A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

Muhammad Shahid, Andreas Rossholm, Benny Lövström

4th International Congress on Image and Signal Processing

2011 Shanghai, China

# A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

Muhammad Shahid, Andreas Rossholm and Benny Lövström
Department of Signal Processing, Blekinge Institute of Technology
SE-37179 Karlskrona, Sweden
Corresponding author email: muhammad.shahid@ieee.org

*Abstract*—**There is a growing need for robust methods for reference free perceptual quality measurements due to the increasing use of video in hand-held multimedia devices. These methods are supposed to consider pertinent artifacts introduced by the compression algorithm selected for source coding. This paper proposes a model that uses readily available encoder parameters as input to an artificial neural network to predict objective quality metrics for compressed video without using any reference and without need for decoding. The results verify its robustness for prediction of objective quality metrics in general and for PEVQ and PSNR in particular. The paper also focuses on reducing the complexity of the neural network.**

## I. Introduction

There has been a huge growth over the last years in multimedia applications for portable devices like mobile phones. A variety of methods for lossy compression for videos has been developed to manage bandwidth and memory usage, introducing specific artifacts impairing the visual quality as perceived by the end user. Therefore, video quality assessment has become important for many stake-holders like the mobile phone industry, network operators and video chatting application developers. The degradation of the video content is measured through various metrics of quality indicators in order to quantify the introduced artifacts, a way forward to improve the visual quality or to compare the competing methods and devices. Traditionally, objective metrics like signal-to-noise ratio (SNR), mean-square-error (MSE) and peak signal-to-noise ratio (PSNR) has been used to measure such distortions. However, end-user perception of the visual quality may not necessarily fall in line with the results of these measures [1]. Subjective testing for the quality evaluation requires human assessors and can provide the actual quality estimation but this method is quite expensive and not applicable in many situations. In order to achieve objective metrics closer to the human perception, Video Quality Experts Group (VQEG) started performing a comprehensive standardization of quality metrics and reported their results in [2] and [3]. Furthermore, the development of quality metrics like SSIM [4], its video adapted version VSSIM [5], VQM [6] and Opticom's PEVQ included in ITU-T Rec. J.247 [7] was witnessed. The aforementioned metrics calculate the quality measure by comparing the original frame with the distorted version of it, a full reference (FR) approach. In real-time applications like video streaming and video chat, access to the original frame is unlikely. Thus, the need of quality assessors which can work in the absence of any reference frame is a natural demand. Some of the recently introduced no-reference quality assessment methods include single feature based prediction [8] or an ensemble of objective features based prediction [9] [10] and [11] enlists many no-reference quality assessment metrics. These techniques involve appropriate processing of the received frame which makes them less usable for real-time applications. Rossholm et. al. [12] proposed a method of predicting video quality using coded bitstream information without decoding the video which makes the method useful for mobiles with limited processing power. The method uses multi-linear regression to map the encoder parameters as inputs for targets of video quality metrics. However, some of the used parameters exhibit non-linear relationship with the quality metrics. Artificial neural networks (ANN) are well known for their ability of handling non-linear problems in general and have been successfully used in the area of video quality assessment [8] [9] [10]. In this paper, we propose a no-reference ANN based video quality predictor which outperforms [12] in several prediction statistics. Video sequences used in this work were encoded by state-of-the-art H.264/AVC codec and have QCIF resolution. The proposed method has been found to be robust, fast and quite precise in terms of the statistics of its results. We think that it can be employed for monitoring visual quality in a network, e.g. 3G/4G cellular networks.

This paper is structured as follows: Section 2 gives information about the features used as input for the proposed model and the corresponding target quality metrics. The proposed model is discussed in section 3. Section 4 provides a discussion on the experimentation and results of this work. Finally, section 5 draws conclusive remarks about this paper.

## II. The inputs and targets for the model

This section gives insight into the choice of encoder parameters used in the proposed method and presents in brief the quality metrics which are predicted using the proposed ANN model.

