International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:3, 2011

Video Quality assessment Measure with a Neural Network

H. El Khattabi, A. Tamtaoui and D. Aboutajdine

***Abstract*—**In this paper, we present the video quality measure estimation via a neural network. This latter predicts MOS (mean opinion score) by providing height parameters extracted from original and coded videos. The eight parameters that are used are: the average of DFT differences, the standard deviation of DFT differences, the average of DCT differences, the standard deviation of DCT differences, the variance of energy of color, the luminance Y, the chrominance U and the chrominance V. We chose Euclidean Distance to make comparison between the calculated and estimated output.

***Keywords*—**video, neural network MLP, subjective quality, DCT, DFT, Retropropagation

I. INTRODUCTION

Video Quality evaluation plays an important role in image and video processing. In order to change the human perception judgment by the machine evaluation, many researches were realized during the last two decades. Among the common methods, the mean squared error (MSE) [9], the peak signal to noise ratio (PSNR)[8]-[14], the discrete cosine transform (DCT)[5]-[6], and the decomposition in wavelets [13]. Another direction in this domain is based on the characteristics of the human vision system [10]-[11], like the contrast sensitivity function. Zhou Wang presented a different approach, Structural Similarity Index (SSIM), by using the structural distortion measurement. One should note that in order to check the precision of these measures, these latter should be correlated with the results obtained using subjective quality evaluations, there exist two major methods concerning the subjective quality measure: double stimulus continuous quality scale (DSCQS) and single stimulus continuous quality evaluation (SSCQE) are defined in ITU-R Rec. BT.500-11 [15]-[3] .

We present the video quality measure estimation via a neural network. This neural network predicts the observers mean opinion score (MOS) by providing height parameters extracted from original and coded videos. The eight parameters are: the average of DFT differences, the standard deviation of DFT differences, the average of DCT differences, the standard deviation of DCT differences, the variance of energy of color, the luminance Y, the chrominance U and the chrominance V.

H. El Khattabi, is with LRIT laboratory associated unite to CNRST, Faculty of sciences, BP 1014, Rabat, Morocco (email:hasnaa.alkhattabi@yahoo.fr)
A. Tamtaoui, is with National Institute Of Post offices and Telecommunications (INPT), Rabat, Morocco (email: tamtaoui@inpt.ac.ma)
D. Aboutajdine, is with LRIT laboratory associated unite to CNRST, Faculty of sciences, BP 1014, Rabat, Morocco (email: aboutaj@fsr.ac.ma)

The network used is composed of an input layer with eight neurons corresponding to the extracted parameters, three intermediate layers (with 7, 5 and 3 neurons respectively) and an output layer with one neuron (MOS). The function trainscg (training scaled conjugate gradient) was used in the training stage. We have chosen DSCQ for the video subjective measure since the extraction of the parameters is performed on the two videos, original and coded.

In the first section we describe the subjective quality measure, in the second section we present the parameters of our work and used the neural network, in the section 3 we give the results of our method and we end by a conclusion.

II. SUBJECTIVE QUALITY MEASUREMENT

A. Presentation

There exist two major methods concerning the subjective quality measure: double stimulus continuous quality scale (DSCQS) and single stimulus continuous quality evaluation (SSCQE).

We have chosen DSCQS [3]-[7]-[15] to measure the video subjective quality, since we deal with original and coded videos. We present to the observers the coded sequence A and the original B, without knowing which one is the reference video. For each sequence a quality score is then assigned, the processing continuation operates on the differences of the two scores using a subjective valuation scale (excellent, good, faire, poor, and bad) linked to a scale of values from 0 to 100.

The results are analyzed in as follows: Positions on the vertical scale are converted to normalized scores in the range 0 to 100. Each pair of scores is then converted to rating difference. The overall difference in quality is given as MOS (mean opinion score), which is computed as the mean value the differences from all observers related to one image pair. The higher the MOS, the more distortion in the image is visible.

