IEEE International Conference on Image Processing (ICIP12), Orlando, FL, Sept. 2012

3D-SSIM FOR VIDEO QUALITY ASSESSMENT

Kai Zeng and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
kzeng@engmail.uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

Effective and efficient objective video quality assessment (VQA) methods are highly desirable in modern visual communication systems for performance evaluation, quality control and resource allocation purposes. Simple VQA algorithms may be developed by direct extensions of still image quality assessment (IQA) approaches on a frame-by-frame basis. Advanced VQA methods take into account the temporal correlation and motion information contained in video signals but often lead to significantly increased computational complexity. Here we use a different approach to examine a video signal by considering it as a three-dimensional (3D) volume image. Specifically, we propose a 3D structural similarity (3D-SSIM) approach, which first creates a 3D quality map by applying SSIM evaluations within local 3D blocks, and then use local information content and local distortion based weighting methods to pool the quality map into a single quality measure. The resulting 3D-SSIM algorithm is computationally efficient and demonstrates highly competitive performance in comparison with state-of-the-art VQA algorithms when tested using four publicly available video quality databases[1].

Index Terms— video quality assessment, structural similarity, 3D volume image quality assessment, information content weighting

1. INTRODUCTION

With the exponential growth of visual communication applications, the demand for effective and efficient video quality assessment (VQA) technologies have been rapidly increasing in recent years. These technologies not only provide useful performance evaluations of visual communication systems, but can also be embedded into these systems as the core components in quality control, resource allocation, and system optimization tasks. Subjective VQA methods are reliable because the human visual systems (HVS) are the ultimate receivers in most applications. However, given the volume of video data being transmitted everyday, they are extremely slow and expensive. Objective VQA approaches provide a practical solution because they can automatically predict perceived video quality without human interactions. Here we mainly focus on full-reference (FR) VQA, where we have full access to the perfect-quality reference video when assessing the quality of a distorted video.

The design of VQA algorithms depends on how a video signal is interpreted. If we consider it as a stack of still images, then a natural approach is to apply still image quality assessment (IQA) algorithms on a frame-by-frame basis and then pool the frame level quality measures into a single quality score. However, this approach missed the temporal correlation between frames as well as the motion information contained in video signals, which are the most critical characteristics that distinguish a video sequence from a stack of independent still image frames. As a result, advanced VQA algorithms take into account temporal correlation or motion information. This can be done by combining multichannel spatiotemporal filtering and spatiotemporal just noticeable difference (JND) models [1, 2]. It can also be implemented by block- or optical flow-based motion estimation followed by weighted pooling based on models of human visual motion perception [3]. More sophisticated method combines both spatiotemporal filtering and motion estimation, and then incorporates both spatial and temporal distortion measures [4].

In this study, we consider a video signal as a 3D volume image and define a "region" in the image as a localized 3D block. We can then generate a 3D quality map by applying a block-wise quality measure within local regions. This is followed by a pooling stage that merge the quality map into an overall quality score. Recently, pooling has become an active research topic in IQA/VQA research. Most existing methods are based on the hypothesis that the regions that are more likely to attract visual attention should be assigned larger weights. The critical issue here is how visual attention is predicted, which may include a spectrum of approaches, ranging from saliency-based low-level vision models [5] to motion detection and object tracking based high-level cognitive methods [6, 4, 7, 8]. In [5], a number of different pooling strategies were compared in the context of IQA, and it was found that the approaches that lead to the most significant performance gain are local information content and local distortion weighted pooling, which are based on the assumptions that the image regions that contain more information (computed based on statistical image models) or more severe distortions are more likely to attract visual attention. Moreover, these methods can be implemented with low computational cost, which is often an important factor in real world deployment of VQA techniques. In this research, we extend these pooling strategies to VQA and find that they lead to consistent gain when tested using several independent video quality databases.

