A VERY LOW COMPLEXITY REDUCED REFERENCE VIDEO QUALITY METRIC BASED ON SPATIO-TEMPORAL INFORMATION SELECTION

Mengmeng Wang, Fan Zhang and Dimitris Agrafiotis

Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB, UK.
{mw13612,Fan.Zhang,D.Agrafiotis}@bristol.ac.uk

ABSTRACT

This paper presents a reduced reference video quality metric that exploits contrast and motion sensitivity characteristics of the HVS to perform a spatio-temporal selection of reference data. Spatio-temporal selection is realised through mapping of wavelet subbands to contrast sensitivity and through motion analysis. The proposed method is integrated with a modified SSIM-based framework to produce STIS-SSIM, a very low complexity reduced reference metric. The metric is shown to offer significant performance improvement over many existing full reference and reduced reference video quality metrics when tested on the LIVE video database.

Index Terms— Video quality assessment, reduced reference, spatio-temporal information selection

1. INTRODUCTION

Video quality assessment (VQA) is a research area of significant interest to the video compression community [1], as it forms the basis of rate-quality optimisation algorithms, quality of service (QoS) evaluation and codec performance assessment. VQA methods are generally classified into subjective and objective approaches. The latter are further categorised as full-reference (FR), reduced-reference (RR) or no-reference (NR) depending on the availability or otherwise of full or partial reference data during the assessment process [2].

Popular FR metrics include peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [3], both of which have been widely employed in various applications. More recent FR metrics include the visual signal-to-noise ratio (VSNR) [4], video quality model (VQM) [5], motion tuned spatial-temporal quality assessment method (MOVIE) [6], spatio-temporal most-apparent-distortion (STMAD) [7] and perception-based video quality metric (PVM) [8].

RR metrics are utilised in those cases where only partial reference data can be used, for example when assessing video quality at the receiving end of a wireless channel. Various spatial and temporal features are commonly extracted to form the reduced reference information [9–11]. RR quality metrics are generally required to have low complexity whilst maintaining a reasonable correlation with subjective opinion scores. One of the most recent RR video quality assessment methods is the STRRED method [12] which employs spatio-temporal entropic differences for performing the quality assessment.

In this paper we present STIS-SSIM, a reduced reference video quality metric that applies adaptive spatio-temporal information selection (STIS) for forming the reduced reference data. STIS selects reference data for a segment of an image sequence by analysing the energy in the medium frequency bands of the wavelet-transformed frame(s) of that segment of the sequence. STIS also takes into account frame differences for ranking the reference data. The selection method is efficiently combined with a modified SSIM metric to form STIS-SSIM. We tested the performance of the proposed RR metric on the LIVE video database where it was found to offer significant improvement in correlation with subjective scores relative to other RR metrics. The computational complexity of the proposed metric was found to be lower compared to that of seven other tested quality metrics.

The rest of the paper is organised as follows. Section 2 describes the proposed quality metric (STIS-SSIM) in detail. In Section 3 the performance of STIS-SSIM is presented including correlation analysis, complexity and significance test. Finally, conclusions are given in Section 4 alongside suggestions for future work.

2. PROPOSED ALGORITHM

The proposed method exploits two properties of the human visual system (HVS) for characterising and ranking reference data. The first property is the high contrast sensitivity displayed by the HVS to medium spatial frequencies, as expressed by the contrast sensitivity function (CSF) [13]. The second property is the high visual attention sensitivity that the HVS shows to motion.

STIS adaptively selects reference data according to their ranking and the desired amount of side information. The latter is expressed as a percentage of the full reference data. The side information consists of the selected reference scalars and corresponding pixel block position coordinates. At the decoder, the position coordinates are used for selecting decoded

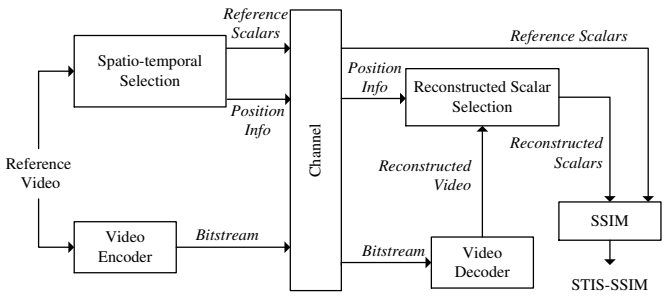

Fig. 1: Block diagram of the proposed reduced reference quality metric.

