A REDUCED REFERENCE VIDEO QUALITY METRIC FOR AVC/H.264

Tobias Oelbaum, Klaus Diepold

Institute for Data Processing Technische Universität München
Munich, Germany
oelbaum@tum.de, kldi@tum.de.

ABSTRACT

A reduced reference video quality metric for AVC/H.264 is proposed. This reduced reference metric evaluates a set of features such as blur or blocking and combines these measurements into one quality information using multivariate data analysis. The metric needs a very low overhead in terms of additional bit rate and is verified with data from accurate subjective testing. It does not only outperform standard PSNR but also shows to be superior compared to two full reference video quality metrics. In addition a method for designing video quality metrics is presented. This method is based on multivariate data analysis, a tool that is widely used in chemo metrics and food science to predict latent variables such as taste by evaluating a set of variables that can easily be measured.

1. INTRODUCTION

Four years after the first version of the upcoming video coding standard AVC/H.264 [1] was released, next to no results exist to demonstrate the prediction capabilities of video quality metrics for AVC/H.264 encoded video data. Up to now most video quality metrics have been verified using MPEG-2 encoded videos, but as AVC/H.264 encoded video has significant different characteristics (e.g. no fixed block sizes, filtering in the decoder loop), those results do not necessarily apply for this new generation of encoded video.
Being the de-facto standard for objective video quality metrics PSNR is still used for comparing AVC/H.264 with other video codecs or for comparing different encoder implementations or coding settings for AVC/H.264. This is in spite of the knowledge, that PSNR values may be misleading [2], meaning that PSNR may not even give an indication about which of two coded videos does have a better visual quality. Video quality metrics such as the approach presented by Gastaldo et al in [3] that were especially designed for MPEG-2 video or a different unique video codec obviously can not predict the quality for an AVC/H.264 encoded video as precisely as if the codec for which this metric was developed was used. Due to tools such as the inloop filter that are an integral part of AVC/H.264, also more generic video quality metrics may have a reduced prediction accuracy for AVC/H.264.
While the adoption of AVC/H.264 video in the market has already reached a remarkable level, very few data about objective methods for measuring the quality that can be gained with this codec has been presented. So apart from conducting a precise but time consuming subjective test the answer to the essential question about the quality of an AVC/H.264 encoded video relies on guessing and assumptions only.
The rest of the paper is organized as follows: In section 2 a short overview about related works is presented. The method used to develop the proposed reduced reference metric is described in section 3 and the model itself is described in detail in section 4. Section 5 presents the results for the proposed method and finally section 6 concludes this paper.

2. RELATED WORKS

2.1 Full Reference Quality Metrics

The most popular video quality metric is the Peak Signal to Noise Ratio (PSNR). This simple metric just calculates the mathematical difference between each pixel of the encoded video and the original video. In fact up to now PSNR is the only video quality metric that is widely accepted and therefore PSNR is the de-facto standard for measuring video quality.
In 2004 the ITU released a recommendation which included four different full reference (FR) metrics (not only the coded video but also the original video is needed for the evaluation) which outperformed PSNR in terms of correlation to results of extensive subjective tests. Among those is the Edge PSNR method [4] developed by Lee et al which was chosen as a comparison point to the metric presented in this contribution. This metric is based on the observation that human observers are especially sensitive to degradations in regions around edges. Therefore this metric evaluates the PSNR only at those pixels that have been classified to belong to an edge region (this classification can be easily done using a edge detection algorithm such as the Canny algorithm [5]). Another FR image metric which has gained a high popularity since it was introduced in 2002 is the so called SSIM (Structural SIMilarity index) [6, 7]. This metric was the second metric chosen for comparison. The SSIM is built on the assumption that the human observer wants to gather the structural information of an image, which is independent of average luminance and contrast and therefore the image quality is closely related to how much this structural information can be retained. The SSIM performs a separate comparison on luminance, contrast and structure in the original and the coded image and uses this information to calculate one overall quality index.