### A. *Selection of the parameters*

While encoding a video, the coding process has to tune and calibrate various parameters to match the demand of specifications like bit-rate, frame rate and the pixel density. Such a parameter setting may be fixed from within a part

TABLE I
THE CORRELATION MATRIX OF PARAMETERS, REPRODUCED FROM [12]

| Parameter | Avg MV | P4x4 [%] | P16x16 [%] | Avg QP | Bits/Frame | P8x8 [%] |
|---|---|---|---|---|---|---|
| Avg MV | 1 | -0.04 | 0.45 | 0.04 | 0.19 | 0.45 |
| P4x4 [%] | -0.04 | 1 | 0.54 | -0.22 | -0.58 | -0.47 |
| P16x16 [%] | 0.45 | 0.54 | 1 | 0.14 | 0.12 | 0.19 |
| Avg QP | 0.04 | -0.22 | 0.14 | 1 | 0.65 | 0.54 |
| Bits/Frame | 0.19 | -0.58 | 0.12 | 0.65 | 1 | 0.47 |
| P8x8 [%] | 0.45 | -0.47 | 0.19 | 0.54 | 0.47 | 1 |

of a frame to the entire frame and ultimately for the whole video sequence under consideration. The coded bitstream carries motion vector arrays, quantized residual coefficients and header information. The idea used in building up this quality predictor is to extract and calculate certain parameters from this bitstream to be deployed as input for the ANN model. Among the many parameters which may be obtained and used for this purpose, the potentially most contributing ones for the quality prediction were selected in [12] and are given in table II.
However, the list of table II can be further simplified with a minor trade-off in the performance. The objective is to decrease the computational load of the ANN quality predictor. By observing the cross-correlation in table I between different parameters and the Pearson correlation coefficient values of prediction by the proposed model shown in Fig. 1, the following analysis is performed in order to reduce the number of parameters in a controlled manner.

- Fig. 1 shows that Bits/frame, P4x4, P8x8 and average QP are the most contributing parameters when each of them is used alone for quality prediction.
- The parameters to be employed for prediction should ideally represent different aspects of the coded video. These aspects could be the motion content dynamics of the video (avg MV), structural formation of the video frames (macroblock size in inter-coded frames) and bitrate (Bits/frame, avg QP) related information. Bits/frame and avg QP are hence chosen for their high individual contribution towards quality prediction (Fig. 1) and to have appropriate input regarding the bitrate information for the given resolution of the video under test. Several H.264 codec applications are not using the macroblocks of size 4x4. Hence, P8x8 is selected from this group.
- Avg MV has a low contribution to the quality prediction when used alone but it has a very low cross-correlation with the parameters avg QP and Bits/frame (table I). But, it is desirable to have motion content dynamics information to make the quality prediction process more robust.
- These arguments lead to the selection of Bit/frame, P 8x8, avg MV and avg QP as the simplified list of parameters to be used for a lesser complex video quality predictor.

### B. The quality metrics

Experiments have been performed using the proposed quality predictor to assess the following quality metrics.

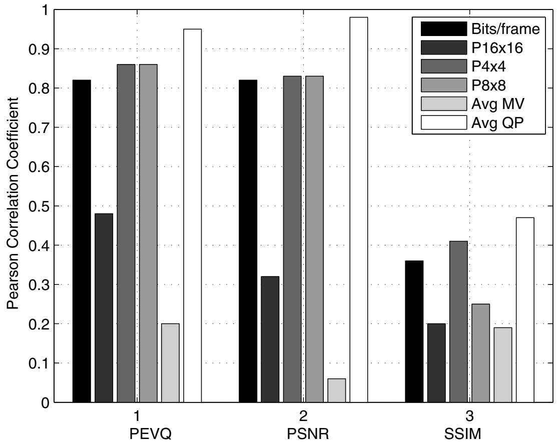

Fig. 1. Pearson correlation coefficient for individual parameter use.

