A MOTION COMPENSATED APPROACH TO VIDEO QUALITY ASSESSMENT

Anush K. Moorthy and Alan C. Bovik

Dept. of Electrical and Computer Engineering,
Univ. of Texas at Austin,
Austin, TX 78712

ABSTRACT

We propose a new full reference video quality assessment algorithm (FR VQA) - the motion compensated structural similarity index (MC-SSIM). MC-SSIM evaluates spatial quality as well as quality along temporal trajectories. Its computationally simplicity makes it a prime choice for practical implementation. In this paper we describe the algorithm and evaluate its performance on a publicly available VQA dataset. We demonstrate that MC-SSIM correlates well with human perception of quality. We also explore its relationship to the human visual system and describe how a simple and efficient implementation of MC-SSIM can be realized.

Index Terms— Video quality assessment, spatio-temporal quality assessment, motion-compensation, structural similarity, H.264/AVC.

1. INTRODUCTION

Creation of algorithms that seek to predict the quality of a video such that the score produced by the algorithm correlates well with human perception of quality is referred to as objective video quality assessment (VQA). Subjective quality assessment refers to assessment of quality of videos by a set of human observers. Such assessment of quality produces a mean opinion score (MOS) which is representative of the *perceived* quality of the video. Evaluation of the performance of objective VQA algorithms involves correlating the scores produced by the algorithm on a set of videos with the subjective mean opinion scores. One such publicly available database of videos which has been widely used for testing VQA algorithms is the VQEG FRTV Phase-I dataset [1].

We classify VQA algorithms as full-reference (FR), no-reference (NR) and reduced reference (RR) algorithms. FR algorithms are those in which the reference and distorted videos are available for quality assessment. In RR algorithms the assumption is that some additional information regarding the reference video and/or distortion inducing process is available to the algorithm. NR refers to blind quality assessment, where the algorithm is supplied only with the distorted video. In each case, the goal of the algorithm is to produce a quality index for the video that correlates well with human perception of quality. In this paper, our aim is the creation of a FR VQA algorithm that achieves this goal.

One approach to VQA is to utilize an image quality assessment (IQA) algorithm that correlates well with the human perception of quality on a frame-by-frame basis. Indeed, in [2], such an approach was used for VQA using the popular image quality assessment (IQA) algorithm - the single scale structural similarity index (SS-SSIM) [3]. Assessing spatial quality alone does not fully capture distortions arising in videos, and it is imperative that VQA algorithms are designed to assess quality along the temporal direction as well. Realizing the importance of motion for VQA, in [2], the authors proposed an elementary weighting scheme using motion vectors from a block motion estimation process. Noticeable improvements were seen. SS-SSIM for VQA was modified in [4], where an alternative weighting scheme was presented for the spatial quality scores.

It is our belief that temporal-based weighting of spatial scores will fail to capture distortions in videos. For accurate VQA, it is not only essential to compute spatial quality, but also necessary to compute temporal quality along motion trajectories. Since algorithmic motion estimation is a difficult problem, it comes as no surprise that most VQA algorithms do not evaluate quality along these temporal trajectories. A recent VQA algorithm that is modeled after the motion processing mechanisms in the human visual system (HVS) was proposed in [5] and is referred to as motion-based video integrity evaluation (MOVIE) index. The MOVIE index uses a spatio-temporal filter bank to decompose the reference and distorted video into sub-bands; quality assessment is then performed over these subbands to produce a score for the video. Even though MOVIE has several features akin to the HVS, its computational complexity makes any practical implementation of the index difficult.

The simplicity of mean squared error (MSE) , coupled with its history of usage in the signal processing community makes MSE a popular choice for VQA, even though it has been pointed out that MSE correlates poorly with human perception of quality [6]. In this paper, our aim is to develop a VQA algorithm that not only offers the simplicity of MSE, but also the performance of a VQA algorithm that correlates well with human perception of quality. The pro-

posed algorithm utilizes the simple SS-SSIM for spatial quality assessment. Temporal quality assessment is carried out using a combination of motion-compensation and SS-SSIM. The computational simplicity of SS-SSIM [7] makes the proposed algorithm - motion compensated structural similarity index (MC-SSIM) highly practical. In this paper, we describe MC-SSIM and relate its design to HVS processing mechanisms. Further, by evaluating the performance of MC-SSIM on the publicly available VQEG FRTV phase-I database, we demonstrate that MC-SSIM performs extremely well in terms of correlation with human perception.

