PERCEPTUAL QUALITY METRIC FOR H.264 LOW BIT RATE VIDEOS

Ee Ping Ong, Weisi Lin, Zhongkang Lu, and Susu Yao, Mei Hwan Loke

Institute For Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email: {epong, wslin, zklu, ssyao, mhloke}@i2r.a-star.edu.sg

ABSTRACT

This paper proposed an objective video quality metric designed for automatically assessing the perceived quality of digitally compressed multimedia videos using H.264 video compression. The rationale in proposing perceptual-based metric is because traditional measure, peak signal-to-noise ratio (PSNR), has been found to correlate poorly with subjective quality ratings, particularly at much lower bit rates. In this paper, computational models have been applied to emulate human visual perception based on a combination of local and global modulating factors. The proposed video quality metric has been tested on CIF and QCIF video sequences compressed using H.264 video compression technique at various bit rates (24-384 Kbps) and frame rates (7.5-30Hz). Performance of the proposed metric with respect to subjective scores will be reported and a comparison with PSNR and also the video structural similarity method (being one of the best video quality metric for high bit rate videos recently reported in the literature) will also be provided in this paper.

1. INTRODUCTION

Visual quality (or distortion) evaluation plays an important and determinative role in shaping most algorithms for image/video manipulations, visual quality control within an encoder, and distortion assessment for decoded signal. Since human eyes are the end receiver of most decoded images/videos, it is desirable to develop visual quality metrics that correlate better with human's visual perception than the conventional PSNR measure, which has been found to correlate poorly with subjective quality ratings [3].

For better prediction of the quality for decoded visual signal, a number of approaches have been tried to model the temporal, spatial and masking characteristics of human vision [7, 9, 14, 17, 18], to evaluate common coding artefacts [5, 19], and also to combine these two paradigms [13, 20]. However, the evaluation so far has been concentrated in TV types of signal (e.g., [14]). Characteristics of human vision have also been modelled for image and video compression [2, 4]).

Although the literature contains large number of perceptual-based video quality metrics, none of these have demonstrated their use for videos compressed with low bit rates, small frame size (e.g. QCIF video format), and low frame rates. However, there are numerous published works [7, 17, 13, 19, 20, 16, 18] (including those in VQEG [14]) that have been designed for TV-sized videos at much higher bit rates and full TV frame rates. In [16], a method known as video structural similarity (VSSIM) method has been demonstrated to perform better than the best method reported in VQEG [14] for TV-sized videos.

There will be increasingly more applications and services of mobile visual signals, particularly with the migration of mobile services from 2G to 3G that is able to provide video streaming/transmission/reception and mobile video conferencing, and therefore this creates the need for measuring/monitoring the quality of video coded at low bit rates and frame rates, particularly by the content service providers during the creation of the digital content to be archived and streamed. Exploration of objective video quality metrics for such "multimedia (MM)" videos is currently in progress in VQEG [15].

In the rest of this paper, Section 2 presents a description of the proposed perceptual video quality metric, Section 3 presents the results, while the last section concludes this paper.

2. PROPOSED PERCEPTUAL QUALITY METRIC

2.1. Perceptual Video Quality Metric

The overall objective video rating for a colour video sequence, Q, is given by a weighted averaging of the objective video quality rating for each colour's $q_j(t)$, for j=1,…,n, where n is the maximum number of colour components:

$$Q = \sum_{j=1}^{n} \alpha_j \left(\sum_{t=i.(f_f/f_r)}^{N} \left[q_j(t) \right] \Big/ N_t \right), \quad i = 1,2,...$$

where α_j denotes the weighting for each colour components, N is the total number of frames in the original video sequence, $N_t = N/\left(f_f / f_r\right)$ is the total number of frames in the coded video sequence, f_r is the frame rate at which the video is being coded, f_f is the full frame rate of the original video sequence, and $q_j(t)$ is the objective video quality rating for colour component of each frame:

$$q_j(t) = D_j(t).F_{BI,j}(t).F_{RF,j}(t)$$

where D is the distortion-invisibility (derived from spatial-textural, colour and temporal maskings), F_{BF} is the block-fidelity, and F_{RF} is the content richness fidelity. The latter two terms are global measures that modulate the final distortion-invisibility value to give the video quality measure for each frame. These global measures are being introduced because it has been observed that the pictorial quality perceived by human visual system is also affected by the overall general impression of the viewed video stream on humans. In addition, recent studies have shown that human visual system awards higher response to more salient image locations and features [6].

