ViMSSIM: From Image to Video Quality Assessment

Cuong Vu
Oklahoma State University
202 Engineering South
Stillwater, OK 74078
cuong.vu@okstate.edu

Sachin Deshpande
Sharp Laboratories of America, Inc.
5750 NW Pacific Rim Blvd.
Camas, WA 98607
sdeshpande@sharplabs.com

ABSTRACT

This paper presents a low complexity algorithm for *video* quality assessment, called ViMSSIM, which extends the *image* quality metric MS-SSIM to take into account visual perception of spatial and temporal quality of the video. First, we use a modified exponential moving average model to pool the MS-SSIM indices of all frames and create the spatial quality index. Next, we apply MS-SSIM to the frame-different images of the distorted video and the reference video. Here each pair i of frame differences is formed by taking the difference between frame *i+1* of the original video and its previous frame i, and the difference between frame *i+1* of the *distorted* video and frame i of the *original* video. Averaging the MS-SSIM indices of these frame-different images leads to the temporal quality index. The final quality index is the average of the computed spatial and temporal index. Testing on the LIVE video database consisting of 150 videos demonstrates that ViMSSIM performs well in predicting video quality. In addition, the proposed algorithm has significantly lower complexity compared to current state-of-the-art benchmark video quality metrics.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metric

General Terms

Algorithm

Keywords

Video quality assessment, image quality assessment, metric, distortion

1. INTRODUCTION

Internet video and mobile video traffic is growing rapidly, according to recent forecasts [1, 2]. In any video transmission system, it is necessary to control the quality of transmitted videos to meet the expectations of end users. While asking human subjects to rate the quality of a given video is most reliable, this method is time consuming and expensive, and thus is impractical in most applications. The ability to quantify the quality of a video which agrees with human perception, therefore, plays an important role. The goal of an objective video quality assessment (VQA) is to build an algorithm which determines the quality of a given video in a manner that highly correlates with human ratings.

Full reference VQA assumes the availability of the original (reference) video, which is considered as perfect quality, while measuring the quality of the processed (and most of the time is distorted) video. Examples of dedicated VQA algorithms are the MOVIE algorithm [3], proposed by Seshadrinathan and Bovik, and the VQM algorithm [4] from Pinson and Wolf. A thorough review of modern objective VQA methods is provided in [5].

Previous work has demonstrated that several image quality assessment (IQA) algorithms, e.g., MS-SSIM [6], VSNR [7], MAD [8], perform quite well in image quality task. However, Chikkerur *et al.* in [5] show that when applying these algorithms to video quality task in a frame-by-frame fashion and averaging the indices of all frames to form the final video quality index, their performances are not good. This result might be due to the fact that frame-by-frame comparison only accounts for spatial information of the video, while the temporal information of the video is ignored. In addition, one may also argue that averaging the indices of all frames is not an appropriate technique for pooling frame quality indices.

In this paper, we present an objective VQA algorithm, called ViMSSIM, which utilizes the MS-SSIM algorithm for IQA. First, instead of taking the average of MS-SSIM indices of all frames, we use a modified exponential moving average model to pool these indices and create the spatial quality index. Next, we apply MS-SSIM to the frame-different images of the distorted video and the reference video to lead to the temporal quality index. The final quality index is the average of the computed spatial and temporal index. We demonstrate the performance of this approach on videos from the LIVE video database [9]. In the next section, we provide details of the algorithm. Results and discussion of the algorithm is presented in Section 3.

2. ALGORITHM

Figure 1 shows the diagram of the proposed algorithm. Two measures implemented in ViMSSIM: (1) Spatial Quality Measure, which computes the spatial quality of the distorted video via applying MS-SSIM to every frame, and then

Movid '12, February 24, 2012, Chapel Hill, North Carolina, USA.

