A RNN-based objective video quality measurement

Xuan Huang (黄 轩)*, Rong Zhang (张 荣), and Jianxin Pang (庞建新)

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China

**E-mail: xuan@mail.ustc.edu.cn*

Received January 6, 2009

Technology used to automatically assess video quality plays a significant role in video processing areas. Because of the complexity of video media, there are great limitations to assess video quality with only one factor. We propose a new method using artificial random neural networks (RNNs) with motion evaluation as an estimation of perceived visual distortion. The results are obtained through a nonlinear fitting procedure and well correlated with human perception. Compared with other methods, the proposed method performs more adaptable and accurate predictions.

OCIS codes: 120.3940, 110.3000, 330.5020, 330.5000, 100.2960, 100.4145.
doi: 10.3788/COL20090711.1004.

Image/video quality assessment is a fundamental problem in the image processing filed[1]. It is important for compression systems and subsequent media services. Video has not only spatial correlations but also temporal correlations, so traditional image quality assessment methods cannot be applied in video directly. A great deal of investigations have been carried out, making significant achievement. And a known framework for performance evaluation of new objective video quality assessment methods has been recommended by the Video Quality Experts Group (VQEG)[2]. Furthermore, VQEG provides ten models to assess video quality[3]. Among them, peak signal-to-noise ratio (PSNR) is one of the most popular models and is still used widely.

Many video quality assessment models have been developed recently. Pessoa *et al*. presented a methodology[4] which worked when the video was processing video by unidirectional transmission systems that used digital interfaces and, ideally, digital transport facilities. Some quality assessment metrics use much simpler transforms such as the discrete cosines transform (DCT)[5] and the separable wavelet transforms[6], which all achieve comparable results. Wang *et al*. introduced a structural similarity (SSIM) measurement[7]. SSIM includes three factors which are luminance, contrast, and structure comparison, and is considered as a significant expression in human visual system (HVS).

Because of the complexity of video sequences, most video quality assessment models have a limitation that they use only one or a few factors to formulate video quality. A recent research in HVS has shown that there are at least five spatial-temporal interactive filters working together for visual perception[8]. The perception is combined with various factors, and the former models are not adaptable in extensive conversation. Perception and vision applications have been widely used in vast areas[9,10]. In this letter, a new substantial version of the algorithm called random neutral network (RNN)[11] is employed for video quality assessment. The RNN model represents the biophysical neural network signal transmitting manner more clearly, and yields strong generalization capabilities. These signals travel as voltage spikes rather than fixed signal levels. This tool has a high level of connectivity[12], which could be called self-description. The model also has fast-learning feature due to its computational simplicity for weight updating process.

The ith RNN neuron, as Gelenbe formulated[11], is defined using the following parameters: Λ_i and λ_i mean the rates of exogenous excitatory and inhibitory signals arriving at the neuron from a source outside of the network, while $W_{ji}{}^+$ and $W_{ji}{}^-$ stand for arrival rates of excitatory and inhibitory signals from neuron j; $k_i(t)$ is the instantaneous potential of the neuron. The neuron's state in a time interval Δt varies that

$$K_i(t+\Delta t) = \begin{cases} K_i(t)-1, & \text{when fired} \\ K_i(t)+1, & \text{excitatory signal arrive} \end{cases}. \quad (1)$$

If $K_i(t)$ is strictly positive, the neuron is excited, and it randomly sends signals according to a Poisson process with rate r_r.

Figure 1 shows the representation of a neuron in the RNN using the model parameters that have been defined above. All the other neurons can be interpreted as the replicas of neuron i.

For steady probability analysis, let $p(k)$ denote the stationary probability distribution, and its definition is $\lim_{t\to\infty} P[k(t)=K]$. The existence of the limit has been proved by Genlenbe[11].

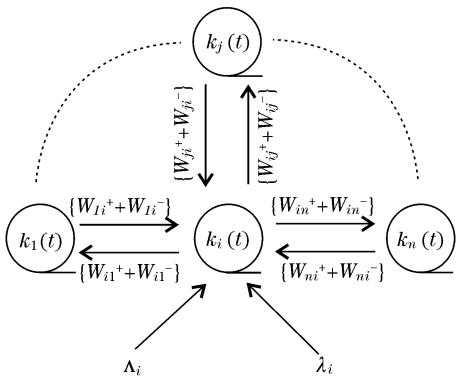

Fig. 1. Neuron model in RNN.