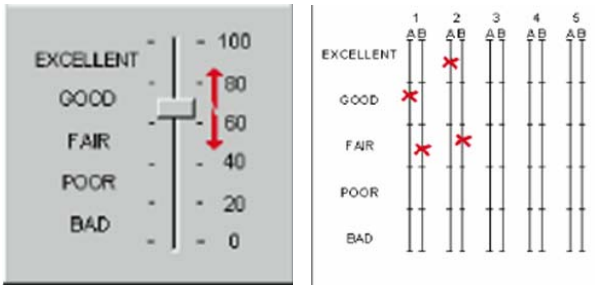


Fig. 1 Quality scale for DSCQS evaluation

B. Experimental

Examples of original sequences and their graduated shading versions that we used:

Akiyo original sequence,
Akiyo Coded / decoded with 24K bits/s,
Akiyo Coded / decoded with 64K bits/s,
Car phone original sequence,
Carphone Coded / decoded with 28K bits/s,
Carphone Coded / decoded with 64K bits/s,
Carphone Coded / decoded with 128K bits/s,

Fig. 2 originals sequences

Each sequence lasts 3 seconds, and each test includes two presentations A and B, coming always from the same source clip, but one of them is coded while the other is the non coded reference video. The observers should note down the two sequences without being aware of the reference video. Its position varies according to a pseudo random sequence. The observers see each presentation twice (A, B, A, B), according to the trial format of table 1.

TABLE I
THE LAYOUT OF DSCQS MEASURE

Subject	Duration(seconds)
Presentation A	8-10
Break for notation	5
Presentation B	8-10
Break for notation	5
Presentation A(second time)	8-10
Break for notation	5
Presentation B(second time)	8-10
Break for notation	5

The number of observers was 13 persons. In order to let them have a valid opinion during the trials, we asked them to watch the original and graduated shading video clips. We did not take into consideration the results of this trial. On the quality scale of figure 1, the observers were writing their notes with a horizontal line to represent their opinion about the quality of a given presentation. The seized value represents the difference in absolute value between the presentations A and B.

III. QUALITY EVALUATION

A. Parameters extraction

The extraction of parameters is performed on blocks for which the size is 8*8 pixels, and the average is computed on each block. The eight features extracted from the input/output video sequence pairs are:

Average of DFT difference (F1): This feature is computed as the average difference of the DFT coefficients between the original and coded image blocks.

Standard deviation of DFT difference (F2): The standard deviation of the difference of the DFT coefficients between the original and encoded blocks is the second feature.

Average of DCT difference (F3): This average is computed as the average difference of the DCT coefficients between the original and coded image blocks.

Standard deviation of DCT difference (F4): The standard deviation of the difference of the DCT coefficients between the original and encoded blocks.

The variance of energy of color (F5): The color difference, as measured by the energy in the difference between the original and coded blocks in the UVW color coordinate system, the UVW coordinates have good correlation with the subjective assessments [1]. The color difference is given by:

$$\Delta E = \sqrt{\Delta U^2 + \Delta V^2 + \Delta W^2} \quad (1)$$

The luminance Y (F6): in the color space YUV, the luminance is given by the Y component .The difference of the luminance between the original and encoded blocks is used as a feature.

The chrominance U (F7) and the chrominance V (F8): in the color space YUV, the chrominance U is given by the U component and the chrominance V is given by the V component. We compute the difference of the chrominance V between the original and encoded blocks and the same for the chrominance U.

The choice of parameters: the average of DFT differences, the standard deviation of DFT differences, and the variance of energy of color, is based on the fact they concern the subjective quality [1] and the choice of the luminance Y, and the chrominance U and V was made to get the information on the luminance and the color to predict the best possible subjective quality.

