Efficient Motion Weighted Spatio-Temporal Video SSIM Index

Anush K. Moorthy and Alan C. Bovik

Laboratory for Image and Video Engineering (LIVE), Department of Electrical & Computer Engineering, The University of Texas at Austin, USA.

ABSTRACT

Recently, Seshadrinathan and Bovik proposed the Motion-based Video Integrity Evaluation (MOVIE) index for VQA.[1,2] MOVIE utilized a multi-scale spatio-temporal Gabor filter bank to decompose the videos and to compute motion vectors. Apart from its psychovisual inspiration, MOVIE is an interesting option for VQA owing to its performance. However, the use of MOVIE in a practical setting may prove to be difficult owing to the presence of the multi-scale optical flow computation. In order to bridge the gap between the conceptual elegance of MOVIE and a practical VQA algorithm, we propose a new VQA algorithm - the spatio-temporal video SSIM based on the essence of MOVIE. Spatio-temporal video SSIM utilizes motion information computed from a block-based motion-estimation algorithm and quality measures using a localized set of oriented spatio-temporal filters. In this paper we explain the algorithm and demonstrate its conceptual similarity to MOVIE; we explore its computational complexity and evaluate its performance on the popular VQEG dataset. We show that the proposed algorithm allows for efficient FR VQA without compromising on the performance while retaining the conceptual elegance of MOVIE.

1. INTRODUCTION

Algorithmic assessment of video quality involves the design of algorithms that can be used to predict the perceived quality of videos. Traditionally, mean squared error (MSE) has been utilized for this purpose. However, as many authors have pointed out, MSE does not correlate well with the human perception of quality.[3,4] As it is the human who is the ultimate viewer of the video, his perception of quality is of utmost importance. Given the peculiarities of the human visual system (HVS), it comes as no surprise that MSE does not perform well in terms of correlation with human perception. It is hence that the area of VQA has seen tremendous activity in the recent past.[5–10]

VQA algorithms have traditionally been categorized as those that predict the quality of a test video given a pristine reference - full reference (FR) VQA; those that predict the test video quality without the presence of the reference - no-reference (NR) VQA and those where the test video stream is embedded with some information from the pristine reference - reduced reference (RR) VQA. MSE, for example is a FR VQA algorithm (albeit a poor one). In this paper, out aim is to propose a new FR VQA algorithm.

Recently, Seshadrinathan and Bovik proposed the Motion-based Video Integrity Evaluation (MOVIE) index for VQA.[7,11] MOVIE utilized a multi-scale spatio-temporal Gabor filter bank to decompose the videos and to compute motion vectors. Apart from its inspiration from the HVS, MOVIE is an interesting option for VQA owing to its performance. However, a practical implementation of MOVIE for VQA is hindered by the computational complexity of the algorithm. In order to bridge the gap between the conceptual elegance of MOVIE and a practical VQA algorithm, we propose a new VQA algorithm - spatio-temporal video SSIM (stVSSIM) - based on the essence of MOVIE.

stVSSIM utilizes the simple single-scale structural similarity index (SS-SSIM)[12] for spatial quality assessment. SS-SSIM correlates well with human perception of quality for image quality assessment (IQA).[13,14] Temporal quality assessment in stVSSIM is achieved by an extension of SS-SSIM to the spatio-temporal domain and this

Further author information: (Send correspondence to A. K. Moorthy)
A. K. Moorthy: E-mail: anushmoorthy@gmail.com, Telephone: 1 512 415 0213
A. C. Bovik: E-mail: bovik@ece.utexas.edu, Telephone: 1 512 471 5370

extension is labeled as SSIM-3D. Motion information is incorporated in stVSSIM using a block-based motion estimation algorithm, as opposed to optical flow as in MOVIE. As of now, there exist very few VQA algorithms that incorporate motion information,[7, 10] even though the HVS has been hypothesized to compute and utilize this information.[15] Computing spatial and temporal quality scores separately is again inspired from the HVS.[15]

In this paper, we first briefly describe MOVIE and SS-SSIM. We then describe SSIM-3D, a method to evaluate spatio-temporal quality of a video. Further, we describe how motion information extracted from block motion estimation is incorporated in the algorithm to form stVSSIM. A brief note on the complexity of stVSSIM follows. Finally, we evaluate the performance of the proposed algorithm on a popular publicly available VQA dataset and compare its performance to leading VQA algorithms. We demonstrate that stVSSIM performs well in terms of correlation with human perception and conclude this paper.