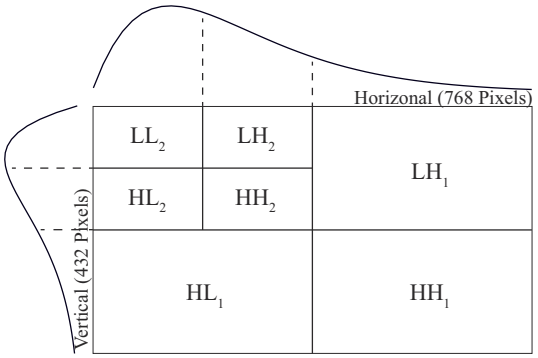

Fig. 2: Mapping of the CSF curve onto DWT subbands based on the viewing configuration used in the LIVE video database.

pixel blocks. These are then compared to the selected reference scalars using a modified SSIM metric. A block diagram describing the STIS-SSIM framework is given in Figure 1. Below we describe the proposed method in detail.

2.1. Spatial Selection

STIS performs spatio-temporal selection of reference data for a given segment of a video. A segment can range from a few frames to the whole sequence. Each frame in the segment is transformed into different frequency bands using the Discrete Wavelet Transform (DWT) with two levels of decomposition.

For a known viewing distance d, the relationship between the CSF and the wavelet subbands is characterised by the following expression [14, 15] :

$$f_{\max,w} = \frac{w_i}{4 \times \arctan(\frac{w}{2 \times d})}, \quad (1)$$

$$f_{\max,h} = \frac{h_i}{4 \times \arctan(\frac{h}{2 \times d})}. \quad (2)$$

Here w and h are the width and height of the display, and w_i and h_i are the horizontal and vertical spatial resolution of the video. $f_{\max,w}$ and $f_{\max,h}$ represent the highest spatial frequencies (in cycles per degree) that can be rendered in the horizontal and vertical directions given a particular viewing configuration (display size, video resolution and viewing distance). Given equations (1) and (2), the CSF function can be mapped to the DWT subbands using the experimental results of [16], with $f_{\max,w}$ and $f_{\max,h}$ corresponding to the first level subbands. Fig. 2 shows this mapping for a two-level DWT decomposition using the viewing configuration employed in the LIVE database subjective tests. The dashed lines indicate the subbands of interest[1].

To select the reference information each frame is firstly segmented into blocks of identical size. Each block is then characterised and ranked by the energy of the wavelet coeficients in the subbands of interest that correspond to that block of pixels. More specifically for each block n we find the subband coordinates (i, j) in level m of the decomposition that correspond to that block and calculate the energy of the coefficients at each of these coordinates (i, j) in subbands LH_m, HL_m and HH_m. We rank blocks according to the maximum of these energies $E(n)$:

$$E(n) = \max_{i,j}(|B_{\mathrm{LH}_m}(i,j)| + |B_{\mathrm{HH}_m}(i,j)| + |B_{\mathrm{HL}_m}(i,j)|). \quad (3)$$

Reference blocks with a higher $E(n)$ value are ranked higher. We have found that applying a mean filter with a 4×4 window to all the DWT subband coefficients in (3) prior to the calculation of $E(n)$ improves the performance.

2.2. Spatio-Temporal Selection

Temporal characterisation of pixel blocks is based on simple frame differences. More specifically we calculate:

$$\mathrm{FD(t)} = \sum_{x,y}(|I_t(x,y) - I_{t-1}(x,y)| + |I_t(x,y) - I_{t+1}(x,y)|), \quad (4)$$

where $\mathrm{FD(t)}$ is the frame difference for current frame t and $I_t(x, y)$ is the luminance value for pixel coordinate (x, y) in frame t.

