequences have the same number of frames for further processing. Although more advanced upsampling methods are proposed in the literature [18, 19], here we use nearest-neighbor interpolation [20] as to emulate a hold-type display (e.g. LCD) [21].

B. Temporal DWT decomposition

In order to detect temporal distortions due to frame rate reduction, both the reference and upsampled videos are temporally decomposed using a 1-D Discrete Wavelet Transform (DWT) at various levels. This processes all luma pixels at the same spatial coordinates over all frames simultaneously using the Haar wavelet. The number of decomposition levels N is related to the ratio between the reference frame rate (FPS_H) and the initial frame rate of the test sequence (FPS_L):

$$N = \left\lceil \log_2 \frac{\mathrm{FPS}_H}{\mathrm{FPS}_L} \right\rceil . \tag{1}$$

For a pixel with spatial coordinates (x, y) within the frame t, its corresponding high frequency (HF) subband coefficient values are denoted as $B_r(x, y, t, n)$ and $B_t(x, y, t, n)$ for the reference and upsampled test videos respectively. Here n is the level of HF subband.

C. Subband comparison and combination

Following wavelet decomposition, subband coefficient values of the reference and test sequences are compared at N levels in order to obtain a HF subband difference $D(x,y,t,n)$:

$$D(x,y,t,n) = |B_r(x,y,t,n) - B_t(x,y,t,n)|\,. \tag{2}$$

A weighted sum of these difference values is computed over all N subband levels:

$$D_c(x,y,t) = \sum_{n=1}^{N} W(f(n)) \cdot D(x,y,t,n), \tag{3}$$

where $W(f(n))$ are weighting values related to the temporal frequency $f(n)$ of the HF subband at level n. Based on the Nyquist criterion [22], the wavelet subband frequency can be calculated from the frame rate of the reference sequence:

$$f(n) = \frac{\mathrm{FPS}_H}{2^n}. \tag{4}$$

The determination of these weighting values are described in Section III-A.

D. Spatiotemporal pooling

It is noted that, in most cases, the distribution of video artefacts are spatially and temporally non-uniform. This becomes more evident when artefacts are introduced by frame rate reduction rather than compression. FRQM therefore employs an effective strategy to obtain global quality indices (frame and sequence levels). This is inspired by the pooling approach in SSIM [23],

In the pooling stage, each video frame is segmented into K non-overlapping square blocks, $\{\mathrm{MB}_1, \mathrm{MB}_2, \cdots, \mathrm{MB}_K\}$. The average combined subband difference is calculated for each block as follows:

$$\bar{D}_c(k,t) = \sum_{(x,y)\in \mathrm{MB}_k} \frac{D_c(x,y,t)}{S^2}, k = 1, 2, \cdots, K, \tag{5}$$

where S is the block size (S=16 is used here[1]). The frame level quality index is then the maximum value across all blocks:

$$Q(t) = \max\left\{\bar{D}_c(1,t), \bar{D}_c(2,t), \cdots, \bar{D}_c(K,t)\right\}. \tag{6}$$

Video quality is calculated by first splitting the video sequence into non-overlapping segments with the same duration, and then, similar to spatial pooling, the maximum of the local mean values is calculated:

$$\mathcal{Q} = \max\left\{\sum_{t=1}^{L} Q(t), \sum_{t=L+1}^{2L} Q(t), \cdots, \sum_{t=GL-L+1}^{T} Q(t)\right\}, \tag{7}$$

where G is the number of segments, T denotes the total number of frames in the reference video, and L represents the number of frames per segment ($L = \mathrm{FPS}_H/5$). The duration of each segment is set as 200ms here, which is based on the visual persistence duration of the human visual system [24].

$\mathcal{Q}$ is then converted to decibel units to give the FRQM quality index for the sequence:

