Estimating PSNR in High Definition H.264/AVC Video Sequences Using Artificial Neural Networks

Martin SLANINA, Václav ŘÍČNÝ

Dept. of Radio Electronics, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic

slanina@ieee.org, ricny@feec.vutbr.cz

Abstract. *The paper presents a video quality metric designed for the H.264/AVC codec. The metric operates directly on the encoded H.264/AVC bit stream, parses the encoding parameters and processes them using an artificial neural network. The network is designed to estimate peak signal-to-noise ratios of the video sequence frames, thus enabling computation of full reference objective quality metric values without having the undistorted video material prior to encoding for comparison. We present the metric framework and test its performance for LDTV (low definition television) as well as HDTV (high definition television) video material.*

Keywords

H.264/AVC, video quality, objective quality metric, HDTV, artificial neural network.

1. Introduction

Digital video processing techniques and methods, however advantageous and efficient, can be characterized with a common issue concerning performance testing and quality assurance.

Let us focus only on the video compression algorithms at this time, and study the impact of the source coding process on the visual quality of the processed video material. Errors and faults appearing during the compressed video transmission and/or storage will not be taken into account. The abovementioned quality assessment issue has its roots in the fact that for digital video processing systems, it is not only the system settings affecting the resulting video quality, but – to a significant extent – the actual video content plays a decisive role. Highly detailed video frames (high spatial activity) with rapid changes over time (high temporal activity) are typically the worst-case scenario.

The word 'quality' has been mentioned in the previous paragraphs with no real explanation of its meaning in the area of video processing. Ideally, the video quality is described and quantified using subjective quality tests, typically employing a group of human observers to rate the quality, identify visibility of distortions, etc. Such tests are well understood, however their usability is limited due to the cumbersomeness of organizing test sessions. The common aim in video quality research is thus in replacing such subjective procedures with objective measurements, i.e. methods capable of evaluating the quality automatically. The benchmark of objective tests is commonly the correlation with subjective test results.

Full reference objective metrics are based on a comparative approach, as the original as well as processed (and distorted) video material is available. They reach from the simplest pixel-based metrics (such as the peak signal-to-noise ratio – PSNR) to the more sophisticated, based on either the 'psychophysical approach' [1], [2] or 'engineering approach' [2], [3]. The problem of full reference metrics is quite well understood and evaluation studies show good results.

On the other hand, no reference metrics, especially for the emerging video compression standards, are still in their infancy. A typical solution for the older, especially DCT-based (discrete cosine transform) algorithms, is in detecting typical artifacts such as blocking, blur, etc. [4]. The H.264/AVC, however, does not have such typical artifacts, as a deblocking filter is employed at its output to adaptively smooth the block areas. Our approach is thus in examining the encoded H.264/AVC data and make a statement on video quality using the encoding parameters present in the bit stream.

In this paper, a method capable of estimating PSNR values of the encoded video frames is presented, only relying on the information present in the encoded bit stream, and removing the necessity of having the undistorted original available for comparison. A similar problem was recently considered in [5], using a different approach and different bit stream parameters.

Section 2 briefly describes the full reference metric whose outputs are desired by our system – the PSNR, sections 3 and 4 present the framework and design details of the no reference system. The test settings are described in section 4 and the performance over different video sets is discussed in section 5.

2. The Full Reference Approach

Among the full reference objective quality metrics, the peak signal-to-noise ratio (PSNR) holds its strong position and is quite often used although it is known to correlate poorly with the subjective scores [1]. However, it is well understood and, above all, fairly easy to implement.

The PSNR is given by

$$\mathrm{PSNR} = 10\log_{10}\frac{m^2}{\mathrm{MSE}} \tag{1}$$

where m is the maximum value a pixel can take and MSE is the mean squared error, defined as

$$\mathrm{MSE} = \frac{1}{MNT}\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{T}\left[f(i,j,k) - \tilde{f}(i,j,k)\right]^2. \tag{2}$$

The symbols M, N, T represent the video frame width, frame height and the number of frames the PSNR is being computed for, respectively, $f(i,j,k)$ and $\tilde{f}(i,j,k)$ are the luma pixel values of the original and the distorted video, respectively.

