Video Quality Assessment Using Radial Basis Function Neural Network

Yao Susu, Lin Weisi, Lu Zhongkang and Ong Eeping
Signal Processing Program
Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613

Abstract: In this paper, a new method for perceptual video quality evaluation is proposed. The method is based on multi-feature of visual perception and various coding artifacts. A radial basis function neural network is used to give discrimination. The proposed method has been tested using full set of 50 Hz VQEG test data, and the results show that a good correlation coefficient of more than 0.90 with the subjective mean opinion score (MOS) is achieved.

1. Introduction

With the development of video coding technologies and establishment of international coding standards, the evaluation of video quality becomes an important issue. This is because video coding at low bit rate inevitably results in visible artifacts due to coarse quantization. To achieve a high performance coding, i.e. good visual quality with low bit rate, it is necessary to investigate human visual perception to the coding artifacts or to build computational models to simulate the Human Visual System (HVS) for predicating image quality. Such a model can be applied to optimize the performance of digital imaging system with respect to the capture, display, storage, compression and transmission of visual information. Peak Signal-to-Noise Ratio (PSNR) and mean square error (MSE) are widely used as objective quality metrics. However, they are pixel based fidelity metrics, which do not always match well with the perceived picture quality. In the past decades, many objective quality metrics for measuring video impairment have been investigated [1-3]. Most of them used perceptual models to simulate the human visual system and weight the impairments according to their visibility. Unfortunately, the HVS is so complex that existed perceptual models could not match to the real HVS well, and thus could not provide accurate rating of video quality. Another approach tried to exploit the properties of known artifacts, such as blocking artifacts, using feature extraction and model parameterization [4-6]. This class of measure method focuses on the particular type of artifacts so it is normally more accurate than perceptual model based metrics. However, it does not possess universality. In this work, we propose a new method for the evaluation of subjective video quality, which uses multi-feature extraction of perceptual properties and coding artifacts. A radial basis function neural network (RBF-NN) is employed to give classification and discrimination. The objective performance of the proposed method has been tested using the 50Hz VQEG test sequences. Good results have been obtained in terms of Pearson line correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) between the objective quality rating and the subjective Mean Opinion Score (MOS).

2. The Multi-Feature Extraction and RBF-NN Based Video Quality Metrics (MFENN-VQM)

The proposed MFENN-VQM is illustrated in Fig. 1. The inputs to the metric are the original and distorted successive frames. Since human eyes are more sensitive to the luminance component than chromatic components we only use Y components in this work. The differences between two inputs are decomposed into multi-resolution and multi-orientation bands for extracting energy features. Other features are extracted from frequency masking, luminance masking, blockness measure and blurring measure. All the features are forwarded to a neural network for discrimination. The neural network is trained using subjective test data provided by the Video Quality Experts Group (VQEG). In the following sub-sections we describe the various components of this metric in details.

2.1 Multi-channel visual decomposition

It is well known that the visual system processes information in a manner of multi-channel tuned to different temporal frequencies, spatial frequencies and orientations. Digital filter banks can simulate this multi-channel system. In this work, we adopt the steerable pyramid transform introduced by Simoncelli *et al* [8]. The transform decomposes the image into several spatial frequency levels within which each level is further divided into a set of orientation bands. The transform has the property of