TABLE II
THE USED PARAMETERS LIST

| Nr. | Parameter | Description |
|---|---|---|
| 1 | Bits/Frame | Bits per Frame |
| 2 | P16x16[%] | Percentage of inter blocks of size 16x16 |
| 3 | P4x4[%] | Percentage of inter blocks of size 4x4 |
| 4 | P8x8[%] | Percentage of inter blocks of size 8x8 |
| 5 | Avg MV | Average motion vector length |
| 6 | Avg QP | Average quantization parameter (QP) value |

1) PSNR. This metric is a quite widely used measure of quality of reconstruction of lossy compression and it is based on the mean-square error of the luminance values of the two frames under comparison.
2) PEVQ. Perceptual Evaluation of Video Quality estimates mean opinion scores (MOS) scores of the video quality by modelling the behaviour of the human visual tract. After successful benchmarks by the VQEG, PEVQ has become part of ITU-T Recommendation J.247 (2008) [7].
3) SSIM. Structural Similarity Index is a technique of measuring the structural similarity between two frames [4]. SSIM is a still used alternative way to evaluate perceptual quality.

## III. THE PROPOSED QUALITY PREDICTOR MODEL

The problem at hand may be addressed by a mapping function Y = F(X), with X = [$x_1$, $x_2$, $x_3$...$x_n$] where $x_1$, $x_2$, $x_3$...$x_n$ are the parameter vectors obtained from the coded bitstream

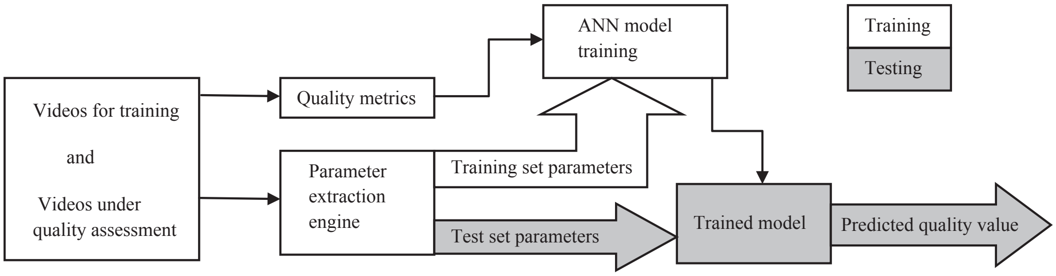

Fig. 2. The general framework of the proposed model

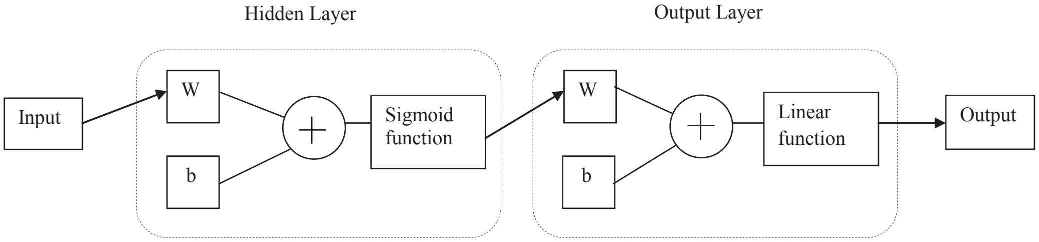

Fig. 3. The ANN architecture

of the videos under test produced by different codec settings. The matrix Y = [$y_1$ $y_2$ $y_3$...$y_m$] holds the corresponding video quality metric vectors $y_1$, $y_2$, $y_3$...$y_m$. So, building up such a function boils down to constructing a model which can regress between the set of parameters obtained from the coded bitstream and the related set of values of the quality metrics. The model should yield a pertinent value from vector y when a specific x vector is injected into it.

The desired mapping model may be designed using one of the many available methods like multi-linear regression in [12] or machine learning methods like artificial neural network [9] or support vector machine [13]. Each approach has its inherited advantages or disadvantages with regards to complexity and performance and any approach may has to trade off one advantage for the other. Linear regression is simple but unable to deal with the possible non-linearities, and machine learning methods are quite intelligent in handling complicated problems but they may become too complex to be implemented in devices with limited processing capacity. In this paper, an ANN based quality predictor is proposed which is relatively simple yet has proven its robustness in measuring video quality. The general framework of the proposed model is as depicted in the Fig. 2.

### A. The architecture of the model

Contrary to the contemporary networks like [9] or [10] used for video quality prediction with no-reference, a reasonably simpler architecture for the ANN model is proposed here, using a two-layer feed-forward network with 10 or 6 sigmoid hidden neurons and one linear output neuron, depicted in Fig. 3. As will be shown in section IV, the work presented has aimed at decreasing the required number of input parameters compared to the number used in [12]. The proposed model performs best with 10 hidden neurons when it uses all the listed parameters and it uses only 6 neuron when it is employed with lesser number of input parameters.