2.2. Distortion-Invisibility

The distortion-invisibility feature measures the average amount of distortion that may be visible at each pixel with respect to a visibility threshold. The distortion-invisibility measure, *D(t)*, for each frame *t* of the video is given by:

$$D(t)=\left\{1\Bigg/\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left[\gamma_1+\frac{\hat{d}(x,y,t)}{\gamma_2+T(x,y,t)}\right]\right\}$$

$T(x,y,t)$ is the visibility threshold at a particular pixel location (x,y) and time interval t, W and H are width and height of the video frame respectively, γ_1 and γ_2 are included to prevent possible division by zero in the equation. Also,

$$\hat{d}(x,y,t)=\begin{cases}0 & \text{if } d(x,y,t)\le(1-s)\cdot T(x,y,t)\\ d(x,y,t) & \text{otherwise}\end{cases}$$

where s is the soft-criterion to avoid clipping of the difference data near the visibility threshold T, and $d(x,y,t)$ is the difference between a frame in the test video I_d and the reference video I_o at the same pixel location (x,y) and time t and is defined as:

$$d(x,y,t)=\left|I_o(x,y,t)-I_d(x,y,t)\right|$$

The visibility threshold T is given by:

$$T(x,y,t)=\begin{pmatrix}M^l(x,y,t)+M^s(x,y,t)-\\ C^{ls}.\min\{M^l(x,y,t),M^s(x,y,t)\}\end{pmatrix}M^t(x,y,t)$$

The visibility threshold $T(x,y,t)$ provides an indication of the maximum allowable distortions at a particular pixel in the image frame which will still not be visible to human eyes. Here, $M^l(x,y,t)$, $M^s(x,y,t)$ and $M^t(x,y,t)$ can be regarded as effects due to colour masking, spatial-textural masking, and temporal masking respectively at a particular pixel located at position (x,y) in the image frame at time interval t in the video sequence, while C^{ls} is a constant term which accounts for the overlapping effect in masking. Masking is a very important visual phenomenon which explains why similar artefacts are disturbing in certain regions of an image frame while they are hardly noticeable in other regions.

The temporal masking M^t attempts to emulate the effect of human vision's characteristic of being able to accept higher video-frame distortion due to larger temporal changes:

$$M^t(x,y,t)=e^{f_s\cdot f_r}T^{md}\left(d_f(x,y,t)\right)$$

where $d_f(x,y,t)$ is the inter-frame difference at a particular pixel location (x,y) in time t between a current frame $I_o(x,y,t)$ and a previous coded frame $I_o(x,y,t-f_f/f_r)$ and is mathematically expressed as:

$$d_f(x,y,t)=I_o(x,y,t)-I_o(x,y,t-f_f/f_r)$$

Here, f_r is the frame rate at which the video has been compressed, f_f is the full frame rate of the original sequence, f_s is a scaling factor, and T^{md} is the profile for all possible inter-frame colour difference used [11].

The colour masking M^l [11] attempts to emulate the effect of human vision's characteristic of being able to accept higher video-frame distortion when the background colour is above or below a certain mid-level threshold.

The spatial-textural masking M^s attempts to emulate the effect of human vision's characteristic of being able to accept higher video-frame distortion when the particular point has richer texture or spatial profile:

$$M^s(x,y,t)=\begin{pmatrix}m(x,y,t)b(x,y,t)\alpha_1+m(x,y,t)\alpha_2\\ +b(x,y,t)\alpha_3+\alpha_4\end{pmatrix}W(x,y,t)$$

Here, α_1, α_2, α_3, and α_4 are constants used to determine the exact profile of the spatial-textural masking.

In the spatial-textural masking, $m(x,y,t)$ is the weighted average colour $g_k(x,y)$ in four different orientations (weighted by weight values w_1 and w_2) and it attempts to capture the textural-masking characteristic of the small local region centred on pixel (x,y,t) and can be defined as:

$$m(x,y,t)=\frac{1}{4}\left|w_1g_1(x,y,t)+w_2g_2(x,y,t)+w_2g_3(x,y,t)+w_1g_4(x,y,t)\right|$$

The different weightings (w_1 and w_2) given to the horizontal and vertical directions and the diagonal directions are for taking into consideration the difference in sensitivity of the human visual system to different spatial orientations.