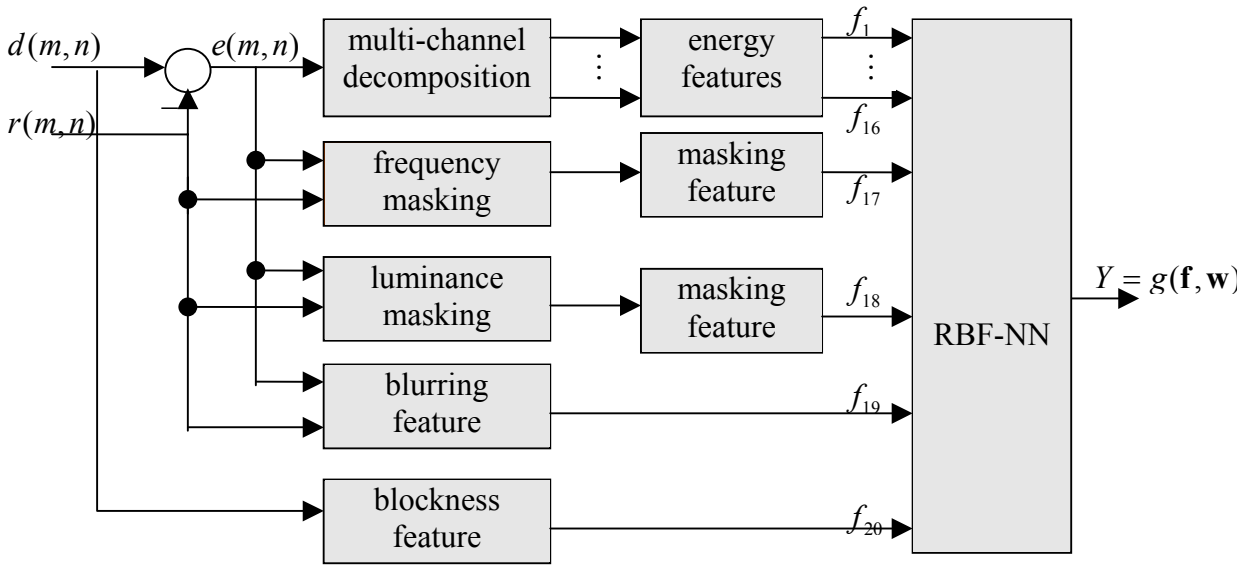


Fig. 1. Block diagrame of the proposed FENN-VQM metric

locally rotationally invariant and self-inverting. The block diagram for the decomposition (both analysis and synthesis) is show in in Fig. 2. Initially, the image is separated into low and high pass subbands, using filters L0 and H0. The lowpass subband is then divided into a set of oriented bandpass subbands and a lower pass subband. This lower-pass subband is then sampled by a factor of 2 in row and clown directions. The recursive construction of a pyramid is achieved by inserting a copy of the shaded portion of the diagram at the location of the solid circle.

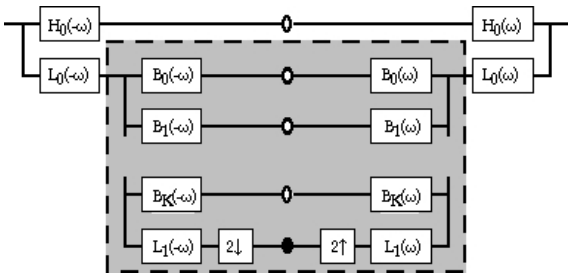


Fig. 2. Streeable pyramid transform

The multi-channel transform generates a set of coefficient values for the input frame. Refer to Fig. 1, the difference between two inputs is represented by $e(m,n) = r(m,n) - d(m,n)$, which is then decomposed into four levels with four orientations bands, where $r(m,n)$ and $d(m,n)$ are original frame and distorted frame, respectively. Total sixteen sub-bands are obtained. These coefficients are squared to form sixteen energy features of the orientation and spatial frequency components, which can be written as follows.

$$f_i = \sum_m \sum_n C_i(m,n)^2 \,, i = 1, \cdots, 16 \,.$$

Where, $C_i(m,n)$ represent the coefficients of steerable transform at each band. Further, f_i is averaged and normalized at each level.

2.1 Frequency masking

Masking is a very important visual phenomenon. Masking explains why similar coding artifacts are disturbing in certain regions such as flat regions of an image while they are hardly noticeable elsewhere such as edges and text regions. Obviously, masking measures, including frequency or activity masking and luminance masking are important features that have been considered in this metric. Due to the effect of masking, a surrounding spatial region of limited extent will affect the visibility of the coding artifacts, especially in the vicinity of edges. In order to incorporate the effects of masking effectively, the block activity of surrounding background should be calculated for every pixel, which would be computationally expensive. For saving computation

cost, we consider masking to be localized to 4 by 4 pixels block. The Surrounding Spatial Region Block (SSRB) with 8 by 8 pixels is used to compute the activity masking. The weighted activity is given as follows:

$$E_a = \frac{1}{64}\sum_{i=1}^{8}\sum_{j=1}^{8} Th(i,j)c(i,j)^2$$

where, $c(i,j)$ is the amplitude of DCT coefficients of the SSRB and $Th(i,j)$ is the visually threshold function of Watson' model [9]. As an example, figure 4 (a) and (b) show the first frames of original "Calendar and mobile" image sequence and distorted frame, respectively. Fig. 4 (c) illustrates the visually threshold function.