2. 3D-SSIM METHOD

The diagram of the proposed method, namely three-dimensional structural similarity (3D-SSIM) algorithm, is shown in Fig. 1. The input reference and distorted videos are first divided into non-overlapping 3D blocks. Within each block, a local 3D-SSIM measure and a local information content measure are computed. The local 3D-SSIM values collected from all blocks form a 3D quality map of the video, which are used to compute a local distortion-based weight map. Both the local information content and local distortion based weights are involved in the weighted pooling stage of the 3D-SSIM map, resulting in an overall 3D-SSIM score.

Let $\mathbf{x} = \{x_i | i = 1, \cdots, N\}$ and $\mathbf{y} = \{y_i | i = 1, \cdots, N\}$ be two sets of pixel values collected from corresponding 3D blocks from the reference and distorted videos, respectively. As in the spatial domain SSIM method [9], the local 3D-SSIM between the 3D

[1] Matlab implementation of the proposed method will be made available online at www.ece.uwaterloo.ca/˜z70wang/research/.

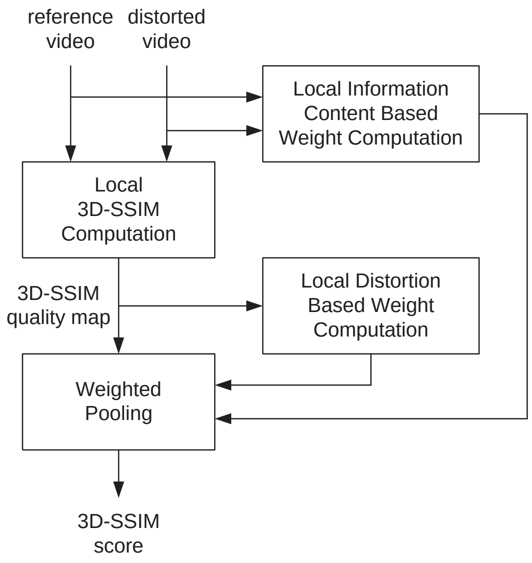
reference video
distorted video
Local Information Content Based Weight Computation
Local 3D-SSIM Computation
3D-SSIM quality map
Local Distortion Based Weight Computation
Weighted Pooling
3D-SSIM score

Fig. 1. Framework of 3D-SSIM algorithm.

blocks is computed as

$$\mathrm{S}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}\,, \tag{1}$$

where where μ_x, σ_x^2 and σ_{xy} represent the mean, variance and covariance of the image blocks, respectively, and C_1 and C_2 are small positive constants to avoid instability when the means and variances are close to zero.

Effective estimation of perceptual information content relies on good statistical models of both natural images and perceptual distortion channels [5]. While sophisticated models such as the Gaussian scale mixtures [5] are available for still images, they often lead to substantially increased complexity, which becomes a major barrier to overcome when applied to large volume video data. To achieve a good comprise between accuracy and simplicity, here we assume a simple model, where Gaussian distributed image source passes through an additive Gaussian channel and the mutual information between the source and received signals is employed to quantify the perceived information content. When this model is applied to local 3D image blocks of both the reference and distorted video signals, a simple computational model of the overall perceptual information content is given by [10]

$$w_{ic}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\log\left[\left(1+\frac{\sigma_x^2}{\sigma_0^2}\right)\left(1+\frac{\sigma_y^2}{\sigma_0^2}\right)\right]\,, \tag{2}$$

where, as in [11], σ_0^2 is a constant that accounts for the noise power of the additive Gaussian channel. This measure is computationally efficient because the values of σ_x^2 and σ_y^2 are readily available in the local 3D-SSIM computation.

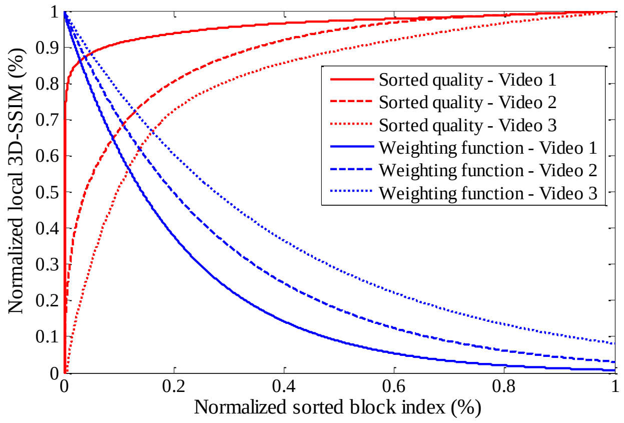

Fig. 2. Samples of sorted local 3D-SSIM curves and local distortion based weighting functions.