The learning algorithm trains the RNN with K input-output pairs (l, Y) and iterates the network parameters W. For an input vector $\mathbf{l} = \{l_1 \cdots, l_k\}$, where l_k is a pair of excitation and inhibition signal flow rates entering each neuron from outside of the network, denoted as $l_k = (\Lambda_k, \lambda_k)$. The partial derivative of the cost function can compute and substitute the updating difference equation in Gelenbe's formulation[11]:

$$W_t + \Delta t(u,\ v) = W_t(u,\ v) \\ -\eta \Sigma_{i=1}^{n} a_i (q_i - y_{ik})[\partial q_i / \partial W(u,\ v)], \quad (2)$$

where η is the learning parameter and is a constant,

$$q_i = \lambda^+(i)/[r(i) + \lambda^-(i)] \quad (3)$$

$$\lambda^+(i) = \Sigma_j q_j W_{ji}^+ + \Lambda(i), \quad (4)$$

$$\lambda^-(i) = \Sigma_j q_j W_{ji}^- + \lambda(i). \quad (5)$$

In Eq. (2), the core algorithm is to calculate the $[\partial q_i / \partial W(u,\ v)]$. Once the K learning values have been used, the whole process is repeated until some convergence conditions are satisfied.

Our model achievement is discribed detailedly as follows. And the whole framework is summarized as Fig. 2. The model uses the VQEG Phase I FR-TV test sequences[2]. These media are stored as YUV format, which contains a color space in terms of one luminance and two chrominance components. The subjective test result comes from double stimulus continuous quality scale (DSCQS) method[2]. In DSCQS, a difference score is defined as the difference between the rates for the reference sequence and the test sequence.

Referring to VQEG documents, the experiment presents 12 quality-affecting parameters that have the highest impact on the quality. The feature vector at frame i is defined as $\mathbf{X}_i = (x_1 \cdots x_{12})$. x_1 is the mean square error (MSE). x_2 is the PSNR. $x_3 - x_5$ denote the SSIM[7] luminance, contrast, and structure. x_6 and x_7 are SI_loss and SI_gain: spatial information (SI) is denoted as f_{SI13} since images are pre-processed using 13×13 filter masks. The feature is computed as standard deviation (std) over the spatial-temporal (S-T) region of $R(i,\ j,\ t)$ samples.

$$f_{\mathrm{s1}} = \{\mathrm{std}[R(i,j,t]\} : i,j,t \in \{\mathrm{S-T\ region}\}. \quad (6)$$

x_8 and x_9 are HV_loss and HV_gain: the image with horizontal and vertical gradients, denoted as HV, contains the $R(i,\ j,\ t)$ pixels that are horizontal or vertical edges (pixels that are diagonal edges are zeroed). Gradient magnitudes $R(i,\ j,\ t)$ are zeroed in both images to compute accurate θ:

$$\mathrm{HV}(i,j,t) = \\ \begin{cases} R\ (i,j,t), m\dfrac{\pi}{2} - \Delta\theta < \theta < m\dfrac{\pi}{2} + \Delta\theta \\ 0, \qquad \text{otherwise} \end{cases}. \quad (7)$$

x_{10} and x_{11} are CI_spread and CI_extreme: chrominance information (CI) is a single feature used to measure distortions in the chrominance signals (blue- and red-difference chroma components C_{B}, C_{R}). The components of a two-dimensional chrominance feature vector, CI, is computed as the mean over the S-T region of the C_{B} and C_{R} samples:

$$f_{\mathrm{coher\_color}} = (\mathrm{mean}[C_{\mathrm{B}}(i,j,t), 1.5 \times \mathrm{mean}(C_{\mathrm{R}}(i,j,t)]), \\ i,j,t \in \{\mathrm{S-T\ region}\}. \quad (8)$$

x_{12} represents CT: contati (CT) is a measurement of localized contrast information which is sensitive to quality degradations such as blurring (contrast loss) and additive noise (contrast gain). One localized contrast feature is computed from the luminance image as

$$f_{\mathrm{cont}} = \{\mathrm{std}[Y(i,j,t)]\} : i,j,t \in \{\mathrm{S-T\ region}\}. \quad (9)$$

These distorted features are used to estimate static spatial correlations, while video also has temporal correlations that determine the intensity of movement within the moving region. As mentioned by Wang *et al.*[7], in moving regions, HVS has low intensity, while it has high intensity in rarely changing scene. In this letter, an adjustment method is employed, and the influence algorithm of motion vector (MV) is explained as follows.