The network is trained with Levenberg-Marquardt backpropagation algorithm which is considered to be a fast method. The performance of the model is evaluated using various error statistics and regression analysis. The already stated number of neurons in the hidden layer has been fixed as a result of experiments under various number of hidden neurons to achieve an efficient system with high performance and no over-fitting. A properly trained network gives reasonable answers when tested on unseen inputs. Typically, a new input leads to an accurate output, if the new input is similar to inputs used in the training set. To this end, the network should be trained to have generalization property. Usually, generalization is achieved by regularization or early stopping. Experiments have been conducted in this regard and it was found out that early stopping makes the training process faster than regularization and hence early stopping was used in this work. The training data is further divided into training set and validation set. The error on the validation set is monitored during training. The validation error normally decreases during the initial phase of training, as does the training set error [14]. However, when the network begins to over-fit the data, the error on the validation set typically begins to increase, and the training is stopped when the validation increases over a fixed number of iterations and the weights and biases at the minimum of the validation error are used for testing on a different data set.

### B. The video sequences data set

The proposed predictor has been tested and trained on a wide range of bit-rates and frame rates of video sequences with a variety of motion content dynamics. The video sequences

TABLE III
THE AVAILABLE SAMPLE SPACE OF VIDEOS

| **Session** | Training (168 videos) | Testing (120 videos) |
|---|---|---|
| **Video sequence** | Foreman, Cart, Mobile, Shine, Fish, Soccer goal, Car phone | Cropped 3G |
| **bitrate (bps)** | 30, 40, 50, 100, 150, 200 | |
| **frame rate (fps)** | 7.5, 10, 15, 30 | |