Also, $g_k(x,y,t)$ is the average colour around a pixel located at position (x,y) of a frame in the original reference video sequence at time interval t and is computed by convolving a 7x7 mask, G_k, with this particular frame in the original reference video sequence:

$$g_k(x,y,t)=\sum_{m=-3}^{3}\sum_{n=-3}^{3}f(x+m,y+n,t).G_k(m+4,n+4,t)$$

The four 7x7 masks, G_k, for $k=\{1,2,3,4\}$, are four differently oriented gradient masks used to capture the strength of the gradients around a pixel located at position (x,y,t).

Here, $b(x,y,t)$ is the average background colour around a pixel located at position (x,y) of a frame in the original reference video sequence at time interval t and is computed by convolving a 7x7 low-pass filter mask, B, with this particular frame in the original reference video sequence:

$$b(x,y,t)=\sum_{m=-3}^{3}\sum_{n=-3}^{3}f(x+m,y+n,t).B(m+4,n+4,t)$$

In addition, $W(x,y,t)$ is an edge-adaptive weight of the pixel at location (x,y) of a frame in the original reference video sequence at time interval t, and it attempts to reduce the spatial-textural masking at edge locations because artefacts that are found on essential edge locations tend to reduce the visual quality of the image frame. Previous research findings have reported that edge information is found to be of primary importance in visual perception [12, 8]. Edge is directly related to the image content that demarcates object boundaries, surface crease, and other important visual events. Distortion at an edge is easier to be noticed than that in other textured regions because edge structure attracts more visual attention from human visual system [8]. Thus, for more accurate visibility threshold estimation, spatial-textural masking in edge and non-edge regions has to be distinguished, as adopted here.

2.3. Block-Fidelity

The block-fidelity feature measures the amount of distortion at block-boundaries in the test video when compared to the original reference (undistorted) video. The blocking effect is one of the significant coding artefacts that often occur in video compression. The block-fidelity measure for each individual frame of the video is defined as:

$$F_{BF}(t) = e^{(f_{BF}) \Big/ \left\{ 1 + \left| \left(B_d^h(t) + B_d^v(t)\right) - \left(B_o^h(t) + B_o^v(t)\right) \right| \Big/ \left(B_o^h(t) + B_o^v(t)\right) \right\}}$$

where the subscript o refers to the original video sequence, d refers to the test video sequence, f_{BF} is a constant, and:

$$B^h(t) = \frac{1}{H(\lfloor W/4 \rfloor - 1)} \sum_{y=1}^{H} \sum_{x=1}^{\lfloor W/4 \rfloor - 1} \left| d^h(4x, y, t) \right|;$$

$$d^h(x,y,t) = I(x+1,y,t) - I(x,y,t)$$

$I(x,y,t)$ denotes the colour value of the input image frame I at pixel location (x,y) and time interval t, H is the height of the image, W is the width of the image, $x \in [1,W]$, $y \in [1,H]$, $t \in [1,N]$, and N is the total number of frames in the video sequence. $B^v(t)$ and $d^v(x,y,t)$ can be computed in a similar way except in the y-direction.

2.4. Content Richness Fidelity

The content richness fidelity feature measures the fidelity of the richness of test video's content when compared to the reference video. This feature closely correlates with human perceptual response which tends to assign better subjective ratings to more lively and colourful images.

The image content richness fidelity feature for each individual frame of time interval t of the video can be defined as:

$$F_{RF}(t) = e^{(f_{RF}) R_d(t) / R_o(t)};$$

$$R(t) = -\sum_{p(i) \neq 0} p(i) \log_e \left(p(i)\right); \qquad p(i) = N(i) \Big/ \sum_{\forall i} N(i)$$

Here, i is the colour value, $N(i)$ is the number of occurrence of i in the image frame, $p(i)$ is the probability or relative frequency of i, and f_{RF} is a constant.

3. PERFORMANCE OF PROPOSED METRIC

3.1. Test Conditions and Subjective Test Method

Ninety test video sequences are generated by subjecting 12 different original undistorted CIF and QCIF video sequences ("Container", "Coast Guard", "Japan League", "Foreman", "News", and "Tempete") to H.264 video compression with different bit rates (from 24 kbps to 384 kbps) and frame rates (from 7.5Hz to 30Hz). The bit rates under test are much lower than those used in [14] after the image size factor has been offset. Each of the video sequence consists of 250 frames.