Previous studies had shown that assigning larger weights to higher distortion regions generally has positive effect on the performance of IQA/VQA algorithms [10, 5, 8]. In Fig. 2, the local 3D-SSIM measures computed from different regions are sorted in ascending order for three different distorted video sequences. It can be observed that the shapes of the ascending curves vary for different video sequences, which may depend on the nature of the videos as well as the type and level of the distortions. It was demonstrated in [8] the usefulness of adapting the weight assignment strategy based on the shape. In this paper, we propose to use a width-adapted exponential weighting function applied upon sorted block index. Assume that there are totally K 3D blocks extracted from the video, and let $\mathbf{y}_k$ be the block with the k-th lowest local 3D-SSIM value. The local distortion-based weighting function is defined upon the normalized index $\alpha_k = k/K$ by

$$w_d(\mathbf{y}_k) = e^{-\frac{|\alpha_k|}{\alpha_0}}\,, \tag{3}$$

where α_0 is a width parameter that controls the speed of falloff of the exponential function. As shown in Fig. 2, the ascending speeds of the sorted local 3D-SSIM curves vary for different video sequences. This motivates us to adapt the weighting function accordingly which can be readily implemented by adjusting α_0. Specifically, we preset an S^* parameter on the normalized 3D-SSIM value and find the corresponding block index α^* value on the sorted 3D-SSIM curve. We then compute the α_0 parameter by

$$\alpha_0 = \beta\alpha^*\,, \tag{4}$$

where β is a scaling parameter to control the relative widths of the sorted 3D-SSIM curve and the weighting function. Examples of the weighting functions computed based on the sorted 3D-SSIM curves are shown in Fig. 2.

Finally, the local 3D-SSIM map is pooled based on both local information content and local distortion based weighting and the overall 3D-SSIM measure of the entire video sequence is given by

$$\text{3D-SSIM} = \frac{\sum_{k=1}^{K}[w_{ic}(\mathbf{x}_k,\mathbf{y}_k)]^{\mu}[w_d(\mathbf{y}_k)]^{\nu}S(\mathbf{x}_k,\mathbf{y}_k)}{\sum_{k=1}^{K}[w_{ic}(\mathbf{x}_k,\mathbf{y}_k)]^{\mu}[w_d(\mathbf{y}_k)]^{\nu}}\,. \tag{5}$$

where μ and ν are two parameters used to control the relative importance of the two weighting functions.

3. IMPLEMENTATION AND EXPERIMENT

The implementation details of the proposed 3D-SSIM algorithm are as follows. As in the default SSIM implementation [12], the in-

Table 1. Test VQA databases. SRC denotes the number of source reference videos and HRC denotes the number of distorted videos created from each source video.

Database	# of video	SRC	HRC	Resolution
VQEG FR-TV I	320	20	16	480i, 576i
IRCCyN/IVC	192	24	7	720×576
EPFL-PoliMI	156	16	9	CIF, 4CIF
LIVE	150	10	15	768×432p

Table 2. PLCC performance comparison of VQA algorithms

Database	VQEG	IRCCyN	EPFL-PoliMI	LIVE
PSNR	0.7683	0.4160	0.7351	0.5621
SSIM [9]	0.8215	0.5012	0.6781	0.5444
SSIM [12] (auto-scale)	0.8113	0.6139	0.6770	0.7177
VQM [13]	0.8170	0.4850	0.8434	0.7236
MOVIE [4]	0.8210	0.4850	0.9210	0.8116
Yu *et al.* [7]	0.8170	0.7680	0.9470	**0.8450**
3D-SSIM (no weighting)	0.8079	0.6212	0.7591	0.7026
3D-SSIM (w_{ic} only)	0.8203	0.7357	0.8136	0.7497
3D-SSIM (w_d only)	0.8295	0.7209	0.9091	0.7832
3D-SSIM	**0.8403**	**0.8194**	**0.9621**	0.8353