2.2 Reduced Reference and No Reference Quality Metrics

Comparably few approaches were presented for reduced reference (RR) quality evaluation and even less for no reference (NR) quality evaluation. For a RR metric only parts of the original video or some extracted properties of this video are needed for evaluation. For a NR metric no information about the original video is needed. One popular approach for a NR image and video quality metric is the insertion of watermarks in the original image and then measuring the amount to which these watermarks can be recovered at the receiver [8, 9]. Wang and Simoncelli showed that natural images have a certain frequency distribution and therefore the frequency distribution of a coded image can be used to predict the visual quality [10]. Recently Callet et al presented an approach using a neural network system to learn how human quality perception is influenced by different image properties [11]. In addition to complete quality metrics there exist several measurements that concentrate on one single image property or a special artifact. Prominent candidates from this field are the blocking measurement introduced by Bovik and Wang [12], or the blur measurement proposed by Winkler [13].

3. MULTIVARIATE DATA ANALYSIS FOR OBJECTIVE VIDEO QUALITY ASSESSMENT

3.1 Challenges in Building Objective Quality Metrics

Video quality metrics that try to model the human visual system (HVS) face the problem that what should be modeled is very complicated and up to the moment not well understood. Measuring the strength of a certain artifact (e.g. blocking, blur) and trying to predict the quality by a linear combination of the measured artifacts introduces the problem that it is not known to which extend a certain artifact affects the perceived video quality. In addition this method ignores the possibility that there may be interferences between certain types of artifacts.

For these two reasons it is proposed to design new video quality models using methods provided by multivariate data analysis. Multivariate data analysis is a tool that is widely used in chemo metrics and food science where the aim is to find the value of a latent variable (e.g. taste) by measuring some fixed variables (e.g. sugar, milk, cocoa) [14]. For the field of video quality assessment this translates to measure the latent variable video quality by measuring fixed variables (or features) such as blocking, blur, activity, continuity or noise. Features selected for the proposed model are presented in the following section.

3.2 Feature Selection

A set of simple no reference feature measurements was selected representing the most common kind of distortions namely blocking, blurriness and noise. One feature measurement was added to measure the amount of detail present in the encoded video. To take into account the time dimension of video four different continuity measurements were performed: predictability (shows how good one frame can be predicted using the previous frame only), motion continuity (measurement for the smoothness of the motion), color continuity (shows how much color changes between two successive images) and edge continuity (shows how much edge regions are changing between two successive images).

- Blur: the blur measurement used is described in [13]. The algorithm measures the width of an edge and then calculates blur by assuming that blur is reflected by wide edges. As blur is something natural in a fast moving sequence this measurement is adjusted if the video contains a high amount of fast motion.
- Blocking: for measuring the blockiness the algorithm introduced in [12] is used. This algorithm calculates the blockiness by applying a FFT along each line or column. The unwanted blockiness can be easily detected by the location in the spectra.
- Noise: to detect the noise present in the video a very simple noise detector was designed. First a prediction of the actual image is built by motion compensation using a simple block matching algorithm. Second a difference image between the actual image and its prediction is calculated and a low pass version of this difference image is produced by first applying a median filter and a Gaussian low pass filter afterward. A pixel is classified to contain noise if the difference value between the original difference image and the low pass difference image exceeds a threshold of 25 (assuming 8 bit values ranging from 0 to 255) for one of the three color planes. This noise detection algorithm is performed on all three channels of an RGB image and the algorithm returns the percentage of pixels that are classified to carry noise.
- Details: to measure the amount of details that are present in a video the percentage of turning points along each line and each row are calculated. This measurement is part of a BTFR metric included in [15]. As the amount of details that are noticed by an observer decreases with increasing motion the activity measurement is adjusted if high motion is detected in the video.
- Predictability: A predicted image is built by motion compensation using a simple block matching algorithm. The actual image and its prediction are then compared block by block. A 8×8 block is considered to be noticeable different if the sum of absolute differences in this block (SAD) exceeds 384. To avoid that single pixels dominate the SAD measurement both images are filtered using first a Gaussian blur filter and a median filtering afterward.
- Edge Continuity: The actual image and its motion compensated prediction are compared using the Edge-PSNR algorithm as described in [4].
- Motion Continuity: Two motion vector fields are calculated: between the current and the previous frame and between the following and the current frame. The percentage of motion vectors where the difference between the two corresponding motion vectors exceeds 5 pixels (either in x- or y-direction) determines the motion continuity.
- Color continuity: A color histogram with 51 bins for each RGB channel is calculated for the actual image and its prediction. Color continuity is then given as the linear correlation between those two histograms.