Spatio-temporal ranking and selection of blocks is then performed according to the spatio-temporal information value of each block, calculated as shown in (5).

$$\mathrm{STI}(n,t) = \frac{E(n,t)}{E_{\max}(t)} \times \frac{\mathrm{FD}(t)}{\mathrm{FD}_{\max}}. \quad (5)$$

$\mathrm{STI}(n, t)$ is the spatio-temporal information value of block n within frame t. $E_{\max}(t)$ and $\mathrm{FD}_{\max}$ are maxima, calculated as shown in (6).

$$\begin{cases} E_{\max}(t) = \max_n\{E(n,t)\} \\ \mathrm{FD}_{\max} = \max_t\{\mathrm{FD}(t)\} \end{cases}. \quad (6)$$

The spatio-temporal selection method is described in detail in Algorithm 1.

[1] It should be noted that, due to the limited spatial resolution of videos in the LIVE database, the number of DWT decomposition levels is such that the selected subbands do not correspond perfectly to the peak region of the CSF.

Algorithm 1 Spatio-temporal selection algorithm.

Input:
Target number of reference scalars: $S_{\max}$;
Spatio-temporal information value of each block: $\mathrm{STI}(n,t)$;
Number of scalars representing each block: S_{block};
Number of blocks in frame: N;
Number of frames in sequence: M.

Output:
Selected blocks with their positions:
block 1 at $(\hat{n}_1,\hat{t}_1)$, block 2 at $(\hat{n}_2,\hat{t}_2)$, $\cdots$, block l at $(\hat{n}_l,\hat{t}_l)$, $\cdots$.

Calculate target number of blocks $B_{\max}=\frac{S_{\max}}{S_{\mathrm{block}}}$;
Calculate selection ratio $r=\frac{B_{\max}}{N\cdot M}$;
Divide sequence into $K=M\cdot\sqrt{r}$ segments, each of which contains $\frac{M}{K}$ consecutive frames;
Calculate target block numbers for each segment: $B_{\max,s}=\frac{B_{\max}}{K}, k=1,\cdots,K$;
Rank $\{\mathrm{STI}(n,t)\}$ for all blocks in descending order and place all blocks in a selection pool;
Set the number of selected blocks $B_{\mathrm{sel}}=0$;
Set the number of selected blocks in each segment $B_{\mathrm{sel,k}}=0, k=1,\cdots,K$;
while $B_{\mathrm{sel}}<B_{\max}$ **do**
    Select the block (assume within segment k) with the maximum STI in the selection pool: $(\hat{n}_i,\hat{t}_i)$;
    $B_{\mathrm{sel}}=B_{\mathrm{sel}}+1$;
    $B_{\mathrm{sel,k}}=B_{\mathrm{sel,k}}+1$;
    Remove all the blocks at the same position in 10 neighbouring frames (5 frames before frame t and 5 frames after, if applicable);
    if $B_{\mathrm{sel,k}}\geq B_{\mathrm{max,s}}$ **then**
        Remove all the blocks within segment K from the selection pool;
    end if
end while

2.3. Integration with SSIM

The proposed spatio-temporal selection method is integrated with a SSIM quality metric [3]. To reduce the size of the side information (the amount of reference data used) we do not include the pixel values of the selected blocks in the reduced reference data, but instead we use the corresponding DC coefficients at level m of the DWT decomposition. The luminance values obtained from the inverse DWT of these reference coefficients are compared with the luminance values obtained by the inverse DWT of the corresponding DC coefficients of the reconstructed (distorted) frames. This reduces the size of the side information to around 6% of the original reference pixels when m equals 2.

In an effort to further reduce the amount of reference data used by STIS-SSIM we observe that in the original SSIM algorithm a Gaussian filter with a window size of 11 is applied to both reference and test frames, before quality indices are calculated. This leads to high correlation between the central pixels in each test block and their neighbouring pixels. Based on this observation, we only include the central DC coefficients of the selected blocks in the reference data. We calculate the SSIM index using the equation below:

$$\mathrm{SSIM}(l)=\frac{(2\mu_o(l)\mu_d(l)+C_1)(2\sigma_{o,d}(l)+C_2)}{(\mu_o^2(l)+\mu_d^2(l)+C_1)(\sigma_o^2(l)+\sigma_d^2(l)+C_2)}. \quad (7)$$

$\mu_o(l)$ and $\mu_d(l)$ represent the average luminance values obtained from the central DC coefficients of reference and distorted block l at position $(\hat{n}_l,\hat{t}_l)$. $\sigma_o^2(l)$ and $\sigma_d^2(l)$ are the variance of these blocks. C_1 and C_2 are two parameters used for stabilising the division with weak denominators. SSIM indices from all selected blocks are averaged to give the final sequence-level quality index as follows:

$$\text{STIS-SSIM}=\sum_{l=1}^{B_{\max}}\frac{\mathrm{SSIM}(l)}{B_{\max}}. \quad (8)$$