2. A BRIEF FORAY INTO MOVIE-LAND

Motion based video integrity evaluation (MOVIE) evaluates the quality of videos sequences not only in space and time, but also in space-time, by evaluating motion quality along motion trajectories.[7]

First, both the reference and the distorted video sequences are spatio-temporally filtered using a family of bandpass Gabor filters. MOVIE uses three scales of Gabor filters. A Gaussian filter is included at the center of the Gabor structure to capture low frequencies in the signal. A local quality computation of the band-pass filtered outputs of the reference and test videos is then undertaken by considering a set of coefficients within a window from each of the Gabor sub-bands. The quality index so obtained is termed as the spatial MOVIE index even though it captures some temporal distortions.

MOVIE uses the same filter bank to compute motion information i.e., estimate optical flow from the reference video. The algorithm used is a multi-scale extension of the Fleet and Jepson[16] algorithm that uses the phase of the complex Gabor outputs for motion estimation. If the motion of the distorted video matches that of the reference video exactly, then the filters that lie along the motion plane orientation defined by the flow from the reference will be activated by the distorted video and outputs of filters that lie far away from this plane will be negligible. In presence of a temporal artifact, however, the motion in the reference and distorted videos do not match and a different set of filter banks may be activated. Thus, motion vectors from the reference are used to construct velocity-tuned responses. This can be accomplished by a weighted sum of the Gabor responses, where positive excitatory weights are assigned to those filters that lie close to the spectral plane and negative inhibitory weights are assigned to those that lie farther away from the spectral plane. This excitatory-inhibitory weighting results in a strong response when the distorted video has motion equal to the reference and a weak response when there is a deviation from the reference motion.

Finally, the mean square error is computed between the response vectors from the reference video (tuned to its own motion) and those from the distorted video. This temporal MOVIE index captures essentially temporal distortions in the video under test.

Application of MOVIE to videos produces a map of spatial and temporal scores at each pixel location for each frame of the video sequence. In order to pool the scores to create a single quality index for the video sequence, MOVIE uses the coefficient of variation.[17]The coefficient of variation serves to capture the distribution of the distortions accurately. The coefficient of variation is computed for the spatial and temporal MOVIE scores for each frame, then the values are averaged across frames to create the spatial and temporal MOVIE indices for the video sequence (temporal MOVIE index uses the square root of the average). The final MOVIE score is a product of the temporal and spatial MOVIE scores. A detailed description of the algorithm can be found in.[7]

3. SPATIAL QUALITY ASSESSMENT: STRUCTURAL SIMILARITY

The single-scale structural similarity index (SS-SSIM) was initially proposed for image quality assessment (IQA).[12] An extension of the metric was later proposed for video quality assessment (VQA).[18] Even though the authors in[18] utilized some motion information, the approach was ad-hoc and this was reflected in the performance. In this paper, SS-SSIM as described below is utilized for spatial quality assessment. Spatio-temporal quality assessment is undertaken using a modified version of SS-SSIM, which we shall consider in the next section.

For two image patches drawn from the same location in the reference - $\mathbf{x} = \{x_i | i = 1, 2, \ldots, N\}$ - and test - $\mathbf{y} = \{y_i | i = 1, 2, \ldots, N\}$ - images; let μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} be the means of $\mathbf{x}$, $\mathbf{y}$, the variances of $\mathbf{x}$, $\mathbf{y}$ and the covariance between $\mathbf{x}$ and $\mathbf{y}$ respectively.

We compute:

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

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

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

where $C_1 = (K_1L)^2, C_2 = (K_2L)^2, C_3 = C_2/2$ are small constants; L is the dynamic range of the pixel values and $K_1 = 0.01$ and $K_2 = 0.03$ are scalar constants. The constants C_1, C_2 and C_3 prevent instabilities from arising when the denominator tends to zero.