(a)

(b)

Fig. 3 (a) Original first frame of "Calendar and mobile" image sequence (b) Distorted frame (src10_hrc16) of fig. 3(a)

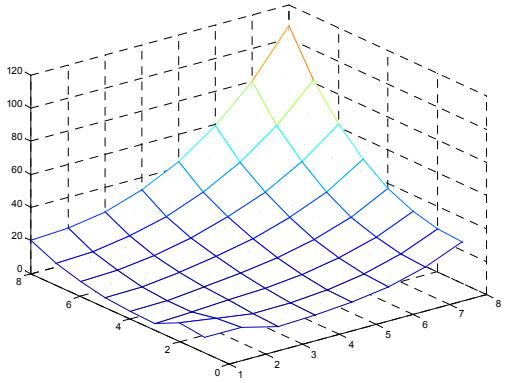


Fig. 3(c) Visually threshold function

Fig. 4. Frequency masking map

Fig. 4 gives the corresponding activity masking map and weighted error image, respectively. The frequency-masking feature is then given by the average of weighted error image.

$$f_{17} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} e(m,n)/E_a$$

2.3. Luminance Masking

Luminance masking occurs when the coding artifacts fall in brighter or darker region. Girod suggested that distortion are most noticeable where the luminance value is between 70 and 90 (centered) approximately on 81) in 8-bit gray-scale images. Luminance-masking weighting function $B_w(m,n)$ can be calculated by the following equations [7]:

$$B_w(i,j)=\begin{cases}\alpha\ln\left(1+\dfrac{\sqrt{\omega}}{1+\sigma}\right), & \text{if } \omega\le\beta\\ \ln\left(1+\dfrac{\sqrt{255-\omega}}{1+\sigma}\right), & \text{otherwise}\end{cases}$$

$$\alpha=\frac{\ln(1+\sqrt{255-\beta})}{\ln(1+\sqrt{\beta})},\ \beta=81$$

ω and σ are the mean and standard deviation of the SSRB. Figure 6 shows the luminance-masking map.

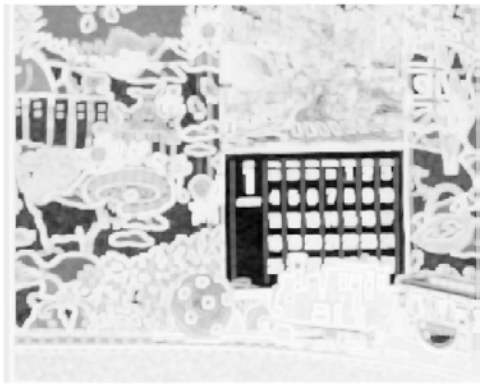


Fig. 5. Luminance-masking map

The brightness-masking feature is thus given by

$$f_{18}=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}B_w(m,n)e(m,n)$$

2.4 Blocking Distortion

The blocking effects, and its propagation through reconstructed video sequences, are the most significant of all coding artifacts. The blocking effect is also a source of a number of other types of reconstruction artifacts, such as stationary area granular noise. Several forms of quantitative quality metrics, or distortion measures, used in image and video coding research, have been developed in recent years, for instance, [5,6]. Given an image $\mathbf{f}=\{\mathbf{f}_{c1}\,\mathbf{f}_{c2}\cdots\mathbf{f}_{cN_c}\}$, where $\mathbf{f}_{cj}$ is the jth column of the image array and N_c is the width of the image, we can calculate the mean square sum of the pixel difference between each of the horizontal block (vertical edge artifacts) boundaries by the following equation, in which 8×8 pixel blocks are assumed as commonly used in video coding standards.