All feature measurements are done for each frame of the video separately and the mean value of all frames are then used for further processing. The above selected measurements are just one example for a set of variables that are used for building such a model. The presented variables were used for their simplicity. Using more complex measurements for artifacts like noise or blur may result in even more accurate models as well as adding measurements for artifacts not considered here (e.g. ringing). For this case only no reference feature measurements are considered, including some feature measurements that require the original video a RR or FR metric could be built.

3.3 Multivariate Calibration

Multivariate calibration is the method of learning to interpret a number of k input sensory signals that contribute to a common output y. For the presented metric the input signals are the above mentioned feature measurements while the output would be the visual quality of the video. The data set used for calibration of the model consisted of four different standard video test sequences (Bus, Football, Harbour, Mobile) at CIF resolution that were encoded according to AVC/H.264 at three (Bus, Harbour) and seven (Football, Mobile) different bit rates ranging from 96 kbit/s to 1024 kbit/s and with a frame rate of 15 or 30 fps. Different encoder settings concerning the number of B-Frames that were inserted (zero to two B-Frames), or the I-Frame periodicity (only one I-Frame or periodic I-Frames) were used. For each of the l calibration sequences the selected feature values f_{mi} ($m \in \{1 \cdots k\}, i \in \{1 \cdots l\}$) were computed, for reference the $l \times k$ matrix containing the feature values is denoted as $\mathbf{F}$.

3.3.1 Correction of the Feature Measurements using MSC

As it is expected that the measured features are not free from multiplicative or additive effects (e.g. the measurement for noise may be correlated with and affected by the amount of details present in the video), multiplicative signal correction (MSC) is performed before starting the multivariate regression. MSC was originally developed to correct measurements in reflectance spectroscopy, but can also help in this context to remove multiplicative and additive effects between different objective features. The MSC corrected value of one feature m for one sequence i is calculated as following:

$$f'_{mi} = c + f_{mi} * d$$

The two variables c and d are obtained by simple linear regression of the feature values of the sequence i compared to the average of the feature values of all l calibration sequences. For a detailed description of MSC see chapter 7.4 in [14]. Consequently the matrix $\mathbf{F}$ becomes $\mathbf{F}'$ after MSC treatment.

3.3.2 Multivariate Regression using Partial Least Squares

The obtained feature values f'_{mi} are then used together with the corresponding subjective ratings y_i that form the column vector $\mathbf{y}$ to

built a regression model using the method of Partial Least Squares Regression (PLSR) which is an extension of the Principal Component Regression (PCR). PCR is a is a bilinear regression method that consists of a Principal Component Analysis (PCA) of $\mathbf{F}'$ into the matrix $\mathbf{T}$ that contains the PCs of $\mathbf{F}'$ followed by a regression of $\mathbf{y}$ on $\mathbf{T}$. For the PLSR the modeling of $\mathbf{F}'$ and $\mathbf{y}$ is done simultaneously to ensure that the Principal Components (PC) gained from $\mathbf{F}'$ are relevant for $\mathbf{y}$.