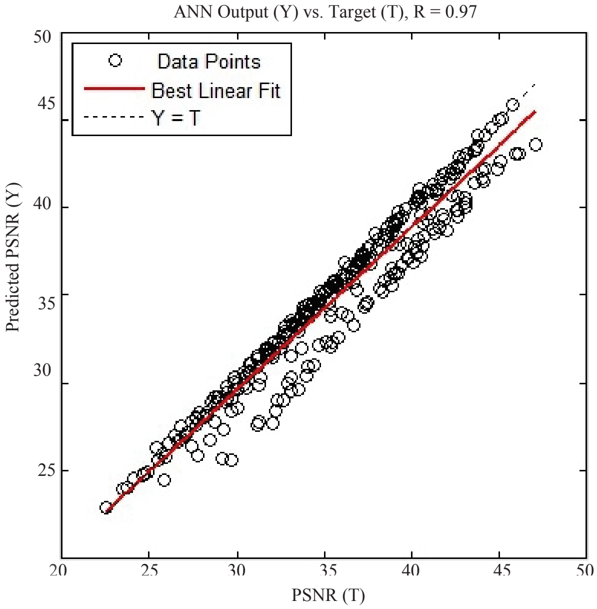

Fig. 4. Regression plot for PSNR assessment

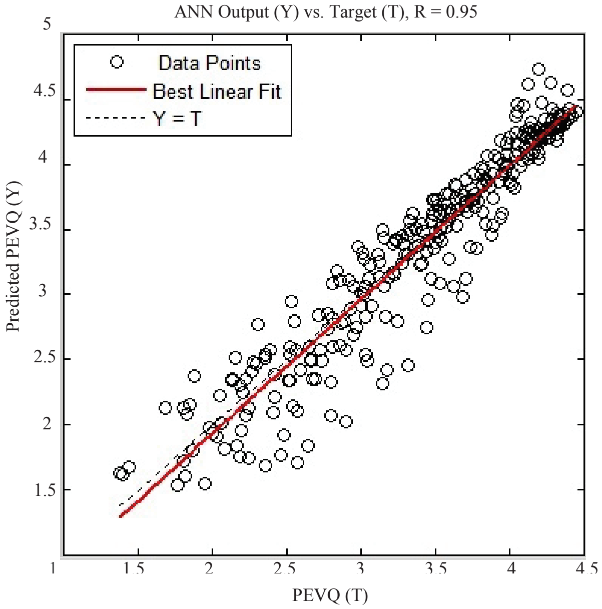

Fig. 5. Regression plot for PEVQ assessment

used in this work were approximately 3 seconds long (90, 45, 30, and 23 frames) in QCIF resolution and they were encoded with the H.264/MPEG-4 AVC reference software, version 12.2 generated by JVT [15] using the baseline profile. In table III, the given 7 video sequences were encoded using 6 different bit-rates and 4 different frame rates as stated in the table III, and this resulted into 7x6x4 = 168 videos for training. These were further divided into actual training set and validation set to have better generalization of the neural network. The trained network was tested on 120 video sequences which were produced by clipping five sequences from different parts of the 3G video sequence, cropping to QCIF resolution and encoding according to the same bit-rates and frame rates as for training set and hence resulted into 5x6x4 = 120 videos. The five parts of the 3G video sequence had different characteristics and motion dynamics and the original letter box aspect ratio of 3G video sequence was not maintained.

## IV. Experiments and Results

The proposed model was trained with the training set for various quality metrics using the parameters listed in table II with 10 neurons in the hidden layer and using the simplified list (Bit/frame, P8x8, avg MV and avg QP value) with 6 neurons in the hidden layer. The network was trained on the training data set and was then tested on the test set as provided in table III. For each quality metric, the experiment was run five times and then an average measure was recorded. Table IV shows the performance of the proposed method and also offers a comparative view with the results obtained by the multi-linear regression (MLR) method. Fig. 4 shows a sample regression graph for PSNR prediction with 6 neurons in the hidden layer using the simplified list of four parameters as input. Similar plots are presented in Fig. 5 and Fig. 6 for PEVQ and SSIM assessment respectively. As the results show, the proposed model is capable of predicting the video quality with a considerable high precision and accuracy. The error statistics and Pearson correlation coefficient values as given in table IV motivate why this model could be preferred over the MLR method for quality prediction. A considerable care has been taken to avoid over-fitting the network and early stopping approach worked well in this regard. Also, the number of neurons in the hidden layer is a crucial parameter to be aware of and it was optimized on the basis of tests performed in the work. Table IV also shows results under column *ANNsimp* obtained from the proposed model using only 6 neurons in its hidden layer and taking simplified list of parameters as its input for quality prediction.These results further substantiate the motivation of using the proposed ANN predictor for quality prediction as even after using 4 out of the 7 parameters used in MLR approach, the ANN predictor works better.

The proposed model is designed to measure compression artifacts but this method can be extended to be employed for

TABLE IV

AN AVERAGE OF THE RESULTS OBTAINED BY USING THE PROPOSED PREDICTOR

| **Quality metric** | PSNR | | | PEVQ | | | SSIM | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method \ Stats** | ANN | ANNsimp | MLR | ANN | ANNsimp | MLR | ANN | ANNsimp | MLR |
| Pearson correlation coefficient | 0.99 | 0.97 | 0.99 | 0.98 | 0.95 | 0.93 | 0.76 | 0.73 | 0.62 |
| Mean square error | 0.91 | 2.3 | 4.74 | 0.03 | 0.065 | 0.12 | 0.00 | 0.00 | 0.00 |
| Standard deviation | 0.88 | 1.28 | 1.43 | 0.18 | 0.25 | 0.33 | 0.04 | 0.04 | 0.04 |
| Mean error | 0.37 | 0.8 | -1.65 | 0.03 | 0.025 | -0.09 | 0.01 | 0.01 | -0.04 |
| Mean of absolute error | 0.61 | 0.96 | 1.77 | 0.13 | 0.18 | 0.27 | 0.02 | 0.04 | 0.05 |