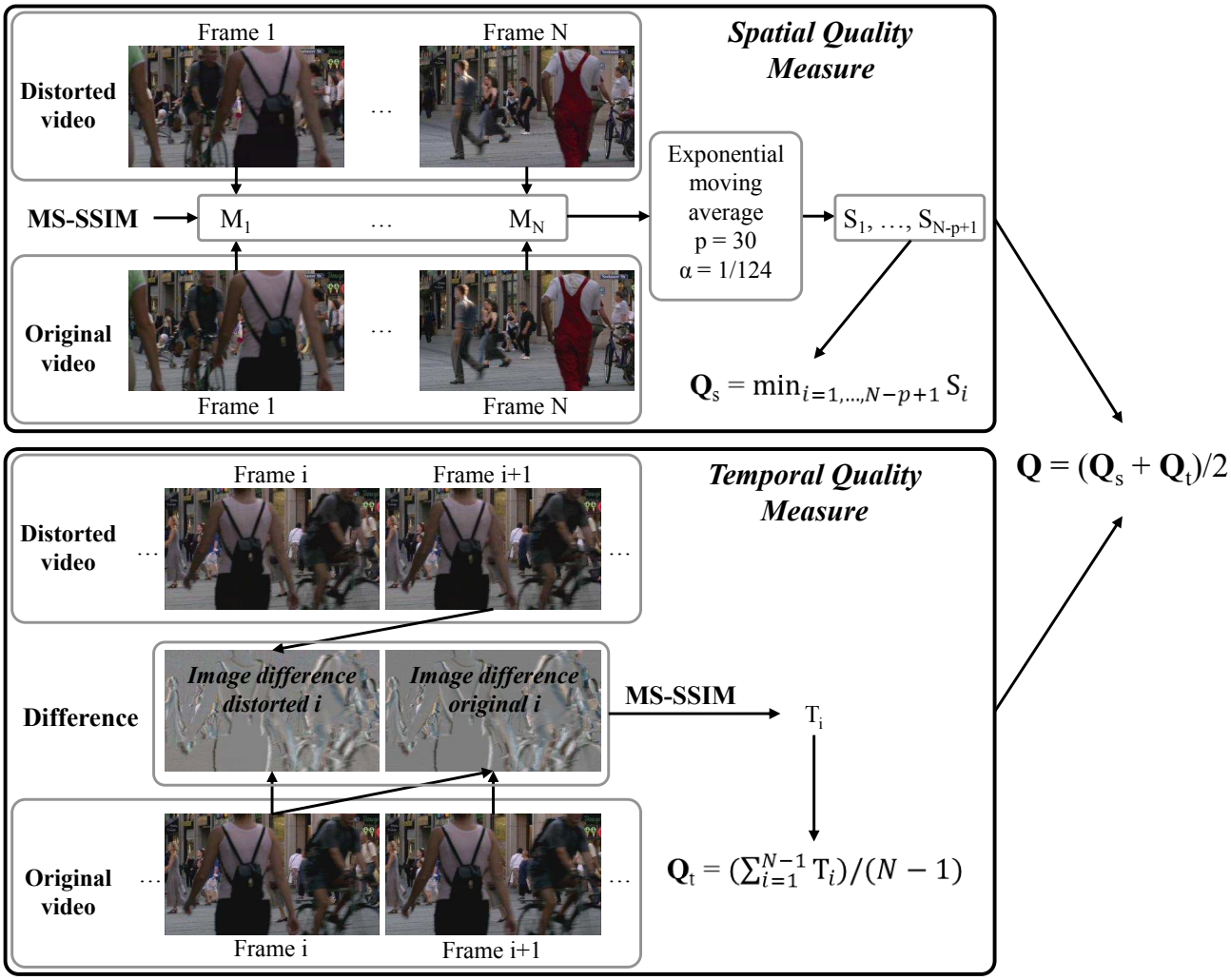

Figure 1: Diagram of the proposed ViMSSIM algorithm.

pools the obtained values using a modified exponential moving average model to create the spatial quality index; and (2) Temporal Quality Measure, which applies MS-SSIM to every pair of frame differences. Here each pair i of frame differences is formed by taking the difference between frame *i+1* of the original video and its previous frame i, and the difference between frame *i+1* of the distorted video and frame i of the *original video* (see, Figure 1). The output MS-SSIM values of all pairs are then averaged to create the temporal quality index. The final ViMSSIM index is computed as the average of the spatial and temporal quality index.