$\mathbf{F}'$ can be modeled as:

$$\mathbf{F}' = \mathbf{1} \bullet \bar{\mathbf{f}} + \mathbf{T} \bullet \mathbf{P}^{\mathbf{T}} + \mathbf{E_f}.$$

With $\mathbf{P}$ being the loadings of the k input features, $\mathbf{T}$ being the scores of the l input sequences. $\bar{\mathbf{f}}$ represents the row vector of the mean values of the features and $\mathbf{E_f}$ is the error in $\mathbf{F}'$ that can not be modeled.

Likewise $\mathbf{y}$ can be modeled as:

$$\mathbf{y} = \mathbf{1} \bullet \bar{y} + \mathbf{T} \bullet \mathbf{Q}^{\mathbf{T}} + \mathbf{E_y}.$$

The prediction $\hat{y}$ for sequence i can then be modeled as:

$$\hat{y}_i = b_0 + \mathbf{f}'_\mathbf{i} * \mathbf{b}$$

$\mathbf{b}$ is the column vector of the single estimation weights b_m, b_0 is the model offset. A detailed description of PLSR can be found in chapter 3.5 of [14]. The process of building the quality model is shown in figure 1.

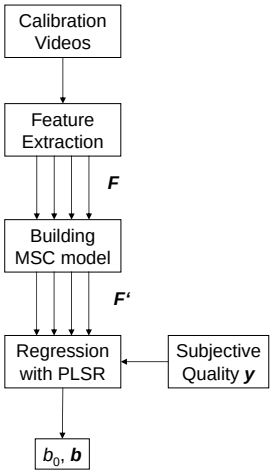

Figure 1: Quality model building process

3.4 Prediction Correction using Additional Quality Information

The NR quality metric gained by the previous steps faces the problem that even the original video may contain a certain amount of blur, blocking or noise and different sequences also have different motion properties. For this reason the overall prediction accuracy of the so far described model is low. But plotting the predicted quality against the quality measured in subjective tests reveals that the prediction accuracy for each single sequence is very high: the data points for one sequence lie on one straight line only with unknown slope s and unknown offset o. The overall prediction accuracy therefore can be improved by estimating slope and offset of these lines by calculating the predicted quality of the original video ($\hat{y}_{orig}$) and of a low quality version of the video ($\hat{y}_{low}$) using the same NR quality metric. While the original video is available and the subjective visual quality of this original is inherently given to be 1 on a 0 to 1 scale with a comparably small error only, an estimation of a low quality video can be produced by e.g. encoding the original with a low bit rate. Apparently the subjective visual quality of this low quality video can only be guessed (here set to 0.25). The possibility of improving the prediction accuracy of the model obviously depends on the accuracy of the estimated low quality video.

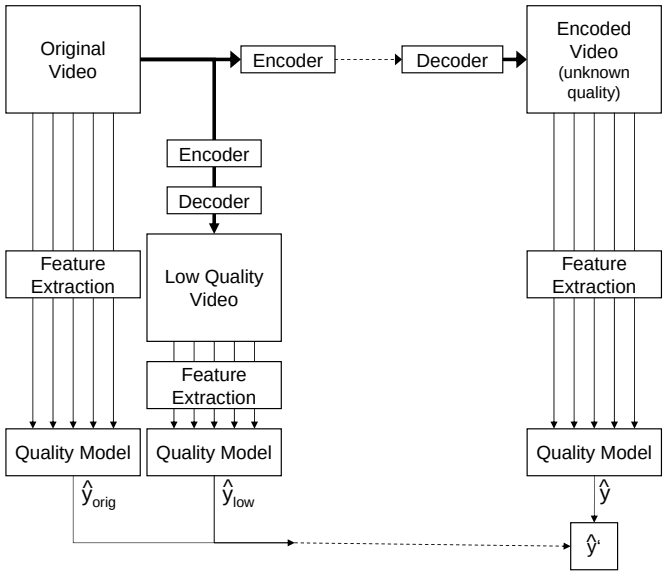

Figure 2: Prediction Model

Including the predicted quality of the original video and the predicted quality of the low quality video, the NR model will become a RR model, even if the additional data that has to be send is very low (not more than two values per sequence). The final prediction $\hat{y}'_i$ is then calculated as

$$\hat{y}'_i = \frac{\hat{y}_i - o}{s} \text{ with } s = \frac{\hat{y}_{orig} - \hat{y}_{low}}{1.0 - 0.25} \text{ and } o = \hat{y}_{low} - 0.25 * s\ .$$

Figure 2 gives an overview over the presented prediction model.