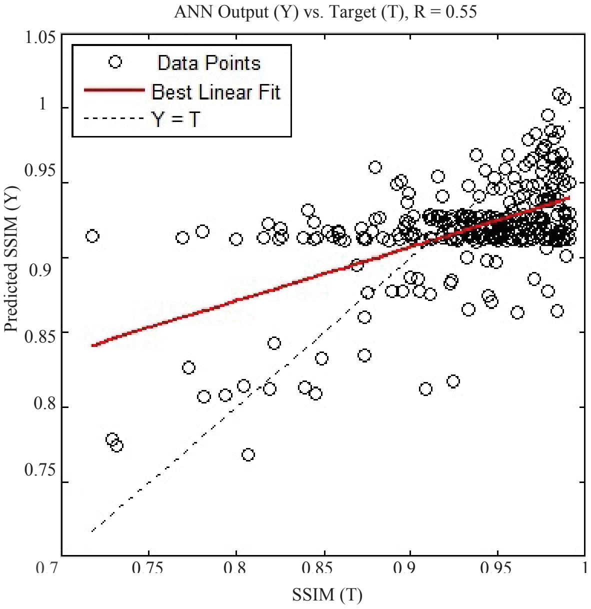

Fig. 6. Regression plot for SSIM assessment

other issues like transmission and network error if the model is expanded with related features. The important consideration in this regard is to train the network on a reasonably large data set to make the model work for acceptable value of predictions.

## V. CONCLUSION

A robust, reference free method to predict mostly used perceptual quality metrics for coded video sequences has been proposed, using only features readily available in the coded bitstream. The method predicts the PSNR and PEVQ metrics quite accurately, while the SSIM metric is predicted with less accuracy. For all three metrics, the proposed method performs better than, or equal to, the earlier proposed method in [12] in all the presented statistics. The proposed method performs better than [12] even when it is used in its reduced complexity form of ANN architecture using lesser number of input parameters, except for the correlation coefficient for PSNR. The main reason why the proposed ANN based model outperforms the linear regression approach is the possible non-linear dependency of the input decoder parameters and the output target quality metric values. These promising results even with a reduced complexity ANN architecture encourage continued development of this neural network based predictor. The accuracy and precision of results obtained in this work for predicting the objective mean opinion scores (MOS) for PEVQ, which has been accepted by ITU-T to be rather close to the human perceptual assessment, show the potential of this predictor to be employed in prediction of MOS of subjective tests. The future work should focus on developing the same model for subjective MOS.

## REFERENCES

[1] S. Winkler, *Digital Video Quality - Vision Models and Metrics*. John Wiley and Sons, 2005.

[2] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” Mar. 2000.

[3] ——, “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii,” Aug. 2003.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactios on Image Processing*, vol. 13, pp. 600–612, April. 2004.

[5] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication, Special issue on Objective video quality metrics*, vol. 19, pp. 121–132, Feberuary. 2004.

[6] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, pp. 312–322, September. 2004.

[7] I.-T. R. J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference,” 2008.

[8] M. Ries, O. Nemethova, and M. Rupp, “Motion based reference-free quality estimation for h.264/avc video streaming,” in *2nd International Symposium on Wireless Pervasive Computing (ISWPC 07)*. IEEE, 2007.

[9] X. Jiang, F. Meng, J. Xu, and W. Zhou, “No-reference perceptual video quality measurement for high definition videos based on an artificial neural network,” in *Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on*. IEEE, 2008.

[10] D. Culibrk, D. K. D, P. Vasiljevic, M. Pokric, and V. Zlokolica, “Feature selection for neural-network based no-reference video quality assessment,” *Lecture Notes in Computer Science*, vol. 5769/2009, pp. 633–642, 2009.

[11] U. Engelke and H.-J. Zepernick, “Perceptual-based quality metrics for image and video services: A survey,” in *Next Generation Internet Networks, 3rd EuroNGI Conference on*, may 2007, pp. 190 –197.

[12] A. Rossholm and B. Lovstrom, “A new low complex reference free video quality predictor,” in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*. IEEE, 2008.

[13] M. Narwaria and W. Lin, “Objective image quality assessment based on support vector regression,” *Neural Networks, IEEE Transactions on*, vol. 21, pp. 515–519, March 2010.

[14] M. Hagan, H. Demuth, and M. Beale, *Neural Network Design*. Boston: MA: PWS Publishing, 1996.

[15] JVT, “H.264/MPEG-4 AVC Reference Software,” *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), 24th Meeting: Geneva, CH, 29*, 2007.