2.1 MS-SSIM

MS-SSIM (Multi-scale Structural Similarity) is the extension of the IQA index SSIM (Structural Similarity Index) [10] to account for the variations of image resolution and viewing distance. For two image patches $\mathbf{x}$ and $\mathbf{y}$ in comparison, the structural similarity of the two patches contains three main components: luminance comparison, contrast comparison, and structure comparison. These three components are defined, respectively, as:

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{1}$$

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2}$$

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3}$$

in which μ_x, μ_y, σ_x, σ_y, and σ_{xy} denote the mean, standard deviation and covariance of the luminance of $\mathbf{x}$ and $\mathbf{y}$, respectively. The constants C_1, C_2, and C_3 are the stabilizing terms of the corresponding components. The SSIM index is defined as:

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = l(\mathbf{x},\mathbf{y})^\alpha c(\mathbf{x},\mathbf{y})^\beta s(\mathbf{x},\mathbf{y})^\gamma \tag{4}$$

For images with a dynamic range L (for 8-bit images $L = 255$), the constants C_1, C_2, C_3 are selected in the form $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, and $C_3 = C_2/2$, where K_1 and K_2 are small. In addition, setting the parameters $\alpha = \beta = \gamma =$

1 makes SSIM become:

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

In MS-SSIM, the image pyramids of M layers (scales) are first built for the distorted and reference image. At layer m, the contrast and structure components were computed according to Equations (2) and (3), and denoted as $c_m(\mathbf{x},\mathbf{y})$ and $s_m(\mathbf{x},\mathbf{y})$, respectively. The luminance component is computed at the highest layer (layer M) only and denoted as $l_M(\mathbf{x},\mathbf{y})$. Combining these components gives the overall MS-SSIM evaluation:

$$\text{MS-SSIM}(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha_M} \prod_{m=1}^{M} [c_m(\mathbf{x},\mathbf{y})]^{\beta_m} [s_m(\mathbf{x},\mathbf{y})]^{\gamma_m} \quad (6)$$

The common selection of the parameters in MS-SSIM are: $K_1 = 0.01$, $K_2 = 0.03$, $M = 5$, $\alpha_M = 0.1333$, and $\beta_{m=1,\ldots,5} = \gamma_{m=1,\ldots,5} = [0.0448, 0.2856, 0.3001, 0.2363, 0.1333]$. An image with higher MS-SSIM index is assumed to have better visual quality. This MS-SSIM has been shown to perform better than SSIM in image quality assessment task.

2.2 ViMSSIM: Spatial Quality Measure

Let $\mathbf{V}_d$ and $\mathbf{V}_r$ denote the distorted video and the reference (original) video, respectively, each contains N frames. Let $\mathbf{f}_{d,i}$ and $\mathbf{f}_{r,i}$ denote the frame i of the distorted and the reference video, respectively, and M_i denote the MS-SSIM index of the pair $(\mathbf{f}_{d,i}, \mathbf{f}_{r,i})$.

Barkowsky *et al.* in [11], via several subjective tests, have made an assumption that the instantaneous impression of video quality is the result of several frames. Using this assumption, we argue that spatial quality of a video is not perceived by human in a frame-by-frame basis, but in a segment-of-frames basis instead. In order to model segments of spatial information which can affect the subject's decision, we apply a modified exponential moving average procedure to the M_i values. Specifically, we create a set of values S_n which are computed as:

$$S_1 = \left(\sum_{i=1}^{p} M_i\right)/p \quad (7)$$

and for $n = 1, 2, \ldots, N-p$

$$S_{n+1} = \alpha M_{n+p} + (1-\alpha) S_n \quad (8)$$

in which p decides how the first value is initialized, and α is the smoothing factor which is usually selected in the form of $\alpha = \eta/(p+1)$. In the computation of S_{n+1} in comparison with S_n, the index M_{n+p} is introduced while the contributions of all previous indices $M_1, \ldots, M_{n+p-1}$ are decreased. The two main properties that we aim for all S_n to have are: (1) Each S_n contains information of at least half a second of the video in order to make an effect on the subject's perception of quality; and (2) In each S_n the new frame does not make immediately strong effect and the contribution of previous frames does not drop too fast. As the frame rate of a video can be as fast as 60fps, we choose $p = 30$ in order to satisfy property (1). The second property requires α to be small; therefore we empirically select $\eta = 0.25$ and thus $\alpha = 1/124$.