4. A REDUCED REFERENCE METRIC FOR AVC/H.264 ENCODED VIDEO

4.1 Subjective Testing

A reduced reference metric using the above described method was built using data from two subjective tests that included AVC/H.264 encoded video. Tests were done on video encoded at CIF resolution and were performed according to the rules given in ITU-R BT-500 [16]. This especially includes:

- Room setup compliant to ITU-R BT-500
- SSIS (Single Stimulus Impairment Scale) evaluation using a discrete impairment scale ranging from 0 to 10 (later rescaled to 0 to 1)
- All test sequences were evaluated by at least 20 naive viewers (students who were not familiar with video coding or video quality evaluation), all screened for visual accuracy and color blindness
- To minimize the contextual effect, which is known to affect results in a single stimulus environment, every encoded sequence was shown twice in the test
- Each test was preceded by an extensive training session to train the subjects on the task of evaluating the video
- Each single test session did not last longer than 25 minutes and an adaptation phase of five sequences was set at the start of each test session (this was not disclosed to the subjects).

The 95% confidence intervals for the subjective ratings were below 0.04 on a 0 to 1 scale, which shows, that the results from the tests are very reliable. Before building the model the data from those tests was split into two parts: only four out of 13 different

Table 1: Weights of objective features

Feature	Weight b
Activity	0.036
Blocking	-0.120
Blur	-0.109
Color Continuity	0.054
Edge Continuity	-0.072
Motion Continuity	0.090
Predictability	0.095
b_0	4.073

sequences were used for calibration of the metric, while the other nine sequences were used for the verification phase.

4.2 The Regression Model

After applying a MSC on the calibration data, a very simple regression model with only one PC can be built by applying a PLSR. The resulting weights b_m of the objective features and the model offset b_0 are given in Table 1. The PLSR on the matrix $\mathbf{F}'$ revealed that the feature 'noise' does not have an influence on the model (the weight for noise would be below 0.005), therefore this feature was removed and only the remaining seven features were taken into account.

4.3 Correcting the Results of the Model

The low quality video needed for the correction step as described in section 3.4 was constructed by encoding the video using the AVC/H.264 reference encoder with a high (fixed) quantization parameter (resulting in low quality). It has to be noted, that not only the coding parameters for producing this low quality video differ quite significant from those used to encode the videos under test, but also a different encoder has been used for this task.

5. RESULTS

Beside PSNR two other FR metrics were calculated for the presented data. The Edge-PSNR metric [4] was chosen as one representative of the methods standardized in ITU-T J.144 [15]. The second FR metric chosen for comparison is the SSIM as presented by Wang in [6].

5.1 Performance Metrics

The metrics that are most often used to measure the performance of an objective quality metric are the Pearson correlation, the Spearman rank order correlation and the outlier ratio. The Pearson correlation (1) gives an indication about the prediction accuracy of the model. A similar task is solved by the Spearman rank order correlation (2). This rank order correlation gives an indication how much the ranking between the sequences under test changes for the model's values compared to the subjective values (prediction monotonicity).

$$r^p = \frac{\sum_k (q_k - \overline{q})(MOS_k - \overline{MOS})}{\sqrt{\sum_k (q_k - \overline{q})^2}\sqrt{\sum_k (MOS_k - \overline{MOS})^2}} \quad (1)$$