A number of different full reference quality algorithms has been presented in the last decades. Some of these algorithms are capable of reaching quite high correlations with the subjective scores.

3. Proposed PSNR Estimation Scheme

As noted above, the PSNR estimation algorithm is supposed to work only with the encoded bit stream of H.264/AVC. Let us observe what parameters available in the encoded bit stream are likely to carry information on the quality of the decoded video:

3.1 Prediction Modes

There are two groups of prediction modes in which data can be predicted in the H.264/AVC – intra prediction modes and inter prediction modes. For intra modes, the prediction is done from neighboring samples within the same image (slice), in the simplest case, just copied in a selected direction. The predicted block size for intra prediction is not fixed, but can be altered depending on the encoder's choice from 16 x 16 pixels, over 8 x 8 pixels down to 4 x 4 pixels as listed the first column in Tab. 1. The IPCM mode enables the encoder to code the data directly with no prediction. The block size and the prediction direction is then signaled in the bit stream. Large predicted blocks are likely to be chosen in smooth areas with low spatial activity. Choosing large blocks requires fewer bits to signal prediction process to the encoder, and thus in some situations the encoder might choose large blocks to spare bits even though the prediction accuracy decreases.

In inter prediction modes, previously encoded and decoded pictures are used as reference. Blocks are then predicted using motion compensation in either one or two directions. Again, the block size is not fixed here. The available block sizes are listed in the second column in Tab. 1. Even though there is a kind of hierarchy described in the standard, the block size used is a sufficient parameter for our application. In the direct mode, the pixel values are simply copied from the reference picture.

Prediction modes	
Intra prediction	**Inter prediction**
Intra 16x16	Direct 16x16
Intra 8x8	Inter 16x16
Intra 4x4	Inter 16x8
IPCM	Inter 8x16
	Inter 8x8
	Inter 8x4
	Inter 4x8
	Inter 4x4

Tab. 1. Prediction modes and prediction block sizes available in the H.264/AVC.

3.2 Quantization

The predicted data are compared to the original image blocks and the differences remain to be encoded. There are several available transform algorithms in the H.264/AVC [6]. What we are interested in is the coarseness of transform coefficient quantization. The standard defines a quantization parameter, ranging from 0 to 51. The higher the quantization parameter, the coarser is the quantization and the lower quality of the decoded video can be expected.

3.3 System Framework

The operation of the PSNR estimating system can be described by a flowchart displayed in Fig. 1. For each frame (or field if processing interlaced video), the encoding parameters are first read from the network abstraction layer NAL [6], which takes care of proper encoded data handling for a selected application. The next step is then decoding variable length codes, which is not explicitly mentioned in the flowchart but can be included in the 'read' block. Having the frame encoding parameters available, the system continues based on the frame type – in case of intra frame, the frame only includes intra coded blocks as listed in the left column in Tab. 1. On contrary, inter frames may include any block type from Tab. 1.

For intra frames, the average quantization parameter is calculated as it can be altered throughout the frame. However, the encoder does not necessarily have to have this feature implemented, which is also the case in our setup (see Sec. 5). The next input parameter for the PSNR evaluation is formed by the ratios of the respective prediction modes throughout the frame. The inputs are then fed to an artificial neural network (ANN) which outputs an esti-

mated PSNR. At this point, PSNR estimation is done for one frame. In case an inter predicted frame is being processed, the whole system operates on the same basis as for intra predicted frames, with one significant difference. For inter prediction, reference frames are used for motion compensated prediction, so the quality of the reference picture (reference PSNR) naturally impacts the quality of the predicted block or frame (estimated PSNR). One additional parameter is thus introduced for each inter prediction mode, defining the average reference PSNR (see Sec. 4).