put reference and distorted video signals first go through an automatic downsampling (or auto-scale) process on a frame-by-frame basis. This is followed by dividing the 3D volume image into non-overlapping $7 \times 7 \times 7$ blocks, within which the local 3D-SSIM measure (1), the local information content weighting function (2), and the local distortion weighting function (3) are calculated. The parameters C_1, C_2 and σ_0^2 are the same as in the default SSIM [12] and VIF [11] implementations. The other parameters are obtained empirically to optimize the performance on the EPFL-PoliMI VQA database and are given by $S^* = 0.95$, $\beta = 0.4$, $\mu = 4.5$ and $\nu = 1$, respectively. The information content weights go through another normalization step so that its value is between 0 and 1 before being plugged into the final computation of the overall 3D-SSIM measure.

The proposed approach was tested on four publicly available VQA databases, as described in Table 1, where the main distortion types include standard video compression (MPEG and H.264) at different bit rates and simulated transmission errors. Pearson linear correlation coefficient (PLCC) and Spearman's rank correlation coefficient (SRCC) between objective and subjective quality scores are adopted as the evaluation criteria, where the subjective scores are in the form of either mean opinion score (MOS) or difference of mean opinion score (DMOS) (difference between the MOS values of the reference and distorted videos). To compute PLCC, a nonlinear regression is carried out between subjective and objective scores using the modified logistic regression model introduced in [11].

The PLCC and SRCC evaluation results are given in Tables 2 and 3, respectively. First, the proposed 3D-SSIM approach in (5) is compared with other pooling options (that are based on the same local 3D-SSIM map), where no weighting or only one of the weighting approaches (w_{ic} in (2) or w_d in (3) only) is applied. Apparently, either information content or distortion based weighting scheme significantly improves upon the no-weighting case and the best results are obtained when both of them are applied. The proposed 3D-SSIM algorithm is also compared with six other VQA approaches, including peak signal-to-noise-ratio (PSNR), direct SSIM [9], SSIM with auto-scaling [12], video quality model (VQM) [13], MOtion-based Video Integrity Evaluation index (MOVIE) [4], and a most recent method proposed by Yu *et. al* [7]. The best results obtained for each database are highlighted in bold. It can be observed that 3D-SSIM appears to be the most reliable measure across all four databases and achieves the best performance in most cases. The scatter plots of 3D-SSIM values versus subjective quality scores over the four databases, together with the nonlinear fitting functions, are shown in Fig. 3.

Table 3. SRCC performance comparison of VQA algorithms

Database	VQEG	IRCCyN	EPFL-PoliMI	LIVE
PSNR	0.7714	0.4510	0.7440	0.5398
SSIM [9]	0.7880	0.5126	0.6770	0.5257
SSIM [12] (auto-scale)	0.7919	0.6058	0.6949	0.6947
VQM [13]	0.7760	0.4820	0.8383	0.7026
MOVIE [4]	0.8330	0.5930	0.9200	0.7890
Yu *et al.* [7]	0.8030	0.7910	0.9450	0.8180
3D-SSIM (no weighting)	0.7804	0.6147	0.7483	0.6810
3D-SSIM (w_{ic} only)	0.8147	0.7143	0.8003	0.7397
3D-SSIM (w_d only)	0.8208	0.7012	0.9016	0.7712
3D-SSIM	**0.8396**	**0.7916**	**0.9608**	**0.8244**

It is worth emphasizing that the highly competitive performance of 3D-SSIM is obtained with vastly reduced computational complexity. Our Matlab implementation of the 3D-SSIM algorithm takes around 4.64 seconds (excluding data loading time) to evaluate a video sequence of 768 × 432 in spatial resolution and 217 frames in length on a computer with Intel Core2 Duo CPU E8600 processor at 3.33GHz. This is estimated to be only less than 1% and 0.1% of the well known VQM [13] and MOVIE [4] algorithms, respectively. This could be a critical advantage in many real world applications.