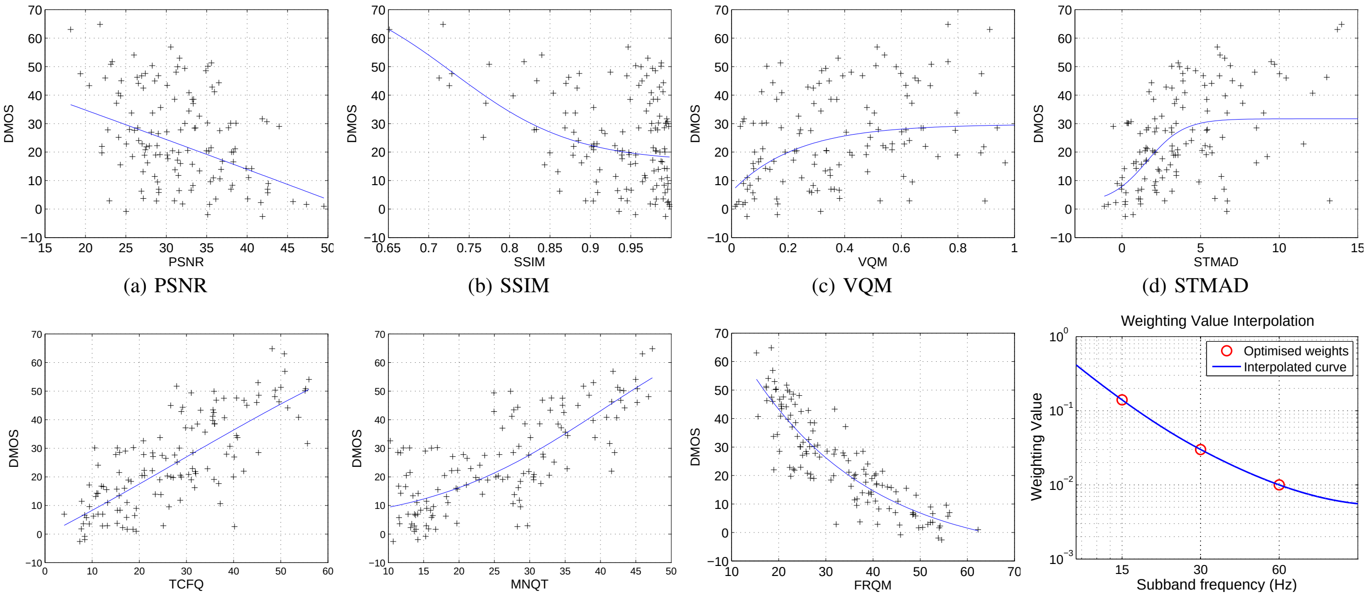
$$\mathrm{FRQM} = 20 \cdot \log_{10}\left(\frac{255}{\mathcal{Q}}\right). \tag{8}$$

III. Results and discussions

Objective quality models are conventionally evaluated by correlating their predictions with subjective opinion scores. The BVI-HFR video database contains 22 unique and diverse sequences (HD) that span a range of frame rates (15, 30, 60 and 120 fps). Mean opinion scores (MOS) for all 88 sequences (22×4) were collected using the single stimulus continuous quality evaluation (SSCQE) methodology. This MOS data is transformed into DMOS (difference) scores by:

$$\mathrm{DMOS}_{\mathrm{FPS}_H/\mathrm{FPS}_L} = \mathrm{MOS}_{\mathrm{FPS}_H} - \mathrm{MOS}_{\mathrm{FPS}_L}. \tag{9}$$

[1] A larger local window than SSIM (8×8) is used here, to account for the increased spatial resolution in the BVI-HFR video database.



(a) PSNR (b) SSIM (c) VQM (d) STMAD

(e) TCFQ (f) MNQT (g) FRQM (h) Interpolation

Fig. 2: (a-g) Scatter plots of subjective DMOS versus the predictions of the tested quality metrics on the BVI-HFR database. The blue curves represent the logistic fitting functions. (h) The interpolation of the weighting values, using cubic spline interpolation [22].

The video database is thus extended to contain 132 (22×6) DMOS scores by selecting different frame rates as a reference, for example we can compare 120fps to 60, 30 and 15 fps; 60fps to 30 and 15fps; 30fps to 15fps.

Alongside FRQM, six existing objective quality metrics[2] are also tested for comparison: PSNR, SSIM [23], VQM [26], STMAD [27], TCFQ [17] and MNQT[3] [10]. The first four are popular generic image and video quality metrics, none of which have been designed specifically to address the impact of frame rate, while TCFQ and MNQT are considered as the state-of-the-art in modelling the relationship between frame rates and perceptual quality. The generic quality metrics were applied between the higher frame rate reference and the corresponding lower frame rate upsampled sequence, using the same upsampling method as Section II-A.

TCFQ and MNQT were originally designed to predict normalised MOS ($\mathrm{MOS}_{\mathrm{FPS}_L}/\mathrm{MOS}_{\mathrm{FPS}_H}$) rather than DMOS [10, 17]. In order to compare with the other tested metrics, the quality values for each test sequence obtained by TCFQ and MNQT have been transformed using the MOS of the corresponding higher frame rate reference by:

$$\mathrm{q}' = \mathrm{MOS}_{\mathrm{FPS}_H} - \mathrm{q}\cdot\mathrm{MOS}_{\mathrm{FPS}_H}. \tag{10}$$

Here q and q' represent the original and transformed quality values respectively for either TCFQ or MNQT. It is noted that the parameters of TCFQ and MNQT were trained on relatively low spatial and temporal resolution videos [10, 17]. To achieve a fair comparison, these parameters were trained again on the whole BVI-HFR video database, using the same procedures as described in [10, 17]. Results of the TCFQ and MNQT approaches using updated parameters are provided in this section, both of which exhibit improved statistical performance compared to the original specified parameters.