3. RESULTS AND DISCUSSION

The performance of STIS-SSIM was evaluated using the LIVE video database [17]. The database contains 150 distorted videos from 10 references with a resolution of 768×432@25/50p. Distortion types include compression artefacts generated by MPEG-2 and H.264 and transmission errors over simulated IP or wireless networks. A two level DWT decomposition was applied as shown in Fig.2 . Video frames were segmented into 64×64 blocks. These correspond to 16×16 blocks of wavelet coefficients at the second level of the DWT decomposition. The central 4×4 coefficients of these blocks from the DC band were used as reference scalars. The performance and complexity of the proposed metric was compared with that of eight existing objective quality assessment methods. These are PSNR, SSIM [3], VSNR [4], VQM [5], MOVIE [6], PVM [8], STMAD [7] and STRRED [12]. The performance analysis rules described in [8, 17, 18] were followed for the assessment, with a weighted least squares approach being used to minimise the error of a logistic fitting function of subjective DMOS and objective quality values.

Table 1 and Table 2 show the correlation performance of the proposed method using different amounts of reference scalars. P represents the number of pixels in the original video frames and $P'=P/256$ stands for the number of reference scalars used if the central DC wavelet coefficients of all pixel blocks are selected as reference.

Table 1: Spatial selection results.

Scalar No.	P'	$0.5P'$	$0.4P'$	$0.3P'$	$0.2P'$
SROCC	0.7021	0.7384	0.7650	0.7852	0.7999
Scalar No.	$\mathbf{0.1}P'$	$0.05P'$	$0.02P'$	$0.01P'$	
SROCC	**0.8033**	0.7895	0.7810	0.7936	

Table 2: Spatio-temporal selection results.

Scalar No.	$0.02P'$	$\mathbf{0.01}P'$	$0.006P'$
SROCC	0.8138	**0.8154**	0.8092
Scalar No.	$0.003P'$	$0.001P'$	$0.0008P'$
SROCC	0.7853	0.7600	0.7434

Table 3: F-test results for the tested video metrics at 95% confidence level.

Metric	PSNR	SSIM	VSNR	VQM	MOVIE	PVM	STMAD	STIS-SSIM (best)
PSNR	-	0	0	-1	-1	-1	-1	-1
SSIM	0	-	0	-1	-1	-1	-1	-1
VSNR	0	0	-	0	-1	-1	-1	-1
VQM	1	1	0	-	0	0	0	0
MOVIE	1	1	1	0	-	0	0	0
PVM	1	1	1	0	0	-	0	0
STMAD	1	1	1	0	0	0	-	0
STIS-SSIM (best)	1	1	1	0	0	0	0	-

Table 1 shows results when only spatial selection is applied, with an equal amount of information from every frame being included in the reference data. It is interesting to note that for the case of spatial selection SROCC (Spearman's rank correlation coefficient) performance increases when less information is used as reference (the best result appears for the case of $0.1P'$ which corresponds to around 0.04% of information). Spatio-temporal selection (STIS-SSIM) improves the performance as shown in Table 2 and reduces further the amount of reference information needed. The best SROCC value with STIS-SSIM (0.8154) is obtained at $0.01P'$ (0.004% of reference information), which is defined as STIS-SSIM (best) for further comparison.

Table 4: Performance of all tested FR and RR metrics on the LIVE database.

Metric	SROCC	LCC	RMSE	Complexity
PSNR	0.5398	0.5613	9.1100	1
SSIM	0.5252	0.5414	9.2876	13
VSNR	0.6881	0.6726	8.0683	65
VQM	0.7748	0.7583	6.6931	681
MOVIE	0.7890	0.8112	6.4439	2206
PVM	0.8045	0.8160	6.3723	632
STMAD	**0.8204**	0.8245	6.2433	808
STRRED (best)	0.8056	n/a	n/a	97
STIS-SSIM (best)	0.8154	**0.8290**	**6.1674**	9

Table 4 presents a summary of the performance results obtained with all tested FR and RR quality metrics using the LIVE database. Performance results are presented in terms of the Linear Correlation Coefficient (LCC), SROCC, and the Root Mean Squared Error (RMSE). It can be seen that STIS-SSIM achieves the best performance of all tested metrics in terms of LCC and RMSE, and the second best in terms of SROCC. A more detailed performance comparison with RR metric STRRED [12] is given in Fig. 3. The graph shown compares SROCC results versus the amount of reference information used by each RR metric. It can be seen that the proposed metric outperforms STRRED.