4. CONCLUSION

We propose a novel VQA algorithm namely 3D-SSIM, which regards a video signal as a 3D volume image and combines local SSIM based quality measure with local information content and distortion based pooling methods. The resulting 3D-SSIM measure is computationally efficient and achieves highly competitive performance when compared with state-of-the-art VQA approaches. One potential drawback of the proposed approach is the memory requirement to store 3D volume data. This problem may be alleviated by dividing the video sequence into segments based on the size of the 3D block involved in the computation. In the future, the proposed method may be improved by incorporating more accurate statistical models in the estimation of local information content and investigating more ad-

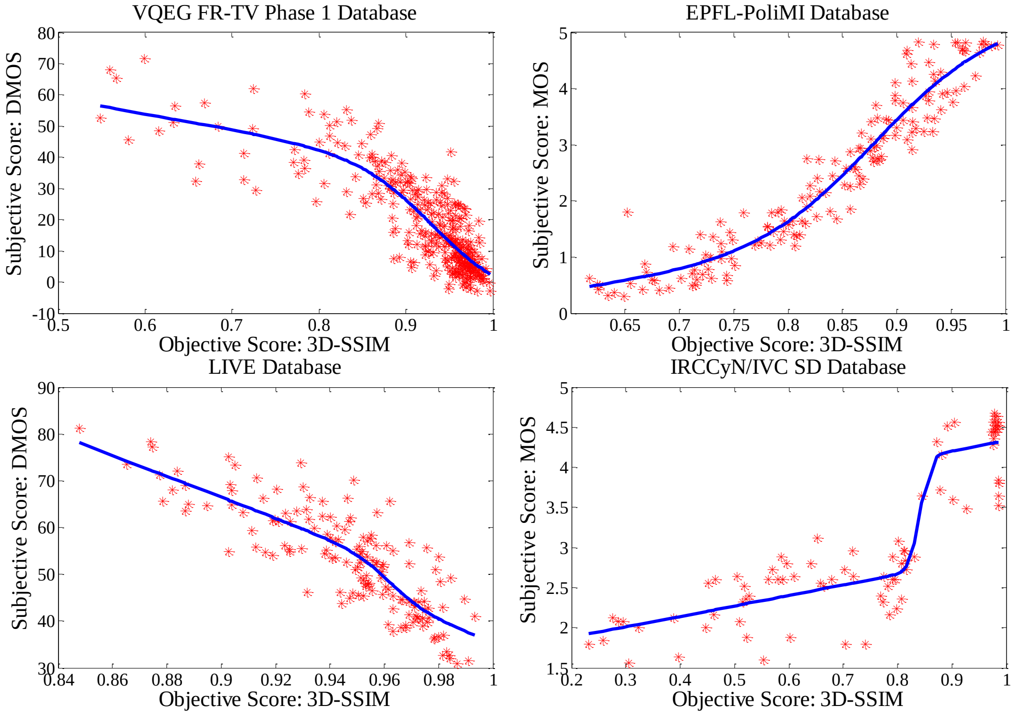

Fig. 3. Scatter plots of 3D-SSIM versus subjective score for four VQA databases.

vanced adaptive strategies for local distortion based pooling.

5. ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and in part by Ontario Ministry of Research & Innovation Early Researcher Award Program, which are gratefully acknowledged.

6. REFERENCES

[1] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[2] Y. Zhao, L. Yu, Z. Chen, and C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 21, no. 12, pp. 1890–1902, Dec. 2011.

[3] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication,* Special issue on objective video quality metrics, vol. 19, no. 2, pp. 121–132, Feb. 2004.

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

[5] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

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

[7] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing attended and global stimuli in perceived video quality assessment," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1269–1285, Dec. 2011.

[8] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Spatio-temporal quality pooling accounting for transient severe impairments and egomotion," in *18th IEEE Inter. Conf. on Image Process. (ICIP)*, Sept. 2011, pp. 2509–2512.

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

[10] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *2006 IEEE Inter. Conf. on Image Process.*, Oct. 2006, pp. 2945–2948.

[11] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "The SSIM index for image quality assessment," `http://www.cns.nyu.edu/~lcv/ssim/`.

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