Here q_k is the predicted value for the video under test and $\overline{q}$ is the mean value of all predictions. MOS_k and $\overline{MOS}$ are the respective subjective values. For the Spearman rank order correlation $r^s \chi_k$ is the rank of q_k and γ_k is the rank of the respective subjective value MOS_k. $\overline{\chi}$ and $\overline{\gamma}$ are the respective midranks.

$$r^s = \frac{\sum_k (\chi_k - \overline{\chi})(\gamma_k - \overline{\gamma})}{\sqrt{\sum_k (\chi_k - \overline{\chi})^2}\sqrt{\sum_k (\gamma_k - \overline{\gamma})^2}} \quad (2)$$

A data point is considered to be an outlier if the difference between measured and predicted quality is higher than 0.05 on a 0 to 1 scale. Note that for the outlier ratio no data fitting was applied for the proposed method while linear fitting was applied for the other three metrics. Linear data fitting has been chosen to fit the predicted values to the actual given data. While higher order fitting is sometimes proposed for this purpose, higher order fitting always carries the danger of fitting the model too much to the actual data and possibly jeopardizing the ability to predict unknown data.

Table 2: Comparison of objective metrics

Model	Pearson Linear Correlation	Spearman Rank Oder Correlation	Outlier Ratio
Proposed RR	0.844	0.799	0.582
PSNR	0.690	0.623	0.833
Edge-PSNR	0.802	0.745	0.833
SSIM	0.763	0.623	0.667

Table 3: Regression line before data fitting

Model	Slope	Offset
Proposed RR	0.969	0.042
PSNR*	0.625	0.183
Edge-PSNR	0.353	0.348
SSIM	0.173	0.821

For easier comparison PSNR was rescaled to PSNR* = (PSNR-15)/30

In addition the slope and offset of the linear regression line before linear data fitting are given in Table 3. This shows how much the model relies on a final fitting stage (an information, that is not given by the correlation measurements) and the ability to finally provide a correct and meaningful quality measurement as without the knowledge of that line no meaningful prediction can be made. For a perfect model the slope of this regression line would be 1.0 with 0 offset.

5.2 Verification

The gained model and the comparison models were compared on the basis of a dataset consisting of nine different video sequences coded at bit rates ranging from 96 kbit/s to 1024 kbit/s. This resulted in a total of 36 data points. The standard test sequences that were used were: City, Crew, Deadline, Foreman, Husky, Ice, Paris, Tempete and Zoom. Detailed results of each metric are given in figures 3 to 6 showing the predicted quality plotted versus the actual visual quality as measured in the subjective test. Error bars show the allowed variation of 0.05 in addition to the calculated values.

6. CONCLUSION

A reduced reference quality metric for AVC/H.264 was built using methods provided by multivariate data analysis. The metric was validated using results from careful conducted subjective tests and no sequence used for calibration of the model was included in the verification phase. The gained metric was verified using a wide variety of different video sequences including sports (Husky) conversational sequences (Deadline, Paris) or news (City, Crew) and results show that this metric produces stable results for sequences different from those that were used to gain the weights. The proposed RR metric provides a slightly higher prediction accuracy compared to two well known FR metrics and clearly outperforms PSNR. In addition the gained model allows a quality prediction by transmitting only two additional values, while most other reduced reference metrics need a much higher amount of additional data to be transmitted. One additional advantage of the proposed metric is, that no final data fitting step is needed, but a 1:1 relationship between the output of the metric and visual quality is given.

It is expected that using more complex features and the inclusion of features that were not regarded for the presented model will result in even more accurate quality models.