SS-SSIM is then computed as:

$$SS - SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{4}$$

At each coordinate, the SSIM index is calculated within a local window. We use a 11×11 circular-symmetric Gaussian weighting function[12] $w = \{w_i | i = 1, 2, \ldots, N\}$, with standard deviation of 1.5 samples, normalized to sum to unity ($\sum_{i=1}^{N} w_i = 1$).

For spatial quality assessment, SS-SSIM is computed on a frame-by-frame basis. The spatial-quality measure is applied on each frame and the frame-quality measure is computed using the percentile-approach.[19] This approach was proposed by the authors in[19] to accommodate for the bias that humans exhibit when asked to rate images. The claim was that humans tend to rate images with low quality regions with greater severity and hence using a percentile approach would enhance algorithm performance. Percentile-SSIM or P-SSIM was shown to perform well in terms of correlation with human perception.[14, 19] Here, P-SSIM is applied on the scores obtained for each frame. Specifically, the frame-quality measure is:

$$S_{frame} = \frac{1}{|\psi|} \sum_{i \in \psi} SSIM(i)$$

where $| \cdot |$ denotes the cardinality of the set and ψ denotes the set of the lowest 6% of SSIM values from the frame and $SSIM(i)$ denotes the SS-SSIM score at pixel location i. A similar technique was used for VQA in.[5]

The spatial score for the video is computed as the mean of the frame-level scores and is denoted as S_{video}. Even though we use the mean here for simplicity, there may be other better pooling strategies.[5] We defer analyzing pooling strategies for the future.

4. TEMPORAL QUALITY ASSESSMENT

Temporal quality evaluation consists of modifying SS-SSIM for a spatio-temporal chunk of the video and then utilizing motion information derived from motion vectors in order to perform a ‘weighting’ of the obtained scores. In this section, we describe the SSIM-3D, which is the extension of the SS-SSIM to the spatio-temporal domain and the weighting scheme used, which results in stVSSIM.

4.1 3D Structural Similarity

SS-SSIM as described in the previous section utilizes a two-dimensional or spatial window to compute the statistics between image patches. In SSIM-3D, a similar quality computation is undertaken utilizing statistics evaluated over volumes. In this case, pixels of the video are indexed by their spatio-temporal location.

Specifically, let x and y denote two 3-dimensional elements or volume chunks extracted from the reference and distorted video around a pixel location (i, j, k); where (i, j) correspond to the spatial co-ordinates and k corresponds to the frame number. At this pixel (i, j, k); a volume chunk is defined as a spatio-temporal region around the pixel which has the dimensions (α, β) spatially and encompasses γ frames temporally. One way to extract such a region is to consider only temporally-earlier frames - which would be the case when a real-time implementation is desired. In this paper, such a chunk is extracted using both temporally earlier and temporally later frames. Hence for pixel (i, j, k), the video chunk consists of the region between the pixels:

$$\begin{gathered}(i - \lfloor\alpha/2\rfloor, j - \lfloor\beta/2\rfloor, k - \lfloor\gamma/2\rfloor),\\ (i - \lfloor\alpha/2\rfloor, j + \lfloor\beta/2\rfloor, k - \lfloor\gamma/2\rfloor),\\ (i + \lfloor\alpha/2\rfloor, j - \lfloor\beta/2\rfloor, k - \lfloor\gamma/2\rfloor),\\ (i + \lfloor\alpha/2\rfloor, j + \lfloor\beta/2\rfloor, k - \lfloor\gamma/2\rfloor)\\ \text{and}\\ (i - \lfloor\alpha/2\rfloor, j - \lfloor\beta/2\rfloor, k + \lfloor\gamma/2\rfloor),\\ (i - \lfloor\alpha/2\rfloor, j + \lfloor\beta/2\rfloor, k + \lfloor\gamma/2\rfloor),\\ (i + \lfloor\alpha/2\rfloor, j - \lfloor\beta/2\rfloor, k + \lfloor\gamma/2\rfloor),\\ (i + \lfloor\alpha/2\rfloor, j + \lfloor\beta/2\rfloor, k + \lfloor\gamma/2\rfloor).\end{gathered}$$

where, $\lfloor\cdot\rfloor$ represents the floor function.