A logistic fitting function was used to fit objective quality indices and DMOS scores [28]. Four correlation statistics, Linear Correlation Coefficient (LCC), Spearman Rank Order Correlation Coefficient (SROCC), Outlier Ratio (OR) and Root Mean Squared Error (RMSE), were calculated for each quality metric [1]. Statistical difference in performance between tested metrics was verified using an F-test on the residual between the actual and predicted DMOS scores [29, 30].

[2] Another popular generic video quality metric MOVIE [25] has not been tested here, due to the large memory required for high frame rate content.

[3] TCFQ and MNQT are based on the same basic model, but combine different video features. We have opted not to use the early version of the TCFQ metric [9], due to its inferior performance on the BVI-HFR database.

The complexity figures of the quality metrics are estimated based on their average execution time (normalised to PSNR), obtained on an Intel Core i7-3770S CPU@3.10GHz PC platform using Matlab R2012a.

A. Parameter determination based on cross validation

In order to obtain optimised weighting values, a twofold cross validation approach [31] was conducted on the BVI-HFR database. 132 pairs of test and reference videos were randomly divided into two subsets with equal size. One of these containing 66 test-reference pairs from 11 source sequences, was firstly used for training the weighting parameters. The remaining subset was employed for benchmarking metric performance. Each subset was used once as the test data. Due to the limited number of video frame rates included in the BVI-HFR database, this training process can only determine the weighting values for the temporal frequencies 15Hz, 30Hz and 60Hz.

The optimum weighting values were determined as those which maximise the correlation between FRQM and the subjective scores of the training data set, with an exhaustive search range from 0 to 1. SROCC was used to assess correlation performance, as it does not rely on data fitting. This training-testing split (each split contributes two sets of optimum weighting values) was repeated 100 times in order to minimise content bias. The other six metrics are tested on the testing subsets using the parameters (if applicable) provided in their published softwares.

Based on the cross validation results, the final optimum weighting values are 0.01, 0.03 and 0.14 for temporal frequencies 60Hz, 30Hz and 15Hz respectively, obtained by taking the medians of the optimum values over these 100 splits (200 trials). These parameters are used in the rest of the paper for testing the overall performance on the BVI-HFR database. Weights for temporal frequencies other than those tested here, can be interpolated using the same method as Fig. 2(h).

TABLE I: Cross validation results on the BVI-HFR database.

Metric	PSNR	SSIM	VQM	STMAD	TCFQ	MNQT	FRQM
μ	0.370	0.301	0.394	0.521	0.753	0.709	**0.896**
σ	0.082	0.073	0.074	0.100	0.044	0.045	**0.028**

The average (μ) and standard deviation (σ) of SROCC values for FRQM and the other tested quality metrics over 100 twofold cross validations in shown in Table I, in which the results are based on the SROCC values from testing subsets. FRQM has the best average (0.896) and standard deviation (0.028) of all the metrics evaluated.

B. Results for the whole BVI-HFR video database

The relationship between the six tested quality metrics and the subjective scores of the BVI-HFR video database can be observed in Fig. 2(a-g), alongside the correlation performance in Table II and F-test results in Table III. FRQM reports a significant improvement in statistical performance over all other tested metrics, demonstrated by the higher correlation coefficients, fewer outliers, lower prediction errors and positive F-test results.

TABLE II: The performance of the tested quality metrics on the BVI-HFR database. The best performer is highlighted in **bold**.

Metric	LCC	SROCC	OR	RMSE	Complexity
PSNR	0.399	0.347	0.644	15.071	1
SSIM	0.414	0.280	0.652	15.033	26
VQM	0.387	0.374	0.644	15.165	108
STMAD	0.520	0.507	0.629	13.998	1256
TCFQ	0.763	0.748	0.492	10.505	1876
MNQT	0.756	0.702	0.561	10.625	1945
FRQM	**0.878**	**0.890**	**0.303**	**7.732**	30

Among all the quality assessment methods evaluated, STMAD, TCFQ and MNQT exhibit much higher computational complexity (over 1000 times greater than PSNR) than the other tested methods. The execution time of FRQM is only 29 greater than PSNR, which is relatively low among the tested quality metrics.

TABLE III: F-test results for the tested quality metrics on the BVI-HFR database. Here a "1" indicates that the metric in the row is superior to the related column (and vice versa for "-1"), while "0" indicates no significant difference.

