A NEW PERCEPTUAL QUALITY METRIC FOR COMPRESSED VIDEO

Abharana Bhat, Iain Richardson and Sampath Kannangara

The Robert Gordon University, Aberdeen, United Kingdom

ABSTRACT

This paper presents a new video quality metric for automatically estimating the perceptual quality of compressed video sequences. Distortion measures such as the mean squared error (MSE) and the peak signal to noise ratio (PSNR) have been found to poorly correlate with visual quality at lower bit-rates. The proposed quality metric (MOSp) predicts perceptual quality of compressed video using sequence characteristics and the mean squared error (MSE) between the original and compressed video sequences. The metric has been tested on various video sequences compressed using the H.264 video compression standard at different bit-rates. Results show that the proposed metric has better correlation with subjective quality compared to popular metrics such as PSNR, SSIM and PSNRplus. The new metric is simple to compute and hence suitable for incorporation into real-time applications such as the standard video compression codecs inorder to improve the visual quality of compressed video sequences.

Index Terms— quality metrics, video quality, mean squared error, perceptual quality, compressed video

1. INTRODUCTION

Measurement of video quality is important in video compression because there is always a trade-off between the amount of compression and video quality. Subjective evaluation is the most accurate way to determine the visual quality of video. However, subjective tests are expensive in terms of time and resources [1], and cannot be easily embedded into practical real-time video applications. Hence several objective video quality metrics have been proposed in the literature. The goal of objective video quality metrics is to predict perceived quality automatically and be in close agreement with subjective test results. Traditionally quality metrics such as mean squared error (MSE) and sum of squared difference (SSD) are employed by video compression systems to choose the best compression options and achieve optimal trade-off between picture quality and data rate [2]. These objective metrics are simple to compute and mathematically convenient to use. However, they are not representative of the distortions perceived by the human visual system (HVS). [3]

Several perceptual based distortion measures have been proposed and analysed as alternatives to the above mathematical measures. These measures focus on modelling the known psycho-visual properties of the human visual system (HVS). Typical HVS-based metrics include the perceptual distortion model [3], just noticible difference (JND) [4], digital video quality (DVQ) [5] and the Structual SIMilarity index (SSIM) [6]. Measures based on complex HVS models are computationally expensive and not practical for real-time video applications. Studies conducted by the Video Quality Experts Group indicate that the performance of HVS-based metrics needs to be improved further [7].

Although the overall correlation between objective quality measures such as MSE and subjective results such as mean opinion score (MOS) is poor, there is a higher correlation between them for one sequence coded at several bit rates with the same codec. This correlation decreases with increase in the number of different video sequences added to the test data set. Previously, authors of [8] have developed a method (PSNRplus) for increasing the correlation between subjective and predicted video quality by estimating the parameters of the linear regression line for each video sequence. The regression parameters were determined using 2 additional instances of the original video. Although this method produced improved results compared to previous methods in the literature, the method requires every sequence to be coded or compressed 3 times inorder to obtain the 2 additional instances hence making this technique unsuitable for real time applications. The accuracy of prediction is highly dependent on the choice of the 2 additional instances used to make the prediction thus reducing the robustness of this technique.

The aim of this research is to develop a perceptual quality metric that can automatically predict the visual quality of compressed video in real time, correlate well with mean opinion score (MOS) and be easily incorporated into standard video compression systems inorder to make coding decisions based on visual distortion (rather than poorly-correlated objective metrics). This paper is organised as follows: Section 2 of this paper describes the subjective evaluation process performed to obtain MOS values of various sequences used to develop the new perceptual metric. In section 3, the proposed perceptual metric is described. Performance of the new metric is evaluated in section 4. Section 5 contains conslusions and future work.

2. SUBJECTIVE VERSUS OBJECTIVE VIDEO QUALITY

As mentioned before, the correlation between objective measures such as MSE and subjective results such as MOS is high for one sequence coded at several bit rates. To investigate this further, we determine the variation of MOS with MSE across various video data using a training data set of seven different CIF video sequences. The sequences were 10 seconds in duration and coded using the H264/AVC compression standard. The sequences used were Carphone, Foreman, Mobile, News, Bus, Paris and Coastguard. These sequences were compressed at QP = {6, 26, 34, 36, 38, 40, 42, 45}. The subjective tests involved 10 naive evaluators and followed the guidelines in ITU-BT.500 [9]. The single stimulus impairment scale (SSIS) evaluation method was used. A grading scale of 0 to 1 was used to rate the quality of the test sequences where 0=bad, 0.25=poor, 0.5=fair, 0.75=good and 1=excellent.