In our implementation, $\alpha = \beta = 11$ and $\gamma = 33$. The spatial dimensions have been chosen to correspond with those from spatial SS-SSIM, while the temporal dimension corresponds to the number of frames spanned by filters from the coarsest scale in MOVIE. Currently our implementation is single-scale (both spatially and temporally); however, a multi-scale extension would involve video chunks having the same support temporally as those from MOVIE.

We then compute the following statistics over this video chunk:

$$\mu_{x(i,j,k)} = \sum_{m=1}^{\alpha}\sum_{n=1}^{\beta}\sum_{o=1}^{\gamma} w(m,n,o)x(m,n,o)$$

$$\mu_{y(i,j,k)} = \sum_{m=1}^{\alpha}\sum_{n=1}^{\beta}\sum_{o=1}^{\gamma} w(m,n,o)y(m,n,o)$$

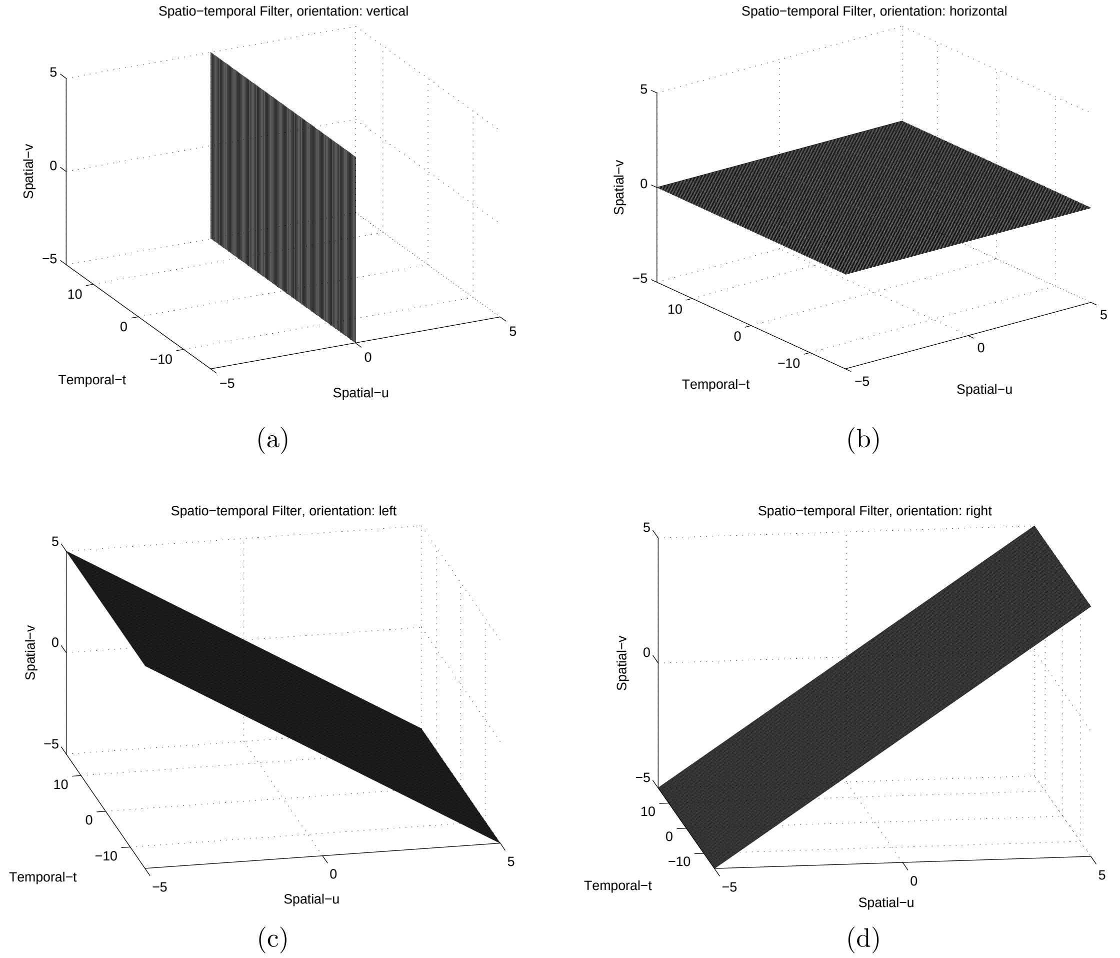
$$\sigma^2_{x(i,j,k)} = \sum_{m=1}^{\alpha}\sum_{n=1}^{\beta}\sum_{o=1}^{\gamma} w(m,n,o)(x(m,n,o) - \mu_{x(i,j,k)})^2$$