The subjective video quality tests of the test video sequences have been carried out as the tests conducted in [1], using Double-Stimulus Impairment Scale variant II (DSIS-II) subjective test method and performed by 20 subjects. The decoded sequences with frame rate lower than 30 fps are displayed with repeated frames on the 30 Hz display device.

3.2. Performance

Performance is measured by comparing the metric output Q with the subjective rating of subjective tests between the original and distorted sequences. Two performance measures have been used for comparison here (as in [14]): (1) Pearson correlation coefficient (r_p), and (2) Spearman rank-order correlation coefficient (r_s). In the case of ideal match between a metric's outputs and subjective ratings, r_p = 1 and r_s = 1.

Pearson correlation, which measures the prediction accuracy, is defined as:

$$r_p = \sum_k (q_k - \overline{q})(MOS_k - \overline{MOS}) \Big/ \sqrt{\sum_k (q_k - \overline{q})^2} \sqrt{\sum_k (MOS_k - \overline{MOS})^2}$$

where $\overline{q}$ and $\overline{MOS}$ are means of q and MOS, and k is the index for the video under test. Spearman rank-order correlation, which measures the prediction monotonicity, is defined as:

$$r_s = \sum_k (\chi_k - \overline{\chi})(\gamma_k - \overline{\gamma}) \Big/ \sqrt{\sum_k (\chi_k - \overline{\chi})^2} \sqrt{\sum_k (\gamma_k - \overline{\gamma})^2}$$

where χ_k is the rank of q_k and γ_k is the rank of MOS_k in the ordered data series, and $\overline{\chi}$ and $\overline{\gamma}$ are the respective midranks.

Table 1 shows the results of the proposed metric with respect to PSNR. The upper bound (UB) and lower bound (LB) of Pearson correlation were obtained with a confidence interval of 95%. Figure 1 shows the scatterplot of subjective ratings versus the PSNR values, Figure 2 shows the scatterplot of subjective ratings versus the video quality ratings estimated using our proposed metric, while Figure 3 shows the scatterplot of subjective ratings versus the VSSIM output. In these two figures, the middle solid line portrays the logistic fit using the 4-parameter cubic polynomial [14], while the upper dotted curve and the lower dotted curve portray the upper bound and lower bound respectively obtained with a confidence interval of 95%.

From Table 1 and Figure 1, 2, and 3, it can be seen that the proposed video quality metric performs much better than the PSNR. Our proposed method also out-performs the VSSIM method. However, this is not surprising as it should be noted that VSSIM method is not designed for low bit rate videos.

	r_p	r_p UB	r_p LB	r_s
PSNR	0.701	0.793	0.578	0.676
Proposed metric	0.916	0.944	0.875	0.920
VSSIM	0.593	0.713	0.440	0.599