Each evaluator took less than 20 minutes to complete the test. The 95% confidence intervals for the subjective ratings were around 0.0476 for the MOS scale of [0,1]. The mean opinion score (MOS) for a sequence was calculated as the average of all scores obtained for the sequence compressed at a certain QP. The mean squared error was calculated as the mean of the squared differences between the luminous values of pixels in original sequence (I) and the reconstructed compressed sequence (Ic) with picture size MxN and T frames per sequence as follows:

$$MSE = \frac{1}{M * N * T} \sum_{t=1}^{T} \sum_{y=1}^{N} \sum_{x=1}^{M} [I(x, y, t) - I_c(x, y, t)]^2 \qquad \textbf{(1)}$$

The graph of MSE versus MOS in Figure 1 shows the characteristic 'hockey stick' shaped curves for four test sequences: Carphone, News, Paris and Bus. The curves are approximately linear from MOS = 1.0 to MOS = 0.1 with a tail off below MOS = 0.1. These curves are ¨hockey stick¨ shaped because at very low bit-rates (below MOS=0.1), the picture quality is very poor and the users rate the video as ¨Bad¨ quality after a certain error threshold with little discrimination in picture quality. Hence we introduce a cut-off at MOS=0.1 and use data points above this cut-off to build our model.

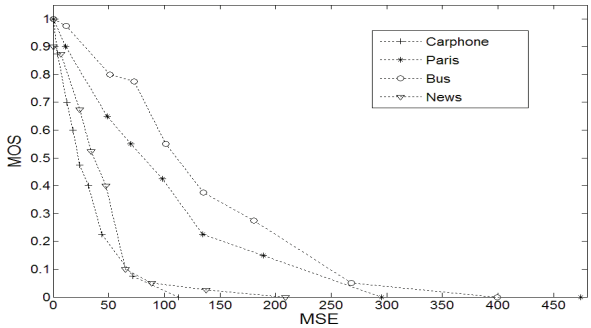


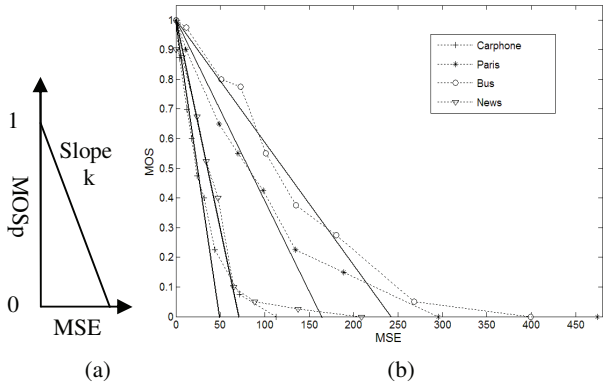
Figure 1. Graph of MSE versus MOS.

3. THE PROPOSED PERCEPTUAL METRIC

Design of the proposed perceptual metric is motivated by: (a) achieving good correlation with MOS (b) maintaining computational simplicity. Based on the observation made from the MOS versus MSE graph in Figure 1 we propose the perceptual metric as:

$$MOS_p = 1 - k(MSE) \qquad (2)$$

for predicting the mean opinion score (MOSp) of a compressed sequence using the mean squared error (MSE) between the original and compressed video sequences, and the slope of the regression line (k) which is calculated automatically from the sequence content. Figure 2(a) illustrates the proposed model which represents the linear relationship between MOS and MSE where the maximum perceived quality (with the value = 1) is observed when there are no pixel errors (MSE = 0). Figure 2(b) shows the proposed model (bold lines) fit to four test sequences: Carphone, Paris, Bus and News.



(a) (b)

Figure 2(a) Proposed Model. (b) Proposed (bold lines) curves for Carphone, Paris, Bus and News Sequences.

3.1. Estimating slope (k) using sequence characteristics

We propose to derive the slope (k) of the regression line in the model from the sequence content. Previous research findings have reported that visibility of artefacts in highly detailed regions is lower than in low detailed regions because a highly detailed region can mask artefacts more effectively [10]. Spatial edges give a good estimate of the amount of detail in a region and are related to object boundaries, surface crease, and other important visual events [11]. Considering this, we use the spatial edge strength as a measure of detail and derive the slope k. In order to obtain the edge information, edge detection filters are applied to the luminance component of the original frame (or field). In this paper, the Sobel edge detecting filters [11] are used due to their simplicity and efficiency. The horizontal edge image and the vertical edge image are separately computed using the Sobel filters, and the edge magnitude image is computed as follows:

$$G(x,y) = |G_{horizontal}(x,y)| + |G_{vertical}(x,y)| \quad (3)$$

where, G is the edge magnitude image and (x,y) is the pixel location. Spatial edge strength is measured using local regions. Hence the edge magnitude image is divided into 16x16 non-overlapping blocks or macroblocks[1], and the edge strength of each macroblock is computed as the average of the edge magnitudes of that macroblock. The edge strength of a frame is calculated as the average of edge strengths of all the macroblocks in the frame and the sequence edge strength is the average value of edge strengths of all the frames. A sequence can be highly detailed or low detailed or be a combination of both. Examples of highly detailed sequences include scenes such as a high speed bus moving through traffic (Bus) and an interview scene with a highly textured background (Paris). On the other hand, scenes such as news reader shot (News, Akiyo) and facial close-up interview (Foreman–no panning) are perceived as low detailed sequences. Video content in general spans from being highly detailed to low detailed based on average edge strength.