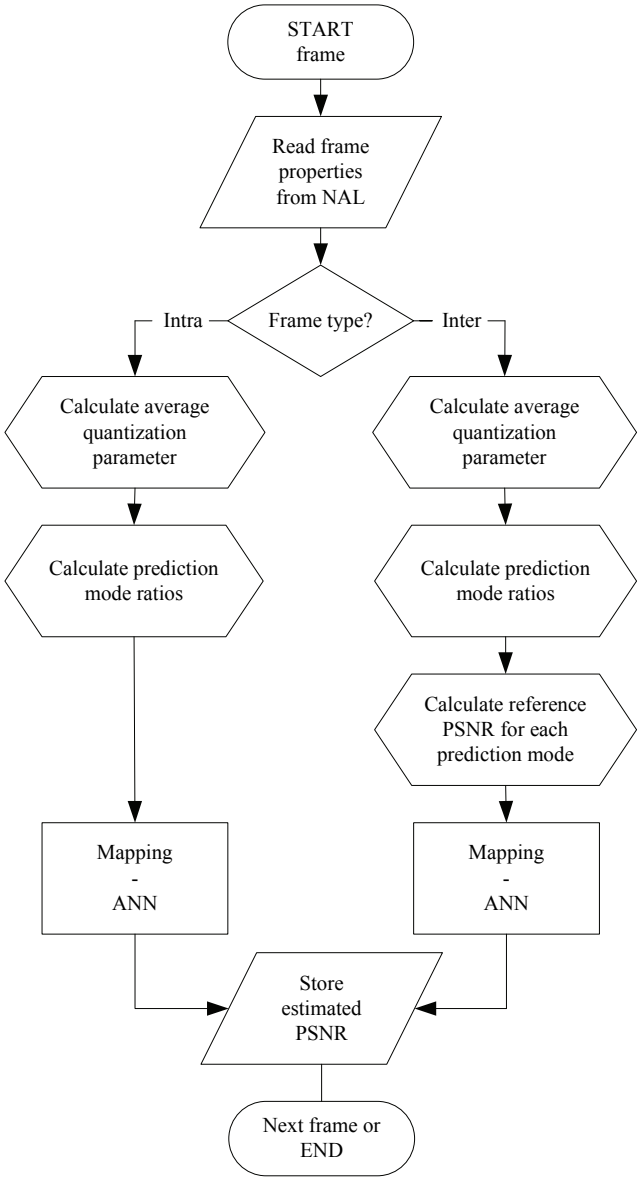

Fig.1. Basic system flowchart for intra and inter predicted frames.

4. System Design Details

This section is intended to more thoroughly analyze some of the system key blocks introduced in Sec. 3.

4.1 Parsing the Bit Stream

As mentioned above, we need a tool to read the required H.264/AVC coded data from the bit stream and to decode the variable length codes used by the standard. Even though a straightforward approach would be to implement these functions directly according to the standard, there is a software available that takes care of all this – the H.264/AVC encoder/decoder reference software [7]. Its code is written in C and is freely available for use and modification. We modified the C code in order to extract the desired parameters. As there is no need to decode the actual pixel values within the frames, a significant part of the decoder can be disabled.

4.2 Reference PSNR

Let us now jump over to the last block prior to ANN operation for inter predicted frames. As noted above, we need to have an estimate of the PSNR in the reference pictures, the motion compensated prediction is done from. For each of the inter prediction modes (right column in Tab. 1) we want one number representing the average reference PSNR computed from the reference pictures.

Assume we have an area consisting of N_1 blocks having the peak signal-to-noise ratio PSNR_1, an area of N_2 blocks with PSNR_2, etc. The corresponding MSE of each area can be expressed from (1) as

$$\mathrm{MSE} = \frac{m^2}{10^{\frac{\mathrm{PSNR}}{20}}}. \tag{3}$$

Calculating overall mean squared error of the whole area using a weighted average yields

$$\begin{aligned} \mathrm{MSE}' &= \frac{N_1\mathrm{MSE}_1 + N_2\mathrm{MSE}_2 + \cdots + N_x\mathrm{MSE}_x}{N_1 + N_2 + \cdots + N_x} \\ &= \frac{\frac{N_1 m^2}{10^{\frac{\mathrm{PSNR}_1}{10}}} + \frac{N_2 m^2}{10^{\frac{\mathrm{PSNR}_2}{10}}} + \cdots + N_x \frac{N_x m^2}{10^{\frac{\mathrm{PSNR}_x}{10}}}}{N_1 + N_2 + \cdots + N_x} \\ &= \frac{m^2 \sum_{i=1}^{x} \frac{N_i m^2}{10^{\frac{\mathrm{PSNR}_i}{10}}}}{\sum_{j=1}^{x} N_j}, \end{aligned} \tag{4}$$