B. Multilayer neural networks

- *Presentation*

Neural networks have the ability to learn complex data structures and approximate any continuous mapping. They have the advantage of working fast (after a training phase) even with large amounts of data. The results presented in this paper are based on a multilayer feedforward network architecture, known as the multilayer perceptron (MLP). The MLP is a powerful tool that has been used extensively for classification, nonlinear regression, speech recognition, handwritten character recognition and many other applications. The elementary processing unit in a MLP is called a neuron or perceptron. It consists of a set of input synapses, through which the input signals are received, a summing unit and a nonlinear activation transfer function. Each neuron performs a nonlinear transformation of its input vector; the net input for unit j is given by:

$$net_j = \sum_i w_{ji} o_i + \theta_j \quad (2)$$

Where w_{ji} is the weight from unit i to unit j, o_i is the output of unit i, and θj is the bias for unit j.

MLP architecture consists of a layer of input units, followed by one or more layers of processing units, called hidden layers, and one output layer. Information propagates, in a feedforward manner, from the input to the output layer; the output signals represent the desired information. The input layer serves only as a relay of information and no information processing occurs at this layer. Before a network can operate to perform the desired task, it must be trained. The training process changes the training parameters of the network in such a way that the error between the network outputs and the target values (desired outputs) is minimized.

In this paper, we propose a method to predict the MOS of human observers using an MLP. Here the MLP is designed to predict the image fidelity using a set of key features extracted from the reference and coded video. The features are extracted from small blocks (say 8*8), and then they are fed as inputs to the network, which estimates the video quality of the corresponding block. The overall video quality is estimated by averaging the estimated quality measures of the individual blocks. Using features extracted from small regions has the advantage that the network becomes independent of video size. Eight features, extracted from the original and coded video, were used as inputs to the network.

- *Network Training Algorithm*

- The weights and the biases are initialized using small random values.
- The inputs and desired outputs are presented to the network.
- The actual outputs of the neural network are calculated by calculating the output of the nodes and going from the input to the output layer.
- The weights are adapted by backpropagating the error from the output to the input layer. That is,

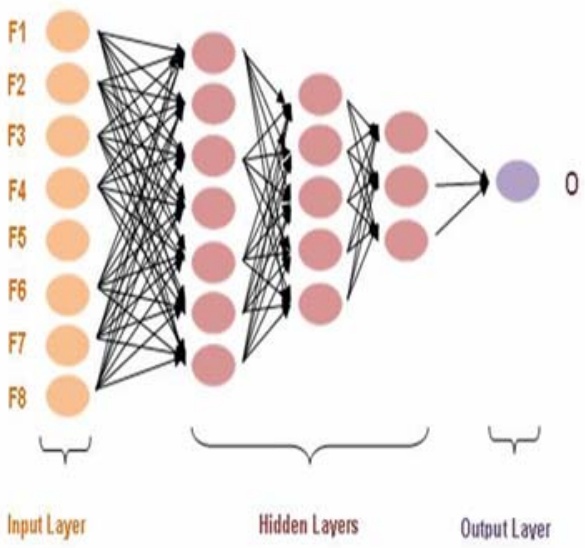
$$w_{ji}(n+1) = w_{ji}(n) + \varepsilon\delta \quad (3)$$

Where the δ is the error propagated from node j, and ε is the learning rate.

This process is done over all training patterns.

- *Architecture*

The multilayer perceptron used here is composed of an input layer with eight neurons corresponding to the eight parameters (F1, F2, F3, F4, F5, F6, F7, F8) ,an output layer with one neuron presenting the subjective quality(MOS),and three intermediate hidden layers. The following figure presents this network:



F1 : The average of DFT, **F2** : The standard deviation of DFT, **F3** : The average of DCT, **F4** : The standard deviation of DCT, **F5** : The variance of energy of color, **F6** : Luminance, **(F7,F8)** : Chrominance (U,V), **O** : MOS

Fig. 3 MLP network Architecture

VI. Experimental results

The aim of this work is to estimate the video quality from the eight extracted using MLP network. We have used sequences coded in H.263 of type QCIF (quarter common intermediate format), whose size is 176*144 pixels*30 frames, and sequences CIF (common intermediate format) whose size is 352*288 pixels*30 frames. We end up with

11880(22*18*30 blocks 8*8) values for each parameter per sequence QCIF and 47520(44*36*30 blocks 8*8) values for each parameter per sequence CIF. The optimization of block quality is equivalent to the optimization of frame and sequence quality [1]. The experiment part is achieved in two steps: Training and test.