Metric	PSNR	SSIM	VQM	STMAD	TCFQ	MNQT	FRQM
PSNR	-	0	0	0	-1	-1	-1
SSIM	0	-	0	0	-1	-1	-1
VQM	0	0	-	0	-1	-1	-1
STMAD	0	0	0	-	-1	-1	-1
TCFQ	1	1	1	1	-	0	-1
MNQT	1	1	1	1	0	-	-1
FRQM	1	1	1	1	1	1	-

IV. Conclusions

In this paper, we have presented an efficient quality assessment method FRQM, which employs a temporal wavelet transform, subband combination and spatiotemporal pooling to predict the difference in perceptual quality due to frame rate reduction. Tested on the BVI-HFR database, the proposed method exhibits relatively low computational complexity while providing superior performance to four generic quality metrics and two state-of-the-art frame rate quality models. The proposed quality metric will facilitate visually lossless temporal resolution adaptation for video acquisition and storage, and can also be integrated with video coding algorithms offering promising bitrate reduction for high frame rate content.

V. References

References

[1] D. R. Bull, *Communicating pictures: a course in image and Video Coding*, Academic Press, 2014.

[2] Recommendation ITU-R BT.2020, "Parameter values for ultra-high definition television systems for production and international programme exchange," Tech. Rep., ITU-R, 2012.

[3] M. Armstrong, D. Flynn, M. Hammond, S. Jolly, and R. Salmon, "High frame-rate television," *BBC Research White Paper 169*, 2008.

[4] S. J. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Photonics West'98 Electronic Imaging*. International Society for Optics and Photonics, 1998, pp. 180–191.

[5] K. Noland, "The application of sampling theory to television frame rate requirements," *BBC Research & Development White Paper 282*, 2014.

[6] A. Mackin, K. Noland, and D. R. Bull, "The visibility of motion artifacts and their effect on motion quality," in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2016.

[7] M. Sugawara, K. Omura, M. Emoto, and Y. Nojiri, "Temporal sampling parameters and motion portrayal of television," in *SID*, 2009, vol. 9, pp. 1200–1203.

[8] M. Emoto, Y. Kusakabe, and M. Sugawara, "High-frame-rate motion picture quality and its independence of viewing distance," *Journal of Display Technology*, vol. 10, no. 8, pp. 635–641, 2014.

[9] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286–298, 2011.

[10] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-star: a perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Trans. on Image Processing*, vol. 23, no. 6, pp. 2473–2486, 2014.

[11] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang, "Perceptual quality assessment of high frame rate video," in *International Workshop on Multimedia Signal Processing*. IEEE, 2015, pp. 1–6.

[12] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2015.

[13] K. Yang, C. C. Guest, K. El-Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. on Multimedia*, vol. 9, no. 7, pp. 1528–1535, 2007.

[14] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 641–651, 2008.

[15] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Trans. on Broadcasting*, vol. 53, no. 1, pp. 441–446, 2007.

[16] S. H. Jin, C. S. Kim, D. J. Seo, and Y. M. Ro, “Quality measurement modeling on scalable video applications,” in *IEEE Workshop onMultimedia Signal Processing*. IEEE, 2007, pp. 131–134.

[17] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, “Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 671–682, 2012.

[18] Y. Guo, L. Chen, Z. Gao, and X. Zhang, “Frame rate up-conversion method for video processing applications,” *IEEE Trans. on Broadcasting*, vol. 60, no. 4, pp. 659–669, 2014.

[19] E. Shechtman, Y. Caspi, and M. Irani, “Space-time super-resolution,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 531–545, 2005.

[20] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, 2 edition, 2002.

[21] H. Pan, X.-F. Feng, and S. Daly, “LCD motion blur modeling and analysis,” in *Proc. IEEE Int Conf. on Image Processing 2005*. IEEE, 2005, vol. 2, pp. II–21.

[22] T. J. Cavicchi, *Digital Signal Processing*, John Wiley & Sons, Inc., 2000.

[23] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, 2004.

[24] M. Coltheart, “Iconic memory and visible persistence,” *Perception & psychophysics*, vol. 27, no. 3, pp. 183–228, 1980.

[25] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quailty assessment of natural videos,” *IEEE Trans. on Image Processing*, vol. 19, pp. 335–350, 2010.

[26] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Trans. on Broadcasting*, vol. 50, pp. 312–322, 2004.

[27] P. V. Vu, C. T. Vu, and D. M. Chandler, “A spatiotemporal most-apparent-distortion model for video quality assessment,” in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2011, pp. 2505–2508.

[28] Video Quality Experts Group, “Final report from the video quality experts group on the validation of objective quailty metrics for video quality assessment.,” Tech. Rep., VQEG, 2000.

[29] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Trans. on Image Processing*, vol. 19, pp. 335–350, 2010.

[30] F. Zhang and D. Bull, “A perception-based hybrid model for video quality assessment,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2016.

[31] D. C. Howell, *Statistical methods for psychology*, Cengage Wadsworth, USA, 7 edition, 2010.