$$B_h=\frac{1}{N_c/8-1}\sum_{k=1}^{N_c/8-1}\left\|\mathbf{f}_{c(8k)}-\mathbf{f}_{c(8k+1)}\right\|^2$$

Similarly, we can obtain the mean square sum of the pixel difference between each of the vertical block (horizontal edge artifacts) boundaries B_v. Then, the blockness feature is given by

$$f_{19}=\sqrt{B_h+B_v}$$

2.5 Blurring Distortion

In order to evaluate burring image, we adopt spatial correlation coefficient as the feature to mark this kind of distortion. To quantify the degree to which a residual image $e(m,n)$ is correlated with an original image $r(m,n)$, we use the magnitude of the correlation coefficient between them

$$C_{er}=\frac{\left|Cov[e(m,n),r(m,n)]\right|}{\sigma_e\sigma_r}$$

Where, Cov refers to covariance, σ_e and σ_r are the standard deviation of images $e(m,n)$ and $r(m,n)$, respectively. By using an absolute value in the numerator, we ensure that $0\le C_{er}\le1$, with 0 indicating no correlation and 1 indicating linear correlation. The covariance is defined as $Cov[e,r]=E[(e-\mu_e)(r-\mu_r)]$, in which μ_e and μ_r denote the average values of $e(m,n)$ and $r(m,n)$, respectively. The correlation coefficient is used as the feature to measure blurring, i.e.

$$f_{20}=C_{er}$$

3. Metric Performance Evaluation

Performance of the objective models was evaluated with respect to three aspects of their ability to estimate subjective assessment of video quality [10]:

- Prediction accuracy – the ability to predict the subjective quality ratings with low error
- Prediction monotonicity – the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings and
- Prediction consistency – the degree to which the model maintains prediction accuracy over the range of video test sequences, i.e., that its response is robust with respect to a variety of video impairments.

4. Experimental Results

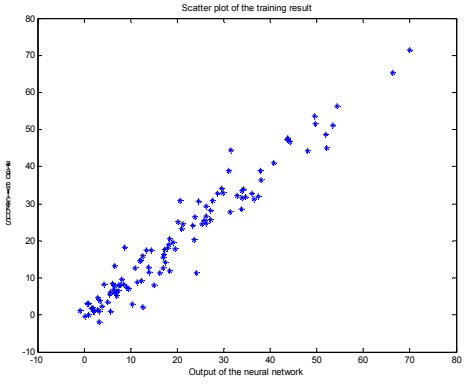
The experiments are conducted in two main steps: 1) training the proposed MFENN-VQM with subjective mean opinion scores (MOS) data of VQEG test sequences. 2) predicting the video quality using test sequences that are both inside and outside training sets. Pearson linear correlation coefficients and Spearman rank-order correlation coefficients between subjective and objective metric output are used to evaluate the performance of the MFENN-VQM.

4.1 Training and Test Sequences

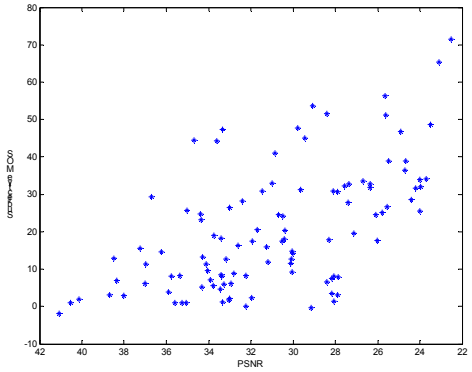
107 sequences are used from VQEG test data. These test sequences are 50Hz, with resolution of 576×720 pixels. 16 test conditions (Hypothetical Reference Circuits or HRCs) are covered in the test sequence, in which 9 HRCs represent a low bit rate range of 768 kb/s-4.5Mb/s (HRCs 8-16) and 9 HRCs represent a high bit rate range of 3Mb/s-50 Mb/s (HRCs 1-9).