$$\sigma^2_{y(i,j,k)} = \sum_{m=1}^{\alpha}\sum_{n=1}^{\beta}\sum_{o=1}^{\gamma} w(m,n,o)(y(m,n,o) - \mu_{y(i,j,k)})^2$$

$$\sigma_{x(i,j,k)y(i,j,k)} = \sum_{m=1}^{\alpha}\sum_{n=1}^{\beta}\sum_{o=1}^{\gamma} w(m,n,o)(x(m,n,o) - \mu_{x(i,j,k)})(y(m,n,o) - \mu_{y(i,j,k)})$$

and then compute SSIM-3D at location (i, j, k) as:

$$SSIM - 3D(i,j,k) = \frac{(2\mu_{x(i,j,k)}\mu_{y(i,j,k)} + C_1)(2\sigma_{x(i,j,k)y(i,j,k)} + C_2)}{(\mu^2_{x(i,j,k)} + \mu^2_{y(i,j,k)} + C_1)(\sigma^2_{x(i,j,k)} + \sigma^2_{y(i,j,k)} + C_2)} \tag{5}$$



(a) (b)

(c) (d)

Figure 1. Filters used for spatio-temporal quality assessment. (a) Vertical ,(b) Horizontal, (c) Left, and (d) Right. The spatial axes are marked as u and v, and the temporal axis is marked as t. For each axis the origin corresponds to the pixel at which the spatio-temporal window is currently centered. The negative direction of the temporal axis corresponds to temporally earlier frames and the positive direction corresponds to temporally later frames (w.r.t. the frame under consideration). The negative and positive directions on the u axis correspond to the left and right of the current pixel respectively. The negative and positive directions on the v-axis correspond to the pixels below and above the current pixel respectively.

where the constants C_1 and C_2 are the same as those for spatial SS-SSIM.

w is a weighting function that depends upon the 'type' of the filter that is being used. As we have mentioned before, the essence of stVSSIM is evaluating spatio-temporal quality along various orientations at a pixel, followed by a weighting scheme which assigns a spatio-temporal quality index to that pixel. In order to evaluate the spatio-temporal quality along various orientations, we utilize the following 'types' of spatio-temporal filters: *vertical, horizontal, left and right.* The spatio-temporal filters utilized for SSIM-3D are seen in Figure 1. The weight w in the above equations is dependent upon the type of filter being used. Hence, at each pixel location (i, j, k), we compute SSIM-3D for each of the filters mentioned above. These four scores are collapsed into a single spatio-temporal score using motion information.

At this point it is imperative to point out that the temporal quality measure (which is computed from the spatio-temporal SSIM-3D) accounts for some spatial quality estimate as well. This is again by design similar to MOVIE. However, the spatial measure is only spatial. In order to emulate MOVIE exactly, one could use SSIM-3D for the spatial-measure (without the temporal weighting described below). However, we have found that the performance improvement is minuscule when this technique is used, and does not justify the increase in complexity. Hence, in order to evaluate spatial quality, we utilize the simple SS-SSIM as described before.

4.2 Incorporating Motion Information

MOVIE utilizes a set of Gabor filter banks to evaluate optical flow at each pixel location. In order to approximate this process and to develop a scheme that is computationally efficient, instead of using optical flow, we utilize block-motion-estimation. Block motion estimation has traditionally been used for video compression, where motion vectors are computed between neighboring frames. The motion vectors so obtained are then utilized to perform motion compensation & frame differencing and the encoded video stream consists of the motion vectors & the quantized frame-differences.[20] For stVSSIM, motion-estimation is performed using the Adaptive Rood Pattern Search (ARPS) algorithm,[21] using 8 × 8 blocks. The block size chosen is motivated by the fact that 8 × 8 blocks have traditionally been used for video compression.[20] However, as variable block-sizes have been incorporated in latest coding standards, it is of interest to evaluate the effect of change in block-size on performance.