The relation between slope and the sequence edge strength is acquired using the seven training sequences mentioned in section 2. We derive the relation between slope and sequence content using the exponential fit as:

$$k = 0.03585 * \exp(-0.02439 * SequenceEdgeStrength) \quad (4)$$

This curve fit is plotted as the dotted line in Figure 3. It is clear from the graph that (4) is a good prediction of slope k. From Figure 3 it can be observed that low-detailed sequences such as, the Carphone sequence with average edge strength of 30.19, produce steeper regression lines in the MSE versus MOS graph. High-detailed sequences such as, the Mobile sequence with an average edge strength of 107.51, have shallower regression lines. This indicates that in low-detailed sequences, a small change in MSE leads to a larger change in MOS when compared to high-detailed sequences for the same amount of change in MSE.

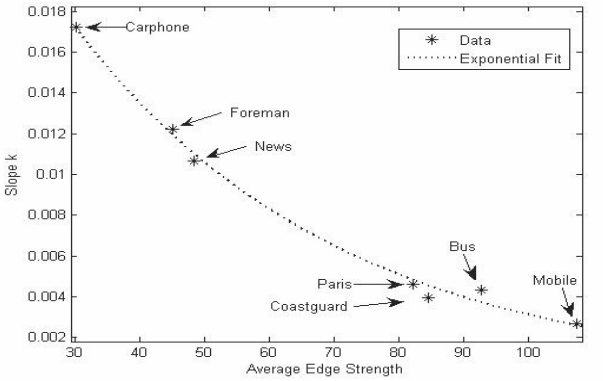


Figure 3. Graph showing relation between slope and the average edge strength of various sequences.

[1] We choose to determine edge strength and MSE at macroblock level in order to facilitate incorporating the metric into block-based video codec mode selection algorithm.

3.2. Sequence level quality evaluation using the proposed metric

During subjective evaluation of video quality by human observers, a judgement is made based on the overall quality of the sequence under test. Video sequences compressed at low bit-rates could have good picture quality in some parts of the sequence while other parts could have poor picture quality. The sequence quality rating in this case will be the average quality. Hence, we propose to evaluate quality at macroblock level first and then combine them into frame-level quality and finally produce a single valued sequence-level quality measure. The proposed metric first computes the edge magnitude image of each original luminance frame and then predicts the perceptual quality (MOSp) at macroblock level. The edge strength of every macroblock is calculated in order to determine the slope k. The MSE between macroblocks of the original and compressed luminance frames are computed. MOSp for every macroblock is computed as:

$$MOSp_{macroblock} = 1 - k_{macroblock}(MSE_{macroblock}) \quad (5)$$

The combined average of MOSp of all the macroblock in a frame gives the frame-level quality measure. The overall quality of the video sequence is given by the average of MOSp of all the frames in the sequence.

4. RESULTS

Performance of a perceptual quality metric depends on how well it correlates with subjective test results. Following the performance evaluation methods adopted by the video quality experts group (VQEG) [7], we use two evaluation metrics to give quantitative measures of the performance of the proposed metric. The first metric is the Pearson's correlation coefficient which measures the prediction accuracy of the new metric with respect to subjective results. For a set of N data pairs (x_i, y_i), the Pearson's correlation (rp) is defined using means $\bar{x}$ and $\bar{y}$ as follows:

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})}} \quad (6)$$

The second metric is the outliers ratio (OR) which is a measure of prediction consistency. A data point is considered to be an outlier if the difference between the predicted value and the actual subjective value exceeds ± 2 times the standard deviation of the subjective results [7]. We compare the proposed metric (MOSp) with MOS and three popular objective metrics: peak signal to noise ratio (PSNR), structural similarity metric (SSIM) [6] and

PSNRplus [8]. Experimental results are illustrated in Table 1 and Table 2. Table 1 gives the Pearson's correlation between the popular metrics and MOS. It is clear from Table 1 that the proposed metric (MOSp) correlates well with MOS for a variety of video sequences ranging from low detailed such as Akiyo and News, to high detailed sequences such as Bus, Mobile and Coastguard. The metric also produces good results with sequences which are a combination of both low-detailed and high-detailed scenes such as Foreman and Tempete sequences.