4.2 Training of the MFENN-VQM

A radial basis function neural network with two-layer structure is adopted for nonlinear approximation. The inputs to the network are 20 feature values. The difference between the subjective data and output of the network is used to tune the weighting function of the connection between layers. Training set is composed of those sequences that are selected from the whole test sequences in a manner of random distribution. Remaining sequences are used for testing. Figure 8 (a) illustrates scatter plot showing the training results. For comparison, the scatter plot of PSNR versus MOS is given in fig. 8 (b). The Pearson correlation between the output of the neural network and the subjective DMOS is 0.9713.



(a)



(b)

Fig. 6. (a) Scatter plot of the output of neural network versus subjective DMOS; (b) scatter plot of PSNR versus subjective DMOS

4.3 Test of the MFENN-VQM

Test of the proposed metric is conducted in three cases. Whole 107 sequences are divided into two parts, one is training set, in which N sequences are randomly selected, and another is test set. First case is to set N=87, remaining 20 sequences are used for test. Second case is to set N=97, then 10 sequences are used for test. In third case, we circularly select 106 sequences for training, and let remaining one sequence be used for test. Table 1 lists the experimental results in terms of Person Linear Correlation Coefficients (PLCC) and Spearman Rank-order Correlation Coefficients (SRCC), in which MFENN-VQM-20 is case one, MFENN-VQM-10 is case 2 and MFENN-VQM-1 indicates case 3. The PSNR, PDM [Stefan's model] (from VQEG final report) are listed in table 2 for comparison. From the experimental results we see that the proposed MFENN-VQM model can give more accuracy prediction for evaluating video quality, compared with PSNR and PDM metrics.

Table 1. MFENN-VQM Correlations

Metrics	PLCC	SRCC
MFENN-VQM-20	0.943	0.901
MFENN-VQM-10	0.977	0.964
MFENN-VQM-1	0.903	0.867

Table 2 Correlation results from VQEG report

Metrics	PLCC	SRCC
PSNR	0.786	0.810
PDM	0.700	0.718

5. Conclusions

In this work we have presented a new video quality metric that is based on multi-feature extraction and radial neural network. The extracted features can reflect some important visual properties such as frequency masking and luminance masking as well as the measurement of coding artifacts. It has been demonstrated that the proposed video quality metrics can achieve high correlations with subjective rating. The further work should focus on the selection of more accuracy features that are related to the human visual system and more other features including color, texture and motion.

References

[1] A.B. Watson, J. Hu, J.F. McGowan III, DVQ: A digital video quality metric based on human vision, Journal of Electronic Imaging, Vol. 10, No. 1, 2001, pp. 20-29.

[2] A.B. Watson, and J.A. Soloman, A model of visual contrast gain control and pattern masking, Journal of the Optical Society of America A, Vol. 14, No. 9, pp. 2379-2391, 1997.

[3] Stefan Winkler, Issues in vision modeling for perceptual video quality assessment, Signal processing Vol. 78, pp. 231-252, 1999.

[4] M. Miyahara, K. Kotani, and V.R. Algazi, Objective picture quality scale (PQS) for image coding, IEEE Trans. Communications, Vol. 46, No.9, pp.1215-1225, 1998.

[5] Shanika A. Karunasekera an Nick G. Kingsbury, A distortion measure for blocking artifacts in image based on human visual sensitivity, IEEE Trans on Image Processing, Vol. 4, No. 6, pp.713-724, 1995.

[6] H.R. Wu, and M. Yuen, A generalize block-edge impairment metric for video coding, IEEE Signal Processing Letters, Vol. 4, No.11, pp.317-320, 1997.

[7] B. Girod, The information theoretical significance of spatial and temporal masking in video signals, Proc. SPIE Conf, Human Vision, Visual Processing, and Digital display, 1989, Vol. 1077, pp.178-187.

[8] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D. J. heeger, Shiftable multi-scle transforms, IEEE Transaction on Information Theory, Special Issue on Wavelet, 38:587-607, 1992.

[9] A.B. Watson, Perceptual optimization of DCT color quantization matrices, Proceedings of ICIP, Vol.1, pp.100-104, 1994,

[10] 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, March 2000.