1) For the ith frame, calculate the video feature vector $\mathbf{X}_i$;

2) compute motion estimation of reference video by H.264 platform, set the MV to MV_i;

3) define the global feature vector as

$$F_{\mathrm{video}} = (\Sigma_{i=1}^{n} \mathbf{X}_i / \mathrm{MV}_i)/(\Sigma_{i=1}^{n} = \ 1/\mathrm{MV}_i). \quad (10)$$

The training process intends to find high dimensionality correlations between data attributes and parameters. A half of the distorted videos, containing 80 samples, are used to train the RNN weights, while other videos are used to assess the model performance. The processed RNN is trained using the learning algorithm given above to minimize the square of sum errors (SSE). When the iteration reaches the maximal limit or the SSE is less than the threshold, the training procedure stops. The implement uses a three-layer RNN which consists of 12 neurons in the input layer (corresponding to the 12 chosen parameters), 35 neurons in the hidden layer, and one output neuron. The SSE threshold is 0.0005. The learning rate is set to 0.1. The maximal number of iterations equals 700. The firing rate of the output neuron is 0.1 and the range of weight initialization is 0.2.

After the training procedure, the neural network reads input data and outputs the evaluation with formed pattern. The output is stored as result data. Then, the model's performance is evaluated with respect to the prediction accuracy mainly. Logistic functions are used in a fitting procedure to provide a nonlinear mapping between the objective and subjective scores. In VQEG plan, the regression of degradation mean opinion score (DMOS) against objective model scores may not adequately represent the relative degree of consistency of subjective scores[7]. Hence, a metric with variance weighted regression analysis is included in order to factor this variability into correlation rates. The logistic function applies a weighted least square procedure to minimize the error of the following function:

$$Y_i^w = W_i \left[\frac{\beta_1 - \beta_2}{1 + \mathrm{e}^{-(X-\beta_3)/\beta_4}} + \beta_2 \right] + \varepsilon_i,\ i = 1, \cdots, n, \quad (11)$$

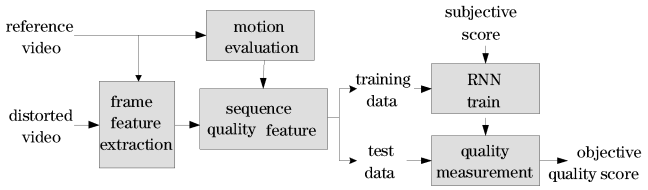

Fig. 2. Proposed video quality assessment framework.

Table 1. Performance Comparison of Twelve Models

Data	All	Low Quality	High Quality	625*	525*
SSIM	0.794	0.790	0.762	0.780	0.803
PSNR	0.804	0.813	0.782	0.826	0.752
P1	0.777	0.867	0.726	0.672	0.806
P2	0.792	0.836	0.695	0.759	0.837
P3	0.726	0.73	0.721	0.808	0.725
P4	0.622	0.584	0.656	0.665	0.657
P5	0.778	0.819	0.701	0.684	0.866
P6	0.277	0.36	0.33	0.347	0.373
P7	0.792	0.761	0.757	0.78	0.789
P8	0.845	0.827	0.666	0.864	0.739
P9	0.781	0.745	0.647	0.76	0.775
RNN	0.885	0.907	0.806	0.887	0.869

*625 and 525 are standard formats of YUV, media files. Both of them contain 720 pixels per horizontal line. The 525 sequences have 486 active lines per frame and the 625 sequences have 576 active lines per frame. The 525 sequences are 260 frames long and the 625 sequences are 220 frames long.

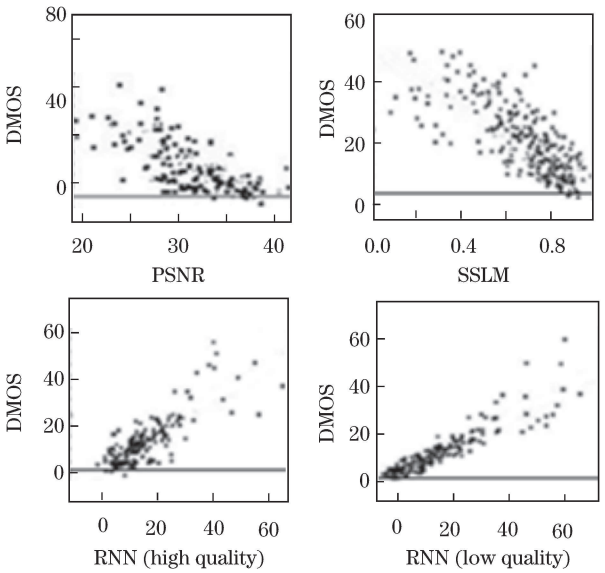

Fig. 3. Result comparison.