Table 1. Pearson Correlation between popular metrics and MOS

Sequences	**PSNR**	**SSIM**	**PSNRplus**	**MOSp**
Training				
Foreman	0.713	0.769	0.925	0.991
Bus	0.747	0.849	0.872	0.985
Paris	0.765	0.798	0.906	0.942
Mobile	0.712	0.702	0.880	0.926
Carphone	0.684	0.853	0.891	0.959
News	0.721	0.771	0.905	0.916
Coastguard	0.657	0.724	0.881	0.991
Non-training				
Husky	0.748	0.765	0.878	0.919
Tempete	0.726	0.794	0.916	0.930
Deadline	0.708	0.834	0.875	0.929
SignIrene	0.787	0.747	0.852	0.896
Mother& Daughter	0.759	0.758	0.894	0.931
Salesman	0.769	0.804	0.881	0.875
Akiyo	0.786	0.811	0.871	0.905

Table 2 illustrates the Pearson's correlation and outliers ratio between MOS and the popular quality metrics when all the video sequences are included. The Pearson's coefficient of MOSp is 0.947, which is the highest amongst the metrics compared in Table 2. The closest to this performance is PSNRplus at 0.886 but it has serious performance limitations (see below). The outliers ratio of MOSp is 0.402, which is the lowest in all the metrics.

Table 2.Comparison of MOSp with popular metrics

Metric	Pearson's Correlation	Outliers Ratio	Elapsed time (seconds)
PSNR	0.709	0.830	2.27
SSIM	0.747	0.776	24.36
PSNRplus	0.886	0.589	12.89
MOSp	0.947	0.402	12.05

Table 2 compares the performance of quality metrics in terms of execution speed. The elapsed time was taken by running each quality metric on the ¨Paris¨ CIF sequence with 150 frames using the MATLAB implementation of the metrics running on a 1.5 GHz, 512 MB RAM desktop PC. Within the class of quality metrics that correlate well with subjective quality, the MOSp metric is twice as fast as the SSIM metric. MOSp is also faster than the PSNRplus in terms of execution speed since PSNRplus requires each sequence to be encoded 3 times in order to determine the regression parameters required for making a prediction [8]. The elapsed times presented in Table 2 exclude the encoding time. Hence it is clear that MOSp performs significantly better than SSIM, PSNR and PSNRplus.

5. CONCLUSION AND FUTURE WORK

A new perceptual quality metric (MOSp) that predicts MOS of compressed video automatically by using sequence characteristics and the mean square error (MSE) has been proposed. Experimental results show that the new MOSp metric correlates well with subjective scores and outperforms existing visual quality metrics such as PSNR, PSNRplus and SSIM. We have started investigating techniques of integrating the new perceptual metric into H264/AVC mode selection algorithm with the aim of achieving better picture quality by making mode decisions based on accurately estimated perceptual quality.

6. REFERENCES

[1] E. Silva and K. Panetta, ¨Quantifying image similarity using measure of enhancement by entropy¨, Proceedings: Mobile Multimedia/Image Processing for Military and Security Applications, SPIE Defence and Security Symposium 2007, Vol. 6579, Paper #6579-0U, Orlando, FL, April 2007.

[2] Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," IEEE Signal Processing Magazine, pp. 23-50, 1998.

[3] S. Winkler, ¨A perceptual distortion metric for digital color video,¨ in Proc. SPIE, vol.3644, May 1999, pp.175-184.

[4] J. Lubin and D. Fibush, ¨Sarnoff JND vision model,¨ T1A1.5 Working group Document, T1 Standards Committee, 1997.

[5] A.B. Watson, J. Hu and J.F. McGowan III, ¨Digital video quality metric based on human vision,¨ Journal of Electronic imaging, vol. 10, no.1, Jan 2001, pp. 20-29.

[6] Z. Wang and A.C. Bovik, ¨A universal image quality index,¨ IEEE Signal Processing Letters, vol. 9, no. 3, Mar. 2002, pp. 81-84.

[7] Video Quality Experts Group, "Final Report from the VQEG on the validation of Objective Models of Video Quality Assessment, Pase II", *www.vpeg.org,* August 2003.

[8] T. Oelbaum, K Diepold and W. Zia, ¨A generic method to increase the prediction accuracy of visual quality metrics¨, PCS 2007.

[9] ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures, ITU-R Std., Rev. 11, June 2002.

[10] E.P Ong, W. Lin, Lu Zhongkang, S. Yao, M. H. Loke, "Perceptual Quality Metric for H.264 Low Bit Rate Videos," IEEE International Conference on Multimedia and Expo, vol., no., pp.677-680, July 2006.

[11] X. Ran and N. Farvardin, ¨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.