In the MLP network training, five video sequences coded at different rates from four original video sequences (news, football, foreman and Stefan) were considered. The values of our parameters were normalized in order to reduce the computation complexity. This project was fully realized under Matlab (neural network toolbox).

The subjective quality of each of the coded sequences is assigned to the blocks of the same sequences. To make easier and accelerate the training, we used the function trainscg (training per scaled conjugate gradient). This algorithm is efficient for a large number of problems and it is much faster than other training algorithms. Furthermore its performances are not corrupted if the error is reduced and does not require enough memory to comply.

We use the neural network for an entirely different purpose; We want to apply it for the video quality prediction. Since no information on the network dimension is at our disposal, we will need to explore the set of all possibilities in order to refine our choice of the network configuration. This step will be achieved via a set of successive trials.

For the test we used 14 coded video sequences at different rates from 6 original video sequences (News, Akiyo, Foreman, Carphone, Football and Stefan). We point out here that the test sequences were not used in the training. The performance of the network is given by the Euclidean Distance, between the estimated output and the computed output of the sequence.

We worked with sequences QCIF and CIF separately in two edges and every edge is applied in two phases: training and test, the first edge consists in training only with sequences QCIF, There were three sequences for the training and six sequences for the test, and our network was able to predict four MOS among these six test sequences. And the second edge consists in making the training with three CIF sequences;

it means training basic increase since the frame of CIF is bigger than that of the QCIF and our network was able to predict seven MOS among eight of the test sequences. We can notice that our network with the training basic increases its performance increases and can indeed predict MOS best possible. The following tables present the computed and estimated (by the network) MOS.

We can observe that our neural network is able to predict the measurements of MOS, since the estimated values approach to the calculated values, and the values of Euclidean Distance are satisfactory.

We remark that the estimated values are not as exact as the ones that are computed; however they belong to the same quality intervals.

TABLE II
MOS COMPUTED AND ESTIMATED FOR QCIF SEQUENCES

Sequences	MOS computed	MOS estimated	Distance
training			
Akiyo24 kbits/s	0.4802	0.4613	0.0189
Akiyo64 kbits/s	0.2328	0.3207	0.0879
Forman41kbits/s	0.3509	0.3811	0.0302
Test			
Carphone28kbits/s	0.3790	0.3825	0.0035
Carphone64kbits/s	0.6690	0.5464	0.1226
Carphone128kbits/s	0.4739	0.4277	0.0462
Akiyo128 kbits/s	0.3711	0.6139	0.2428
Forman64kbits/s	O.6492	0.6373	0.0119
Forman128kbits/s	0.5508	0.6153	0.0645

TABLE III
MOS COMPUTED AND ESTIMATED FOR CIF SEQUENCES

Sequences	MOS computed	MOS estimated	Distance
training			
Football1.2Mbits/s	0.1257	0.1280	0.0024
Football 387kbits/s	0.3374	0.3351	0.0022
News387kbits/s	0.1862	0.1860	0.0002157
Test			
Forman1.2Mbits/s	0.0979	0.1257	0.0278
Forman388kbits/s	0.2056	0.1270	0.0786
Football358kbits/s	0.3177	0.3363	0.0186
News1.2Mbits/s	0.1194	0.1831	0.0638
News129kbits/s	0.2520	0.2826	0.0306
Stefan1.2Mbits/s	0.1754	0.3364	0.1610
Stefane388kbits/s	0.2795	0.3369	0.0574
Stefane280kbits/s	0.3520	0.3371	0.0149

IV. Conclusion

The idea of this work is based on the fact that we try to substitute the human eye judgment by an objective method that makes easier the computation of the subjective quality, without the need of people presence. That saves us an awful lot of time, and avoid us the hassle of bringing over people. Sometimes we need to calculate the PSNR without the use of the original video, that's why we are adding in this work the PSNR estimation.