where x is the total number of different PSNR values in the examined area. Substituting the result of (4) back into (1) gives the overall reference PSNR as

$$\begin{aligned} PSNR' &= 10\log_{10} \frac{m^2 \sum_{j=1}^{x} N_j}{m^2 \sum_{i=1}^{x} \frac{N_j}{10^{\frac{PSNR_j}{10}}}} \\ &= 10\log_{10} \sum_{j=1}^{x} N_j - 10\log_{10} \sum_{i=1}^{x} \frac{N_i}{10^{\frac{PSNR_i}{10}}}. \end{aligned} \tag{5}$$

The expression in (5) is evaluated for each prediction mode and the PSNRs of the associated reference pictures. If the video is encoded in Main, Extended or High profile, the inter prediction may be bi-directional. In such case, (5) is used first to compute one value of PSNR for each predicted block – as there is one reference in each direction, the x is equal to 2.

5. Video Sequences

The artificial neural networks have been trained using a set of 10 short video sequences in CIF format (352 x 288 pixels). The first frames of the training sequences are displayed in Fig. 2, the sequences are available at [8].

The sequences were encoded using the H.264/AVC reference encoder in Main profile at Level 3.0 with seven different configurations each [7]. The altered parameters are listed in Tab. 2.

Setting number	Target bitrate [kbps]	Initial quantization parameter
1	100	25
2	1000	25
3	100	35
4	1000	35
5	100	45
6	1000	45
7	5000	45

Tab. 2. Encoder configuration for training video sequences.

To verify the PSNR estimation algorithm for a different set of video sequences with a different resolution, two sets were used, including video sequences in 720p HDTV resolution (1280 x 720 pixels) and in full HD resolution 1080p (1920 x 1080 pixels). The sequences are again freely available [9].

The variable encoder parameters for the 720p video sequences are the same as those for the training sequences in CIF resolution (Tab. 2). As the encoding of 1080p sequences is quite demanding, fewer configurations were used for this format as listed in Tab. 3. The HDTV sequences were encoded at Main profile, but the Level had to be changed to 3.1 and 5.0 for the 720p and 1080p sequences, respectively.

Setting number	Target bitrate [kbps]	Initial quantization parameter
1	5000	25
2	1000	35
3	100	45

Tab. 3. Encoder configuration for 1080p video sequences.

Fig. 2. Video sequences in CIF resolution used for network training.

Fig. 3. HDTV video sequences used for evaluation and verification of network generalization ability.

6. Results

Having the set of low resolution training video sequences created and desired parameters extracted, different network configurations were trained on training set of parameters using the least mean squares algorithm with Bayesian regularization [10].

The intra and inter predicted frames were treated separately, which means different networks were designed and trained for each frame type. Neural networks were

trained on the CIF resolution training sequences and their performance was verified for HDTV set of video sequences. The optimization criterion for the training was the mean squared error of the real and estimated PSNRs over the training set.

6.1 Intra Frames

For the intra frames, there are only five input parameters to the artificial neural network. The simplest network configuration is the linear unit, which is capable of representing any configuration of linear neurons [10]. Surprisingly, even such a simple configuration gives quite nice results. For the 720p video sequence test set, the linear unit reached a correlation coefficient between the real and the estimated PSNR values 0.9681 and a mean squared error (MSE) of 3.014. Anyway, as the performance of the system for a whole video sequence vastly depends on the estimation accuracy of the first frame in the sequence, we will use a more complicated network configuration in our consideration. Out of the several tested configurations, a three layer network with five units in the first layer, two units in the hidden layer and one linear unit in the output layer was selected. Such network reached a correlation of 0.9715 and MSE equal to 2.088.