Once motion vectors for each pixel (i, j, k) is available, spatio-temporal SSIM-3D scores are to be weighted. In MOVIE, such weighting is performed by evaluating the distance of the plane formed by the motion vectors in the frequency domain to the centers of each of the filters.[7] Here, instead of weighting the spatio-temporal values using floating point numbers, we perform a *greedy* weighting. Specifically, the spatio-temporal score at pixel (i, j, k) is the score produced by that 'type' of filter which is closest to the direction of motion at pixel (i, j, k). For example, if the motion vector at a pixel were $(u, v) = (2, 0)$, the spatio-temporal score of that pixel would equal to the SSIM-3D value produced by the *horizontal* filter. In cases where the motion vector is equidistant from two of the filter planes, the spatio-temporal score is the mean of the SSIM-3D scores from the two filters. When the motion vector is zero, the spatio-temporal score is the mean of all four SSIM-3D values.

We wish to draw the attention of the reader to some finer points of such a weighting scheme. By avoiding actual multiplication of (possibly floating point) weights, we enable a faster implementation. Further, in cases where a mean is to be computed, the mean is computed between either two or four values, in which case we divide each score by a factor of 2 or 4. It is obvious that such an implementation is easily accomplished by a shift in the binary representation of the score. The reader is bound to question the complexity of utilizing a motion-estimation algorithm. In the next sub-section, we shall describe an approach that does not require the computation of these motion-vectors, further reducing complexity.

Thus, evaluation of the spatio-temporal quality score at each pixel location involves the use of SSIM-3D using four different filter directions, and a greedy weighting based on motion information obtained from block motion estimation. At the end of this process, we have a temporal quality score at each pixel location in the frame. Again, the strategy used to pool these scores is the one proposed in[14] and used for collapsing the spatial scores. Specifically, for each frame, the temporal score is:

$$T_{frame} = \frac{1}{|\psi|} \sum_{i,j \in \psi} SSIM - 3D(i, j, k)$$

where $| \cdot |$ denotes the cardinality of the set and ψ denotes the set of the lowest 6% of SSIM values from the frame and $SSIM - 3D(i, j, k)$ denotes the SSIM-3D score at pixel location (i, j, k).

The temporal score for the video is computed as the mean of the frame-level scores and is denoted as T_{video}. The final score for the video is then given by $T_{video} \times S_{video}$. The temporal and spatial scores are computed for each 16^{th} frame in the video similar to MOVIE.

4.3 Complexity

stVSSIM utilizes SS-SSIM for spatial quality assessment. It can easily be shown that the computational complexity of SS-SSIM is $O(MN)$, where M, N are the dimensions of the frame of the video whose quality is being evaluated. For SSIM-3D, the number of multiplies required for each computation of $\mu_{x(i,j,k)} \& \mu_{y(i,j,k)}$ are $\alpha \cdot \beta \cdot \gamma$ and that for $\sigma^2_{x(i,j,k)}, \sigma^2_{x(i,j,k)} \& \sigma_{x(i,j,k)y(i,j,k)}$ are $2 \cdot \alpha \cdot \beta \cdot \gamma$ each. Since no floating point multiplies are performed for the temporal weighting, there are no more multiples added here. We neglect the complexity of computing the final SSIM-3D term as well as computing which filter output forms the SSIM-3D score as these are negligible

Table 1. Performance of stVSSIM on the VQEG dataset. SROCC = Spearman rank ordered correlation coefficient. LCC = linear correlation coefficient. OR = outlier ratio.

Algorithm	SROCC	LCC	OR
SS-SSIM (no weighting)	0.788	0.820	0.597
SS-SSIM (with weighting)	0.812	0.849	0.578
MOVIE	0.833	0.821	0.644
stVSSIM	0.840	0.843	0.616

compared to the above mentioned calculations. Hence, the net complexity of SSIM-3D is the same as that of SS-SSIM: $O(MN)$. Further, we sort the scores to pool them and sorting can be achieved with a worst case complexity of $O(MNlog(MN))$.