Another important assumption made in [11] is that when doing video quality rating task, subjects focus on the maximum distortion visible to them. Therefore, based on this assumption, in our metric we compute the spatial quality of the video based on the worse segment. Thus the spatial quality index, denoted as Q_s, is computed as the minimum value over all S_n, $n = 1, \ldots, N-p+1$:

$$Q_s = \min_n S_n \quad (9)$$

2.3 ViMSSIM: Temporal Quality Measure

Many techniques have been proposed to capture temporal (motion) information in a video. The simplest approach of modeling temporal change is taking the difference between two consecutive frames. Here we modify this technique to quantify how temporal information changes in the distorted video as compared to the reference video. Specifically, let $\mathbf{D}_{d,i}$ and $\mathbf{D}_{r,i}$ denote the images which represent the temporal information at frame (time) i of the distorted video and the reference video, respectively. $\mathbf{D}_{d,i}$ and $\mathbf{D}_{r,i}$ are computed as:

$$\mathbf{D}_{r,i} = \mathbf{f}_{r,i+1} - \mathbf{f}_{r,i} \quad (10)$$

$$\mathbf{D}_{d,i} = \mathbf{f}_{d,i+1} - \mathbf{f}_{r,i} \quad (11)$$

Notice in Equation (11) that $\mathbf{D}_{d,i}$ is computed as the difference between the frame $i+1$ of the *distorted* video and the frame i of the *reference* video. Figure 3 demonstrates the advantage of using this method. In this figure, image (c) is computed as the difference between frame $i+1$ (Figure 2(d)) and i (Figure 2(c)) of the distorted video. On the other hand, image (b) is computed as the difference between frame $i+1$ of the *distorted* video (Figure 2(d)) and frame i of the *original* video (Figure 2(a)). This demonstration shows that the image in Figure 3 (b) (image $\mathbf{D}_{r,i}$ as computed in Equation (11)) can capture more distortion of temporal information than the image in Figure 3 (c) (the difference between two consecutive frames of the distorted video). A zoom-in portion of the image in Figure 3(b) and Figure 3(c) is shown in Figure 3(e) and Figure 3(d), respectively, for better visual examination.

We next compute the MS-SSIM index of the pair ($\mathbf{D}_{d,i}$, $\mathbf{D}_{r,i}$), denoted as T_i. This quantity T_i represents the temporal quality of the distorted video at the frame (time) i (where $i = 1, \ldots, N-1$). The temporal quality index, denoted as Q_t, is computed as the average of these T_i values:

$$Q_t = \left(\sum_{i=1}^{N-1} T_i\right)/(N-1) \quad (12)$$

The final quality index Q, or our ViMSSIM index, of the video is computed as the average of the spatial and the temporal quality index:

$$Q = (Q_s + Q_t)/2 \quad (13)$$

3. RESULTS AND ANALYSIS

We test the performance of our ViMSSIM algorithm using the LIVE video database [9]. This database contains 10 high quality videos as reference (original videos), and 150 distorted videos (15 distorted video for each original video). All videos have a resolution of 768×432. Four distortion types presented in the database are MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP network, and through error-prone wireless network. In the following analysis, we name these 4 distortion types as MPEG-2, H.264,

Figure 2: (a) and (b): two consecutive frames of an original video; (c) and (d): two corresponding consecutive frames of a distorted video.

IP, and Wireless, respectively. The overall ratings of videos from LIVE are reported as differential mean opinion scores (DMOS).

Six well-known quality assessment methods were used in comparison: PSNR, VSNR [7], MS-SSIM [6], MAD [8], VQM [4], and MOVIE [3]. Note that PSNR, VSNR, MS-SSIM, and MAD are image quality assessment methods, and here are extended to video by applying them on a frame-by-frame basis and averaging the scores across all frames. We evaluate the performances of these algorithms based on two criteria: (1) The Pearson Linear Correlation Coefficient (CC), which measures how well an algorithm's predictions correlate with the subjective scores; and (2) The Spearman Rank Order Correlation Coefficient (SROCC), which measures the relative monotonicity between the predictions and subjective scores.

Before evaluating the performance of an algorithm, it is common to apply a logistic transform to the predicted ratings to bring the predictions on the same scale as the MOS or DMOS values, and to account for the nonlinear relationship between the predictions and opinion scores. A logistic fitting function is commonly used to nonlinearly map between the predictions and subjective scores. We adopt the logistic function suggested by the Video Quality Experts Group [12], which is given by

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp(\frac{x-\tau_3}{\tau_4})} + \tau_2, \tag{14}$$

where τ_1, τ_2, τ_3, and τ_4 are the model parameters chosen to minimize the MSE between the predicted values and the subjective scores.