Table 3 shows the results of a significance test that we conducted for all tested metrics. More specifically we performed an F-test (we followed the rule in [8, 17]) on the residual between the average DMOS of the LIVE subjective assessment and the predicted DMOS given by the tested objective quality metrics. Each value in Table 3 indicates the significance of the difference in performance between the quality

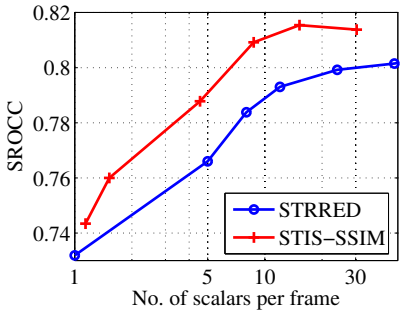

Fig. 3: STIS-SSIM vs STRRED.

metrics stated in the top row and left most column of the table at the 95% confidence level. A "1" suggests that the metric in the row is superior to that in the column, a "-1" suggests the opposite. A "0" indicates that there is no significant difference between the performance of the two metrics. According to the results of Table 3, MOVIE, PVM, STMAD and STIS-SSIM (best) are statistically equivalent, and they all outperform PSNR, SSIM and VSNR.

Finally, for most application scenarios of RR metrics, it is also important to analyse the computation complexity of the metric. Table 4 gives a summary of the relative complexity of all the tested quality metrics. Complexity was measured as the average execution time on an Intel Core i7-2600 CPU @3.40GHz PC, and is normalised relative to the execution time of PSNR. The complexity of the reduced reference metrics was calculated for the level of reference that offered the best performance. All test metrics were realised in Matlab except MOVIE (realised in C). It can be observed that the proposed method is less complex than most test metrics.

4. CONCLUSIONS

In this paper we presented STIS-SSIM, a novel reduced reference video quality metric that exploits the contrast and motion sensitivity characteristics of the HVS for reducing the amount of reference information needed and for producing good results. The proposed method has low computational complexity and provides superior performance to most existing full reference and reduced reference quality metrics when tested on the LIVE database. Future work will focus on integrating the selection part of the metric (STIS) with more advanced perception-based video quality metrics.

5. REFERENCES

[1] H. R. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)*, CRC Press, Inc., Boca Raton, FL, USA, 2005.

[2] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification,

review, and performance comparison," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165–182, June 2011.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, April 2004.

[4] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, Sept 2007.

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

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

[7] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Sept 2011, pp. 2505–2508.

[8] F. Zhang and D. R. Bull, "Quality assessment methods for perceptual video compression," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 39–43.

[9] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 1010–1013.

[10] M. Rohani, A. Nasiri Avanaki, S. Nader-Esfahani, and M. Bashirpour, "A reduced reference video quality assessment method based on the human motion perception," in *Telecommunications (IST), 2010 5th International Symposium on*, Dec 2010, pp. 831–835.

[11] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 1, pp. 71–83, Jan 2008.

[12] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 4, pp. 684–694, April 2013.

[13] R. J. Snowden, P. Thompson, and T. Troscianko, *Basic vision: an introduction to visual perception*, Oxford ; New York : Oxford University Press, 2006.

[14] D. Bull, *Communicating Pictures: A course in Image and Video Coding*, Academic Press Inc, 2014.

[15] Y. Zhang, E. Reinhard, and D. Bull, "Perception-based high dynamic range video compression with optimal bit-depth transformation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Sept 2011, pp. 1321–1324.

[16] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, Andrew B. Watson, Ed., pp. 179–206. MIT Press, Cambridge, MA, USA, 1993.

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

[18] ITU-T Study Group 12, "Evaluation of new methods for objective testing of video quality: objective test plan," Tech. Rep., ITU-T, 1998.