Finally, the motion estimation process used here is ARPS[21] and is (relatively) computationally intensive. In order to reduce the complexity associated with computing block-motion estimates, we describe an alternative strategy to extract motion estimates. In a practical setting, the user of a VQA algorithm rarely has access to the pristine YUV video. Generally, a compressed video is processed by a black box and the video at the output needs to be evaluated for its fidelity. In this situation, the user of the VQA algorithm has a compressed reference with respect to which he wishes to evaluate video quality. At this point of time, one solution is to compute block motion estimates and then to apply stVSSIM. However, since most modern video compression algorithms utilize the motion-compensated frame-differencing model for compression, the motion estimates computed by the compression algorithm are present in the compressed reference video stream. All that is left for the VQA algorithm to do, is to extract these motion estimates and utilize them instead of the motion estimates from a block based motion estimation algorithm. This procedure eliminates a major bottleneck in stVSSIM and enables its practical deployment.

Further reduction in complexity of stVSSIM may be achieved by incorporating techniques detailed in,[22] which have shown to reduce complexity without compromising on the performance of SSIM.

5. PERFORMANCE EVALUATION

In order to evaluate the proposed algorithm, we utilize the popular Video Quality Experts Group (VQEG) FRTV Phase-I dataset.[23] The VQEG dataset consists of two sets of videos (525 and 625), each of which has 160 distorted videos. The distorted videos were created from 10 reference videos for each set. Each video has associated with it a differential mean opinion score (DMOS) that was computed using subjective scores obtained from a large-scale user study. The DMOS is representative of the perceived quality of the video.

We utilize different performance measures to evaluate the algorithm: Spearman's rank-ordered correlation coefficient (SROCC), linear (Pearson's) correlation coefficient (LCC) and outlier ratio (OR). LCC and OR are computed after transforming the algorithm scores using a logistic function. The logistic function used is the same as the one proposed by the VQEG:[23]

$$logistic(x) = \frac{\beta_1 - \beta_2}{1 + exp(\frac{x-\beta_3}{|\beta_4|})} + \beta_2$$

where, x is the quality score produced by the VQA algorithm and $\beta_1 \dots \beta_4$ are parameters which are estiamted through a non-linear curve fitting procedure as described in.[23] Table 1 lists the performance of stVSSIM on the entire VQEG dataset (320 distorted videos). The table also lists the performance of video-SSIM - without weighting and with weighting as described in the original implementation of video-SSIM.[18] Further, we also list the performance of MOVIE.[7] As can be seen, stVSSIM performs competitively with the state of the art algorithms.

One criticism that has been associated with the VQEG dataset is the presence of non-natural sequences.[7] The presence of non-natural sequences complicates the evaluation of algorithm performance. Even though SSIM is not explicitly based on human visual system models it has been shown[24] that there exists a relationship between the structure term of SSIM and another IQA index - Visual Information Fidelity (VIF).[25] VIF is based

Table 2. Performance of stVSSIM and MOVIE on natural sequences only. SROCC = Spearman rank ordered correlation coefficient. LCC = linear correlation coefficient.

Algorithm	SROCC	LCC
PSNR	0.739	0.718
SSIM (no weighting)	0.802	0.810
MOVIE	0.860	0.858
stVSSIM	0.865	0.877

on natural scene statistics (NSS), and utilizes the gaussian scale mixture (GSM) model which is widely used to model NSS.[26] Hence, in order to allow for a fair evaluation of performance and list the performance of stVSSIM for natural sequences only in table 2. Table 2 also lists the performance of MOVIE for natural-only sequences. stVSSIM performs exceedingly well in terms of correlation with human perception.