2. MOTION COMPENSATED STRUCTURAL SIMILARITY INDEX

Consider two videos which have been spatio-temporally aligned. We denote the reference video as $R(x, y, t)$ and the test video as $D(x, y, t)$ where the tuple (x, y, t) defines the location in space (x, y) and time t. The algorithm is defined for digital videos and hence the space coordinates are pixel locations and the temporal coordinate is indicative of the frame number. The test video is the sequence whose quality we wish to assess. Our algorithm is designed such that if $D = R$, i.e., if the reference and test videos are the same, then the score produced by the algorithm is 1. Any reduction from this perfect score is indicative of distortion in D. Also, the algorithm is symmetric, i.e., MC-SSIM (R, D) = MC-SSIM (D, R). We assume that each video has a total of N frames and a duration of T seconds. We also assume that each frame has dimensions $P \times Q$.

2.1. Spatial Quality Assessment

Spatial quality is evaluated in the following way. For each frame t from R and D and each pixel (x, y), the following spatial statistics are computed:

$$\mu_{R(x,y,t)} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} R(i,j,t)$$

$$\mu_{D(x,y,t)} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} D(i,j,t)$$

$$\sigma^2_{R(x,y,t)} = \frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(R(i,j,t) - \mu_{R(x,y,t)})^2$$

$$\sigma^2_{D(x,y,t)} = \frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(D(i,j,t) - \mu_{D(x,y,t)})^2$$

$$\sigma_{RD(x,y,t)} =$$

$$\frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(R(i,j,t)-\mu_{R(x,y,t)})(D(i,j,t)-\mu_{D(x,y,t)})$$

For spatial quality computation, w_{ij} is a $N \times N$ circular-symmetric Gaussian weighting function with standard deviation of 1.5 samples, normalized to sum to unity with $N = 11$.

Finally, we compute SSIM using equation 1 (next page).

In equation 1, $C_1 = (K_1 L)^2, C_2 = (K_2 L)^2$ are small constants; L is the dynamic range of the pixel values and $K_1 << 1$ and $K_2 << 1$ are scalar constants with $K_1 = 0.01$ and $K_2 = 0.03$. The constants C_1, C_2 and C_3 prevent instabilities from arising when the denominator is close to zero.

This computation yields a map of SSIM scores for each frame of the video sequence. The scores so obtained are denoted as $S(x, y, t), (x = \{1 \dots P\}, y = \{1 \dots Q\}, t = \{1 \dots N-1\})$.

2.2. Temporal Quality Assessment

We first estimate motion by applying a block-based motion estimation algorithm. The algorithm is applied on a frame-by-frame basis, where motion vectors are obtained for frame i from its preceding frame $i-1$. We seek to characterize the distortion in D, and hence motion estimation is performed only on the reference video. The block size is set at 8×8. For simplicity, assume that P and Q are multiples of the block size. The motion vectors so obtained are of integer pixel lengths.

In order to evaluate quality, we proceed as follows. For a frame i and for block (m_R, n_R) ($m_R = \{1, 2 \dots P/b\}$, $n_R = \{1, 2 \dots Q/b\}$), in video R, we compute the motion-compensated block (m'_R, n'_R) in frame $i-1$ by displacing the $(m_R, n_R)^{th}$ block by an amount indicated by the motion vector. A similar computation is performed for the corresponding $(m_D, n_D)^{th}$ block in D, thus obtaining the motion-compensated block (m'_D, n'_D). We then perform a quality computation between the blocks $B_R = (m'_R, n'_R)$ and $B_D = (m'_D, n'_D)$. This quality computation is performed using SS-SSIM. For such temporal quality computation, the window used in SSIM, w_{ij} is a $N \times N$ rectangular window normalized to sum to unity with $N = 8$. Hence, for each block we obtain a quality index corresponding to the perceived quality of that block, and for each frame we obtain a quality map of dimension $(P/b, Q/b)$. We denote the temporal quality map thus obtained as $T(x, y, t), (x = \{1 \dots P/b\}, y = \{1 \dots Q/b\}, t = \{1 \dots N-1\})$. A schematic diagram explaining the algorithm is shown in Fig. 1.