where initial estimations of parameters are Y_i is the ith DMOS, $\beta_1 = \max(Y_i)$, $\beta_2 = \min(Y_i)$, $\beta_3 = \text{mean}(x)$, $\beta_4 = 1, w_i = 1/\sigma(Y_i)$, $\sigma(Y_i)$ means the variance of the ith DMOS value, $Y_i^w = W_i \times Y_i$, $\varepsilon_i^w = w_t \times \varepsilon_i$, ε_i is the ith residual.

As previously described weight regression, X_i is the ith fitted value with objective scores. And the weighted correlation r_w can be computed via

$$r_{\mathrm{w}} = \frac{\Sigma_{i=1}^{n} w_i (X_i - \bar{X}_w)(Y_i - \bar{Y}_w)}{\sqrt{\Sigma_{i=1}^{n} \mathrm{w}_i (X_i - \bar{X}_w)^2 \Sigma_{i=1}^{n} w_i (Y_i - \bar{Y}_w)^2}}, \qquad (12)$$

where $\bar{X}_w = \Sigma_{i=1}^{n} X_i w_i / \Sigma_{i=1}^{n} w_i$, Y_i is the ith DMOS, $w_i = 1/\sigma_Y^2$, $\bar{Y}_w = \Sigma_{i=1}^{n} Y_i w_i / \Sigma_{i=1}^{n} w_i$, σ_Y^2 is the standard deviation of the ith DMDS value.

After the whole process on data set, the nonlinear regression correlations of performance comparison for all 12 models (included RNN, PSNR, SSIM, and the VQEG models P1 to P9[2]) are given below.

Table 1 and Fig. 3 shows the objective comparisons on all test video sequences of the VQEG Phase I proponent and the proposed method with weighted adjustment, respectively. The results with weighted adjustment indicate that the proposed RNN method is better than all the other models, about 10% higher than the average. The performances of P0, P1, P2, P3, P4, P5, P7, P8, P9, and SSIM are statistically equivalent.

For further works, we will investigate some other ways. For example, the RNN-based theory could cooperate with multiple-layer perception (MLP) based tools which concern the human visual perception under multi-scales.

In conclusion, we design a new objective video quality assessment system. The method's key feature is the use of RNN and multiple distorted factors instead of one factor for quality evaluation. Experiments on VQEG FR-TV Phase I test data set show that this method has good correlation with subjective video quality assessment and is more adaptable and accurate in human visual predictions.

References

1. J. Pang, R. Zhang, H. Zhang, X. Huang, and Z. Liu, Chin. Opt. Lett. **6,** 491 (2008).
2. Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment" (March 2000) available online at http://www.vqeg.org/ (September 24, 2009).
3. Video Quality Experts Group, "Methodology for the subjective assessment of the quality of television pictures", (September 1998) available online at http://www.vqeg.org/ (September 24, 2009).
4. A. C. F. Pessoa, A. X. Falcão, A. E. F. e Silva, R. M. Nishihara, and R. A. Lotufo, SMPTE Journal **108,** 865 (1999).
5. A. B. Watson, Proc. SPIE **1913,** 202 (1993).
6. Y.-K. Lai and C.-C. J. Kuo, J. Vis. Commun. Image Represent. **11,** 17 (2000).
7. Z. Wang, L. Lu, and A. C. Bovik, Signal Process.: Image Commun. **19,** 121 (2004).
8. Z. Yu and H. R. Wu, in *Proceedings of ICSP2000* 1088 (2000).
9. Y. Zhang, M. Li, J. Qiao, and G. Liu, Acta Opt. Sin. (in Chinese) **28,** 2104 (2008).
10. L. Pan, K. Tu, P. Liu, X. Zou, and X. Shao, Chinese J. Lasers (in Chinese) **35,** (s2) 355 (2008).
11. E. Gelenbe, Neural Comput. **5,** 154 (1993).
12. H. Bakircioğlu and T. Koçak, Eur. J. Operat. Res. **126,** 319 (2000).