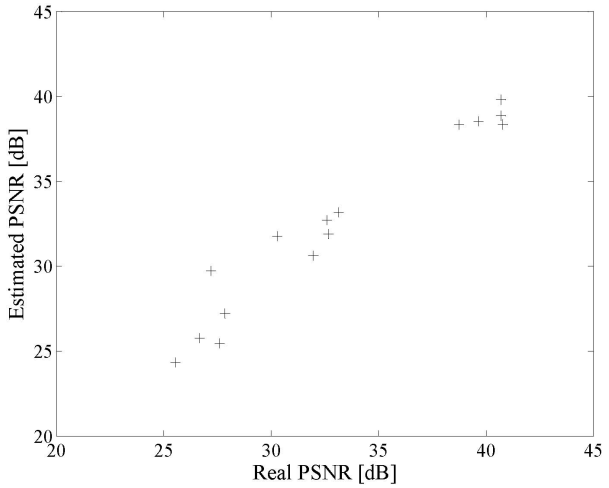

Fig. 4. Scatter plot diagram: Real versus estimated PSNR values for intra coded frames. 720p test set, three-layer network.

The scatter plot diagram of the real and estimated PSNR values for the three-layer network is shown in Fig. 4. Please note that even though there are 35 compressed video sequences available in the 720p test set, there are fewer values in the diagram as for some configurations the first frames were encoded in the same way.

6.2 Inter Frames

The PSNR estimation problem seems to be more difficult as it turned out that a simple linear neuron is not capable of reaching satisfactory results. The configuration of a multi-layer perceptron (MLP) was thus used instead. The MLP is supposed to be made up from three layers, with the first layer having as many units as there are inputs to the network. The number of units in the second (hidden) layer is arbitrary and should be selected to give best results for the desired application. The number of units in the third (output) layer is determined by the desired number of network outputs – in our case it will only have one unit as the only output we require is the estimated PSNR.

Tab. 4 lists the results for the hidden layer having one to six neuron units. In the computation of MSE of the training set, only inter frames are considered. For MSE and correlation coefficient of the test set, even the intra frames are taken into account and the results represent the whole considered video sequences.

It is obvious the results differ very slightly for the varying network configurations. The changes are more likely to be caused by different initial network weight setting rather than limiting network capabilities. As we are using a regularized training algorithm, the risk of overfitting is minimized and the networks are usable even when the number of hidden units is larger than necessarily needed. Fig. 5 shows a scatter plot diagram of the real and the estimated PSNR values for the whole set of 720p testing sequences. A similar diagram for 1080p sequences is displayed in Fig. 6. In both diagrams, an MLP with three units in the hidden layer was used.

	CIF	720p		1080p	
Hidden units	Training MSE	Test MSE	Test Corr.	Test MSE	Test Corr.
1	1.644	4.369	0.915	4.489	0.914
2	1.654	4.258	0.916	5.213	0.902
3	1.554	4.502	0.914	4.673	0.923
4	1.537	4.390	0.916	3.920	0.923
5	1.488	4.866	0.911	4.659	0.922
6	1.496	3.793	0.923	3.879	0.928

Tab. 4. PSNR Estimation results for high definition video sequences.

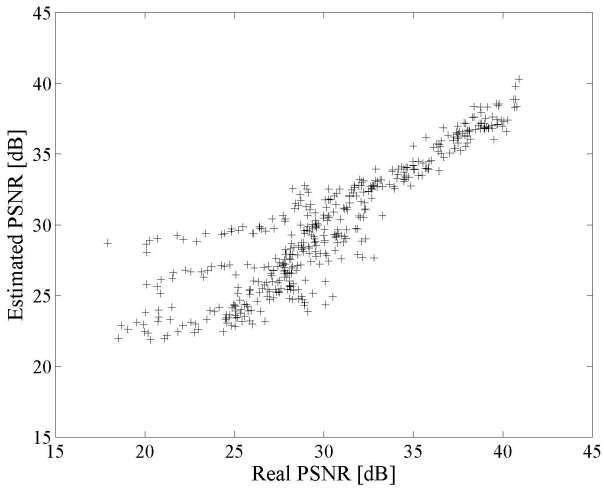

Fig. 5 Scatter plot diagram: Real versus estimated PSNR values for inter coded frames. 720p test set, three-layer network, four units in the hidden layer.