In order to pool the obtained scores, we utilize the technique from [8], where regions which disproportionately affect attention are given higher weight. Specifically, for each frame t, we compute:

$$T(t) = \frac{1}{|\xi|}\sum_{x,y \in \xi} T(x, y, t)$$

and

$$S(t) = \frac{1}{|\xi|}\sum_{x,y \in \xi} S(x, y, t)$$

$$S(x,y,t) = SSIM(R(x,y,t), D(x,y,t)) = \frac{(2\mu_{R(x,y,t)}\mu_{D(x,y,t)} + C_1)(2\sigma_{RD(x,y,t)} + C_2)}{(\mu^2_{R(x,y,t)} + \mu^2_{D(x,y,t)} + C_1)(\sigma^2_{R(x,y,t)} + \sigma^2_{D(x,y,t)} + C_2)} \quad (1)$$

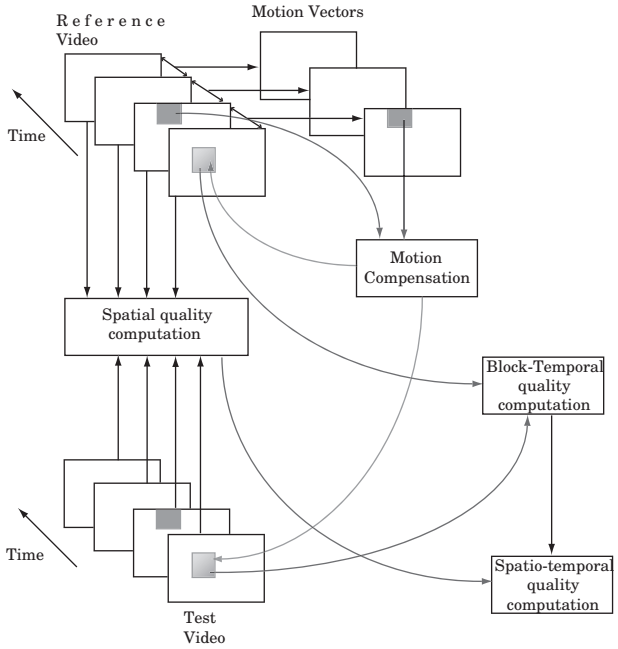

Fig. 1. Temporal quality computation: for the current block (dark grey) in frame i, the motion compensated block from frame $i-1$ (light grey) is recovered for both the reference and test video sequences using motion vectors computed from the reference sequence. Each set of blocks so obtained is evaluated for their quality.

where, ξ denotes the set consisting of the lowest 6% of the quality scores of each frame and $|\cdot|$ denotes the cardinality of the set. $S(t)$ and $T(t)$ are then averaged across frames to produce the spatial and temporal quality scores for the sequence - S and T. The final quality score for the video is $S \times T$.

Temporal quality is assessed not only on the 'Y' (luminance) component, but also on the color channels 'Cb' and 'Cr'. For each of the channels, motion estimation is performed to extract corresponding motion vectors and the algorithm as described in the previous section is applied. The final temporal quality score for the video is computed as:

$$T^{final} = 0.8 \times T^Y + 0.1 \times T^{Cb} + 0.1 \times T^{Cr}$$

where, T^Y, T^{Cb} and T^{Cr} are the temporal quality scores on each of the three color channels obtained as described above. A similar quality computation is undertaken for each of the three channels to assess spatial quality as well. The final spatial quality is computed as:

$$S^{final} = 0.8 \times S^Y + 0.1 \times S^{Cb} + 0.1 \times S^{Cr}$$

where, S^Y, S^{Cb} and S^{Cr} are the spatial quality scores on each of the three color channels obtained as described above. The weights assigned to each of the channels are exactly as in [2], though incorporating color in VQA remains an interesting avenue of research.

2.3. Relation to HVS

The efficient coding hypothesis states that the purpose of the early sensory processing is to recode incoming signals, so as to reduce the redundancy in representation [9]. Time-varying natural scenes possess high spatial and temporal correlation. Given the redundancy in videos and the efficient coding hypothesis, the principle that the visual pathway tries to improve efficiency of representation is compelling. It has been hypothesized that the Lateral Geniculate Neucleus (LGN), which lies in the area called the thalamus, performs such a temporal decorrelation [10]. The amount of feedback that the thalamus receives from the visual cortex is suggestive of possible feedback regarding motion estimates to thalamus, which will enable a reduction in redundancy at the LGN. The motion compensated approach used here for VQA roughly mirrors this process. SS-SSIM, which is used as the quality index, has some interesting relationships with natural scene statistics (NSS) which are used widely for studying HVS mechanisms [11].