Table 1: Performance of proposed video quality metric, VSSIM method, and PSNR

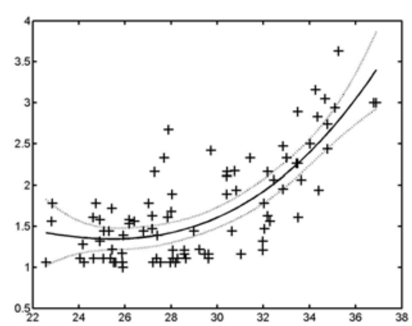

Figure 1: Scatterplot of subjective ratings vs PSNRs

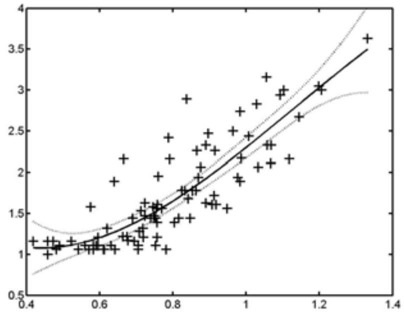

Figure 2: Scatterplot of subjective ratings vs proposed metric's output

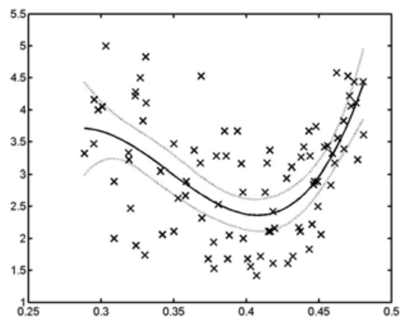

Figure 3: Scatterplot of subjective ratings vs VSSIM's outputs

5. CONCLUSION

This paper describes an objective perceptual video quality metric to automatically assess the perceived quality of a stream of video images compressed using H.264 video compression. The proposed method attempts to emulate human visual perception by introducing computational models based on block-fidelity, content richness fidelity, spatial-textural, colour, and temporal maskings. This method has been tested on H.264 digitally coded CIF and QCIF video sequences (at 24~384 Kbps and 7.5~30Hz) and shown to achieve significantly better correlation with subjective viewing results when compared to the PSNR and the video structural similarity (VSSIM) measure (being one of the best perceptual video quality metric (for high bit rate videos) that has been recently reported in the literature). Such an objective video quality metric will be extremely useful as it can replace the use of performance measure such as the traditionally used PSNR which has been found to correlate poorly with subjective quality ratings and also subjective tests which is not only time-consuming but also tedious and expensive to perform.

6. REFERENCES

[1] Baroncini, V., ISO/IEC JTC 1/SC29/WG 11, 4240, Sidney, July 2001.

[2] Chou, C.H., Li, Y.C., "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile", *IEEE Trans. Circuits & Syst. for Video Tech.*, Vol. 5(6), 1995, pp. 467-476.

[3] Girod, B., "What's wrong with mean-squared error", *Digital Images and Human Vision*, MIT Press, 1993, pp. 207-220.

[4] Jayant, N., Johnston, J., and Safranek, R., "Signal compression based on models of human perception", *Proceedings of IEEE*, Vol. 81, 1993, pp. 1385-1422.

[5] Karunasekera, S.A., Kingsbury, N.G. "A distortion measure for blocking artifacts in image based on human visual sensitivity", *IEEE Trans. on Image Processing*, Vol. 4(6), 1995, pp.713-724.

[6] Li, Z., "A saliency map in primary visual cortex", *Trends in Cognitive Science*, Vol. 6(1), 2002.

[7] Lindh, P., and Lambrecht, C., "Efficient spatio-temporal decomposition for perceptual processing of video sequences", *International Conference on Image Processing*, 1996.

[8] Marr, D., Vision: A computational investigation into the human representation and processing of visual information, W.H. Freemanand Co., 1982.

[9] Miyahara, M., Kotani, K. and Algazi, V.R., "Objective picture quality scale (PQS) for image coding", *IEEE Transactions on Comm.*, Vol. 46(9), 1998, pp. 1215-125.

[10] Nothdurft, H.C., "Salience from feature contrast: Additivity across dimensions", *Vision Res.*, Vol. 40, 2000, pp. 1183-1201.

[11] Ong, E.P., Yang, X., Lin, W., Lu, Z., and Yao, S., "Perceptual quality metric for compressed videos", *International Conference on Acoustic, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005.

[12] Ran, X., and Farvardin, N., "A perceptually motivated three-component image model – Part 1: Description of the model", *IEEE Trans. on Image Processing*, Vol. 4(4), 1995, pp. 401-415.

[13] Tan, K.T., Ghanbari, M., "A multi-metric objective picture-quality measurement model for MPEG video", *IEEE Trans. on Circuits & Syst. for Video Tech.*, Vol. 10(7), 2000, pp. 1208-1213.

[14] VQEG (Video Quality Expert Group), "Final Report from the Video Quality Expert Group on the validation of Objective Models of Video Quality Assessment", *www.vqeg.org*, 2000.

[15] VQEG, "Multimedia Group Test Plan – Draft Version 1.5e", Stockholm, Sweden, 30 Sep. 2005.

[16] Wang, Z., Lu, L., and Bovik, A.C., "Video quality assessment based on structural distortion measurement", *Signal Processing: Image Comm.*, Vol. 19(2), pp. 121-132, Feb. 2004.

[17] Watson, A.B., Hu, J., and McGowan III, J.F., "DVQ: A digital video quality metric based on human vision", *Journal of Electronic Imaging*, Vol. 10(1), 2001, pp. 20-29.

[18] Winkler, S., "A perceptual distortion metric for digital color video", *SPIE Proc. Human Vision and Elect. Imaging IV,* Vol. 3644, pp. 175 – 184, 1999.

[19] Wu, H.R. and Yuen, M., "A generalize block-edge impairment metric for video coding", *IEEE Signal Processing Letters*, Vol. 4(11), 1997, pp.317-320.

[20] Yu, Z, Wu, H.R., Winkler, S. and Chen, T., "Vision-model-based impairment metric to evaluate blocking artifacts in digital video", *Proc. IEEE*, Vol. 90(1), pp. 154-169, 2002.