We have tried to find a method that will allow us to compute the video subjective quality via a neural network by providing parameters (the average of DFT differences, the standard deviation of DFT differences, the average of DCT differences, the standard deviation of DCT differences, the variance of energy of color, the luminance Y, the chrominance U and the chrominance V) that are able to predict the video quality. The values of our parameters were normalized in order to reduce the computation complexity. This project was fully realized under Matlab (neural network toolbox). All our

sequences are coded in the H.263 coder. It was very hard to get a network able to compute the quality of a given video. Regarding the testing, our network approaches the computed value. Several tests have been conducted to find the architecture of a neural network that would give us better results. And similarly several experiments have been tried to search the adequate number of parameters. The same criteria have been used for both parameters and architecture, which is based on the error between the estimated value and the calculated value at the network output in the training step. Since we used the supervised training, we do impose to the network an input and output. We obtained bad results when we worked with a minimum of parameters (five and four parameters), as well as several parameters (eleven parameters).

We met some problems at the level of time, because the neural network takes a little more time at the level of the training step, and also at the level of database. Therefore our objective for the next work is to reduce the number of parameters by sequence and increase database.

REFERENCES

[1] F-H Lin, R. M. Mersereau: Rate-quality tradeoff MPEG video encoder. *Signal Processing Image Communication* 14 1999 297-309.

[2] Z. Wang, A. C. Bovik, (2006), *Modern Image Quality Assessment*, Morgan & Claypool Publishers, USA.

[3] M. Pinson, S. Wolf, (July 2003), *Comparing subjective video quality testing methodologies*. SPIE Video Communications and Image Processing Conference, Lugano, Switzerland.

[4] J. M. Zurada, (1992), Introduction to artificial neural system, PWS Publishiner Company.

[5] J. Malo, A. M. Pons, and J. M. Artigas,(1997), *Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain, Image and Vision Computing*, Vol. 15, pp. 535-548.

[6] A. B. Watson, J. Hu, and J. F. McGowan, (2001), *Digital video quality metric based on human vision*, Journal of Electronic Imaging, Vol. 10, No. I, pp. 20-29.

[7] H.M. Sun, Y.K. Huang, (2009), *Comparing Subjective Perceived Quality with Objective Video Quality by Content Characteristics and Bit Rates,* 2009 International Conference on New Trends in Information and Service Science, niss, pp.624-629.

[8] Q .Huynh-Thu,M. Ghanbari (2008) ,*Scope of validity of PSNR in image/video quality assessment,* Electronics Letters, vol. 44,No.13,pp.800–801.

[9] Z .Wang, A.C.Bovik (2009), Mean squared error: love it or leave it?, IEEE Signal Process Mag, vol.26, No.1,pp.98–117.

[10] H. R.Sheikh, A.C.Bovik, G.d. Veciana,(2005), *An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics*, IEEE TRANSACTIONS ON IMAGE PROCESSING,vol. 14, NO. 12,pp. 2117- 2128.

[11] D.Juan,Y.Yinglin,X.Shengli,(2005),*A New Image Quality Assessment Based On HVS*,Journal Of Electronics ,vol.22,No.3,pp.315-320.

[12] A.Bouzerdoum,A.Havstad,A.Beghdadi,(2004),*Image quality assessment using a neural network approach*,the Fourth IEEE International Symposium on Signal Processing and Information Technology,pp. 330-333.

[13] **A.**Beghdadi,**B.**Pesquet-Popescu,(2003),*A new image distortion measure based on wavelet decomposition*,Proc.Seventh Inter.symp.Signal. Proces. its Appricatiom , Vol. 1, pp. 485- 488.

[14] Slanina, M. Ricny, V.,(2008), *Estimating PSNR without reference for real H.264/AVC sequence intra frames* , Radioelektronika, 2008 18th International Conference,pp.1-4.

[15] **ITU-R BT.500-1**, **(2002**), *Methodology for the subjective assessment of the quality of television pictures*