2.4. Computational Complexity

The essence of the proposed algorithm is SS-SSIM. The computational complexity of SS-SSIM is $O(PQ)$. Sorting of scores required for percentile pooling can be performed with a worst-case complexity of $O(PQlog(PQ))$. The major bottleneck in MC-SSIM is the motion-estimation phase. However, we can completely avoid this bottleneck by re-utilizing motion vectors computed for compressed videos.Specifically, assume that we have a pristine compressed video which passes through a 'black-box'. We wish to assess the quality of the video at the output of the black box with respect to the compressed original. Since most video compression algorithms utilize a motion-compensated approach to video compression, the motion vectors used for this purpose are available in the compressed stream. Thus, in order to assess quality in such a situation, we simply decompress the reference and test video and read out the motion vectors from the reference video. MC-SSIM computation can then be performed as detailed above. In this case, the complexity of MC-SSIM is not much greater than that for SS-SSIM due to the novel motion-vector re-use. Further, as shown in [7], the SSIM index can be simplified without sacrificing performance.

Algorithm	SROCC	LCC	OR
PSNR	0.782	0.779	0.678
Proponent P8 (Swisscom)[1]	0.803	0.827	0.578
SS-SSIM (no weighting) [2]	0.788	0.820	0.597
SS-SSIM(weighted) [2]	0.812	0.849	0.578
SW-SSIM (dense - Y only)	0.837	0.810	0.622
MOVIE(Y only) [5]	0.833	0.821	0.644
MC-SSIM	0.848	0.853	0.597

Table 1. Evaluation of algorithm performance: Spearman Rank Ordered Correlation Coefficient (SROCC), Linear Correlation Coefficient (LCC) and Outlier Ratio (OR)

3. RESULTS

In order to evaluate the performance of MC-SSIM we utilize the Video Quality Experts Group (VQEG) dataset, which consists of 320 distorted videos along with the associated reference videos and subjective differential mean opinion scores (DMOS). The measures of performance are the Spearman rank rrdered correlation coefficient (SROCC), the linear (Pearson's) correlation coefficient (LCC) and root mean squared error (RMSE). LCC and RMSE are computed after transforming MC-SSIM scores using the logistic function as prescribed by the VQEG [1]. The results are seen in table 1. The table also lists scores for SS-SSIM [3] and speed weighted SSIM (SW-SSIM) [4]. In [3] and [4], the authors use a sparse sampling - i.e., not all pixel locations are sampled in a frame. Even though this may offer computational benefits; using such a sampling technique may hamper results. In any case, at this juncture it is unclear if such a sampling system will allow for a fair comparison of algorithms. Hence, for SW-SSIM, we list the scores when the frame is densely sampled - i.e., each pixel location from a frame is utilized. We find that for SS-SSIM a change in the sampling does not alter results much. Further, we also list the performance of the recently proposed MOVIE index. As seen in table 1, MC-SSIM extremely well in terms of correlation with human perception.

4. CONCLUSIONS

In this paper we proposed a new computationally efficient video quality assessment (VQA) algorithm, the Motion Compensated Structural Similarity Index (MC-SSIM). The algorithm was explained in detail its relationship to the human visual system was studied. The performance of the algorithm was evaluated on a publicly available VQA dataset and the algorithm was shown to correlate extremely well with the human perception of quality.

5. REFERENCES

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

[2] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assesssment based on structural distortion measurement," *Signal Processing: Image communication*, , no. 2, pp. 121–132, Feb. 2004.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[4] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[5] K. Seshadrinathan, *Video quality assessment based on motion models*, Ph.D. thesis, The University of Texas at Austin, 2008.

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

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

[8] A. K. Moorthy and A. C. Bovik, "Perceptually significant spatial pooling techniques for image quality assessment," *SPIE Conference on Human Vision and Electronic Imaging*, January 2009.

[9] J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network: Computation in neural systems*, vol. 3, no. 2, pp. 213–251, 1992.

[10] D. Dong and J. Atick, "Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Network: Computation in Neural Systems*, vol. 6, no. 2, pp. 159–178, 1995.

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