Table 1 shows the performances of the proposed algorithm and six other methods on the LIVE video database. We also include in this table the individual performance of the spatial quality index (see, Section 2.2), which we call S-ViMSSIM, and the temporal quality index (see, Section 2.3), which we call T-ViMSSIM.

In this table, the performance of S-ViMSSIM shows a noticeable improvements in comparison with MS-SSIM. This result supports our argument that averaging the indices of all frames might not be an appropriate technique for pooling frame quality indices. Our modified exponential moving average model implemented in S-ViMSSIM, on the other hand, captures the information in each segment of frames separately, and thus might better simulate how human perceives spatial visual quality of a video.

Considering each distortion category separately, ViMSSIM holds good performance on Wireless and H.264 (among the two best performers regarding both CC and SROCC). It is also the best performer on MPEG-2 regarding CC. Even though ViMSSIM does not give good result on IP category, overall our proposed algorithm performs better than all the others in this comparison. while the second best performer is T-ViMSSIM. Among the rest algorithms, overall T-ViMSSIM is the second best while MOVIE shows quite competitive results. To the best of our knowledge before

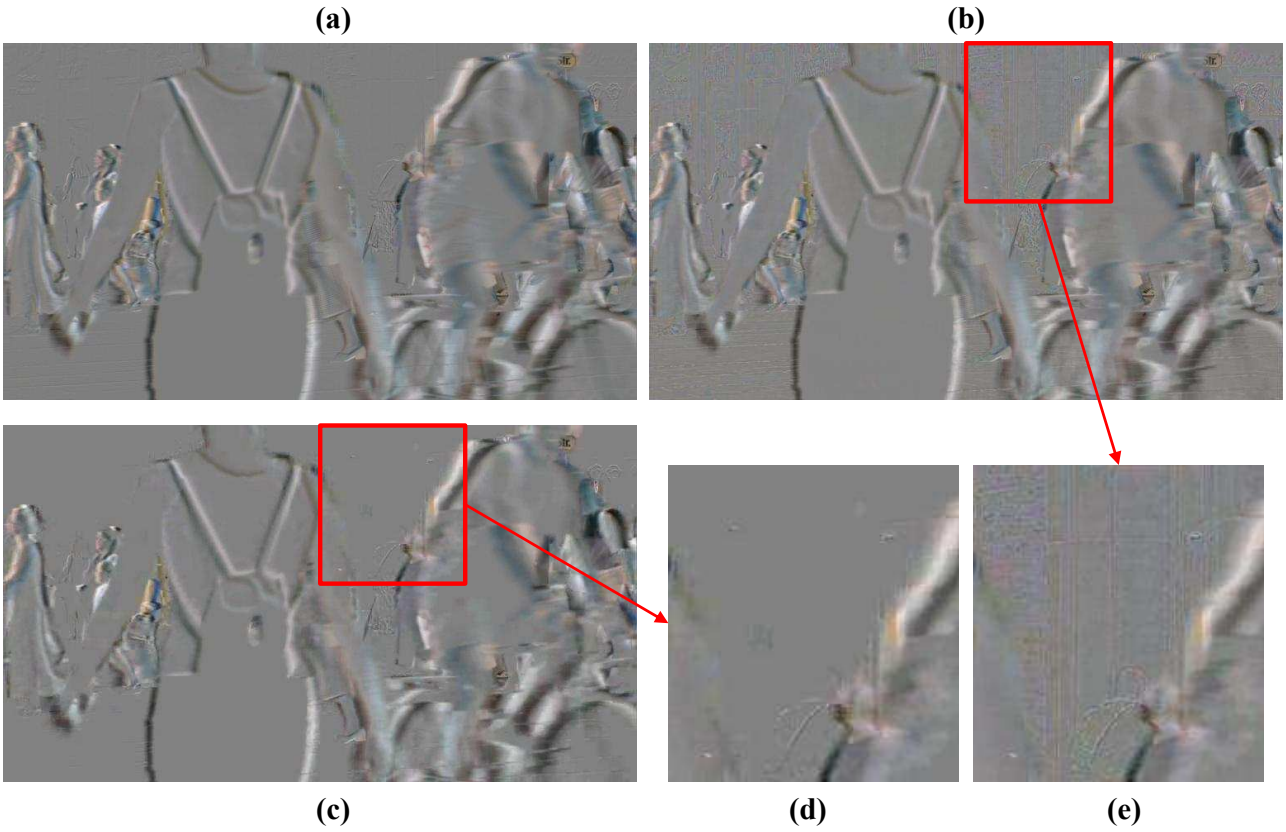

Figure 3: (a) The difference between the images in Figure 2(b) and Figure 2(a) (image $D_{r,i}$ as computed in Equation (10)); (b) The difference between the images in Figure 2(d) and Figure 2(a) (image $D_{d,i}$ as computed in Equation (11)); (c) The difference between the images in Figure 2(d) and Figure 2(c) (two consecutive frames of the distorted video). Image (b) better captures the distortion of temporal information than image (c). A zoom-in portion of each image (b) and (c) is shown in (e) and (d), respectively.

developing the ViMSSIM algorithm, MOVIE was the best performer on the LIVE video database.

However, as MOVIE uses a 3D Gabor analysis, it is potentially slow. Running on our Quadcore 2.9GHz and 4G RAM computer, MOVIE takes approximately 5 hours to compute the quality index of a 10-second video (25fps, 768×432). In contrast, our proposed ViMSSIM algorithm takes approximately 120 seconds over the same condition. Even though a thorough analysis is needed in order to make a solid conclusion for speed comparison, this result is reasonable due to the low complexity of ViMSSIM. This advantage has significant benefits in mobile video scenario in which bandwidth, power, and processor are somewhat limited.

4. CONCLUSION