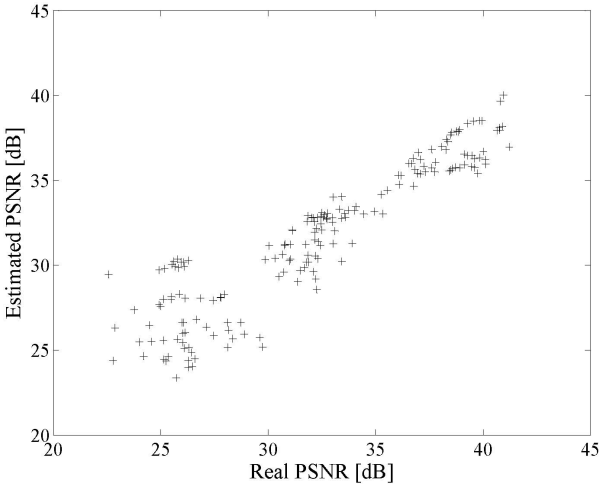

Fig. 6. Scatter plot diagram: Real versus estimated PSNR values for inter coded frames. 1080p test set, three-layer network, four units in the hidden layer.

Notice the declination of the results from the ideal values in the left part of Fig. 5. Obviously, there are sequences in the test set for which the algorithm gives wrong outputs and the peak difference can reach up to 10 decibels which is an unacceptable value. This issue might be solved by a different selection of the training sequence set. The network outputs are much more accurate for the video sequences with high signal-to-noise ratios.

7. Conclusion

We have presented a system capable of estimating PSNR for real H.264/AVC compressed video sequences. The results show the approach is quite universal in terms of video resolution as it was designed using low resolution video sequences and with increasing resolution it still performs reasonably well. As the system is designed quite universally, it should be possible to train the artificial neural network to estimate different target values. Experiments with subjective score estimation will follow.

Acknowledgements

The research described in the paper was financially supported by the Czech Grant Agency under grant No. 102/ 08/H027, and by the Czech Ministry of Education by the research program MSM 0021630513.

References

[1] WINKLER, S. *Digital Video Quality: Vision Models and Metrics*. Chichester: Wiley, 2005.

[2] WU, H. R., RAO, K. R. *Digital Video Image Quality and Perceptual Coding.* Boca Raton: Taylor & Francis, 2006.

[3] WANG, Z., LU, L., BOVIK, A. C. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, February 2004, vol. 19, no. 2, p. 121-132.

[4] WANG, Z., BOVIK, A. C., EVANS, B. L. Blind measurement of blocking artifacts in images. In *Proceedings of 2000 International Conference on Image Processing*, vol. 3, p. 981-984, 2000.

[5] EDEN, A. No-reference estimation of the coding PSNR for H.264-coded sequences. *IEEE Transactions on Consumer Electronics*, May 2008, vol. 53, no. 2, p. 667-674.

[6] ITU-T Recommendation H.264, *Advanced Video Coding for Generic Audiovisual Services*. Geneva: The International Telecommunication Union, 2006.

[7] SUEHRING, K. *The H.264/MPEG-4 AVC reference software – JM11.* [online] Available: http://iphome.hhi.de/suehring/tml/download/

[8] Arizona State University, Video Traces Research Group. *CIF Sequences*. [online] Available: http://trace.eas.asu.edu/yuv/cif.html

[9] HAGLUND, L. *The SVT High Definition Multi Format Test Set.* Sveriges Television AB, 2005 [online] Available: http://www.ebu.ch/en/technical/hdtv/test_sequences.php

[10] DEMUTH, H., BEALE, M. *Neural Network Toolbox for Use With MATLAB. User's Guide*, version 4. Natick: Mathworks, Inc., 2000.

About Authors...

Martin SLANINA was born in 1982. He received his master's degree from the Brno University of Technology in 2005. Since then, he is pursuing his Ph.D. degree at the Department of Radio Engineering at the same university. His research is focused on television technology, image and video processing and video quality assessment in particular.

Václav ŘÍČNÝ was born in 1937. He is a professor at the Department of Radio Electronics, Brno University of Technology. His research interest includes, in particular, video and television technology, analogue and digital processing and digital measurement.