6. CONCLUSION

We proposed a computationally efficient video quality assessment (VQA) algorithm, spatio-temporal video SSIM (stVSSIM) index inspired by MOVIE. stVSSIM was shown to retain the conceptual elegance of MOVIE as well as its superior performance, while allowing for a practical implementation. stVSSIM utilized the simple single-scale structural similarity index (SS-SSIM) for spatial quality assessment. Temporal quality was evaluated using the extension of SS-SSIM to the spatio-temporal domain - SSIM-3D. Motion information from a block motion estimation process allowed for mimicking the motion-based weighting scheme of MOVIE. Further, we described how block motion estimation could be avoided completely, thereby reducing computational complexity. The algorithm was evaluated on the popular video quality experts' group (VQEG) FRTV Phase-I dataset and was shown to perform extremely well in terms of correlation with human perception.

REFERENCES

[1] Seshadrinathan, K. and Bovik, A. C., "Motion-based perceptual quality assessment of video," *Human Vision and Electronic Imaging XIV. Proceedings of the SPIE* **7240** (2009).

[2] Seshadrinathan, K. and Bovik, A. C., "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing* **(to appear)** (2009).

[3] Girod, B., "What's wrong with mean-squared error?, Digital images and human vision, A. B. Watson, Ed.," 207–220 (1993).

[4] Wang, Z. and Bovik, A. C., "Mean squared error: Love it or leave it? - a new look at fidelity measures." IEEE Signal Processing Magazine (January 2009).

[5] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* , 312–313 (Sept. 2004).

[6] Masry, M., Hemami, S., and Sermadevi, Y., "A scalable wavelet-based video distortion metric and applications," *IEEE Transactions on circuits and systems for video technology* **16**(2), 260–273 (2006).

[7] Seshadrinathan, K., *Video quality assessment based on motion models*, PhD thesis, The University of Texas at Austin (2008).

[8] Liu, T., Wang, Y., Boyce, J. M., Yang, H., and Wu, Z., "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment* **3** (April 2009).

[9] Barkowsky, M., J. Bialkowski and, B. E., Bitto, R., and Kaup, A., "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment* **3**, 266–279 (April 2009).

[10] Ninassi, A., Meur, O. L., Callet, P. L., and Barba, D., "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment* **3**, 253–265 (April 2009).

[11] Seshadrinathan, K. and Bovik, A. C., "A structural similarity metric for video based on motion models," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* , 869–872 (Apr. 2007).

[12] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Processing Letters* **13**, 600–612 (Apr. 2004).

[13] Sheikh, H. R., Sabir, M. F., and Bovik, A. C., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (Nov. 2006).

[14] Moorthy, A. K. and Bovik, A. C., "Visual importance pooling for image quality assessment," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment* **3**, 193–201 (April 2009).

[15] Sekuler, R. and Blake, R., [*Perception*], Random House USA Inc (1988).

[16] Fleet, D. and Jepson, A., "Computation of component image velocity from local phase information," *International Journal of Computer Vision* **5**(1), 77–104 (1990).

[17] Frank, H. and Althoen, S. C., [*Statistics: Concepts and Applications*], ch. The coefficient of variation, 5859, Cambridge, Great Britan: Cambridge University Press (1995).

[18] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assesssment based on structural distortion measurement," *Signal Processing: Image communication* , 121–132 (Feb. 2004).

[19] Moorthy, A. K. and Bovik, A. C., "Perceptually significant spatial pooling techniques for image quality assessment," *Human Vision and Electronic Imaging XIV. Proceedings of the SPIE* **7240** (January 2009).

[20] Richardson, I., "H. 264 and MPEG-4 video compression," (2003).

[21] Nie, Y. and Ma, K., "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transactions on Image Processing* **11**(12), 1442–1449 (2002).

[22] Rouse, D. and Hemami, S., "Understanding and simplifying the structural similarity metric," *15th IEEE International Conference on Image Processing, 2008. ICIP 2008* , 1188–1191 (2008).

[23] "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment."

[24] Seshadrinathan, K. and Bovik, A. C., "Unifying analysis of full reference image quality assessment," *15th IEEE International Conference on Image Processing, 2008. ICIP 2008* , 1200–1203 (2008).

[25] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).

[26] Simoncelli, E. and Olshausen, B., "Natural Image Statistics and Neural Representation," *Annual Review of Neuroscience* **24**(1), 1193–1216 (2001).