In this paper, we have presented a video quality assessment algorithm, ViMSSIM, which extends the image quality metric MS-SSIM to take into account visual perception of spatial and temporal quality of the video. We propose in our algorithm a modified exponential moving average model to pool frame-based MS-SSIM indices, and use the outputs of this model to create the spatial quality index. We also use a simple but efficient technique to measure temporal distortion via applying MS-SSIM to frame-different images. The final quality index is the average of the computed spatial and temporal index. Testing on the LIVE video database has demonstrated the great potential of ViMSSIM, in term of both performance and speed, in predicting video quality.

5. REFERENCES

[1] Cisco Inc., "Cisco visual networking index: Forecast and methodology, 2010-2015," June 2011.

[2] Cisco Inc., "Cisco visual networking index: Global mobile data traffic forecast update, 2010-2015," February 2011.

[3] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.

[4] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[5] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165 –182, june 2011.

Table 1: Performance of ViMSSIM and other quality assessment algorithms on the LIVE video database. The two best performances are bolded.

	Wireless	IP	H.264	MPEG-2	All data
CC					
PSNR	0.6690	0.4645	0.5492	0.3891	0.5621
VSNR	0.6992	0.7341	0.6216	0.5980	0.6896
MS-SSIM	0.7170	0.7219	0.6919	0.6604	0.7441
MAD	0.7649	0.6730	0.6302	0.6898	0.7206
VQM	0.7325	0.6480	0.6459	**0.7860**	0.7236
MOVIE	**0.8386**	**0.7622**	0.7902	0.7595	0.8116
S-ViMSSIM	0.7515	**0.7503**	0.7837	0.7834	0.7796
T-ViMSSIM	0.8219	0.6890	**0.8810**	0.7709	**0.8122**
ViMSSIM	**0.8327**	0.7322	**0.8117**	**0.7978**	**0.8260**
SROCC					
PSNR	0.6574	0.4167	0.4585	0.3862	0.5398
VSNR	0.7019	0.6894	0.6460	0.5915	0.6755
MS-SSIM	0.7285	0.6534	0.7051	0.6617	0.7361
MAD	0.7616	**0.7125**	0.6580	0.6509	0.7341
VQM	0.7214	0.6383	0.6520	**0.7810**	0.7026
MOVIE	**0.8109**	**0.7157**	0.7664	**0.7733**	0.789
S-ViMSSIM	0.7340	0.6521	0.7713	0.7694	0.7690
T-ViMSSIM	0.7951	0.6650	**0.8580**	0.7499	**0.7984**
ViMSSIM	**0.8111**	0.6774	**0.8559**	0.7630	**0.8211**

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, November 2003.

[7] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," in *IEEE Transactions on Image Processing, 16*, 2007.

[8] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, 2010.

[9] A. C. Bovik K. Seshadrinathan, R. Soundararajan and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[10] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[11] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 266 –279, april 2009.

[12] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, Phase II," August 2003, http://www.vqeg.org.