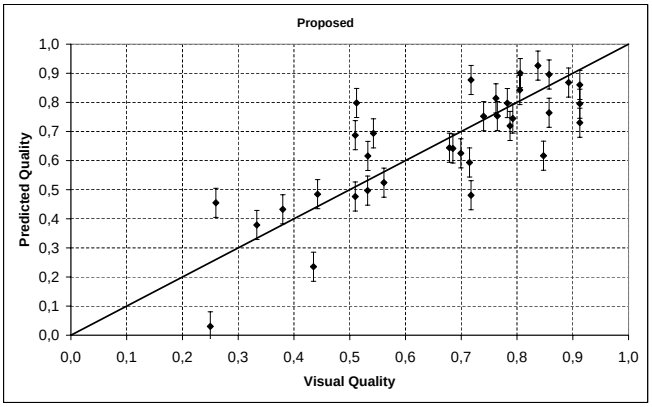

Figure 3: Proposed RR Metric - no data fitting

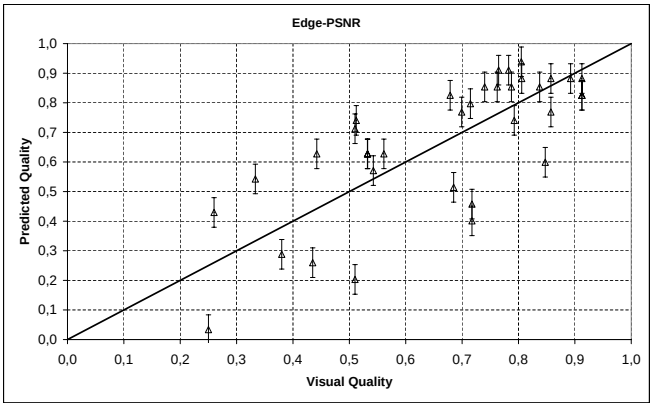

Figure 5: Edge-PSNR - linear data fitting

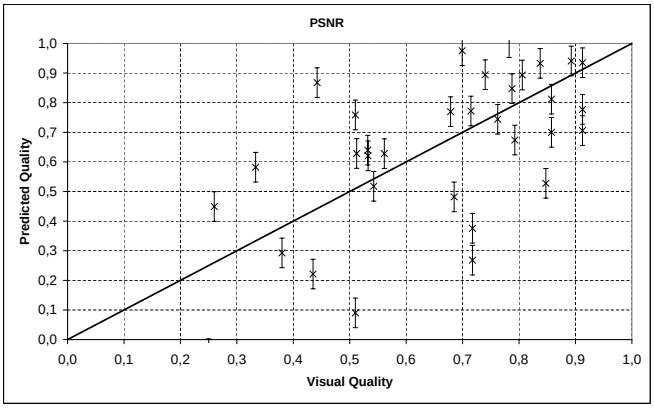

Figure 4: PSNR - linear data fitting

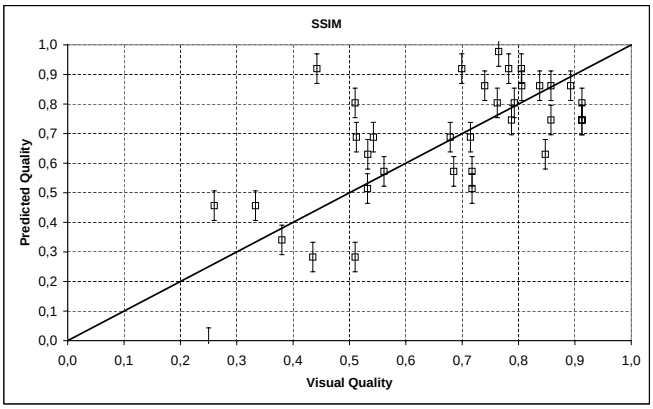

Figure 6: SSIM - linear data fitting

REFERENCES

[1] *ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG4-AVC), Advanced Video Coding for Generic Audiovisual Services*, ITU, ISO Std., Rev. 4, Jul. 2005.

[2] U. Benzler and M. Wien, "Results of SVC CE3 (quality evaluation)," ISO/IEC JTC1/SC29/WG11, Tech. Rep. M10931, Jul. 2004.

[3] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective assessment of MPEG-2 video quality," *Journal of Electronic Imaging*, vol. 11, pp. 365–374, Jul. 2002.

[4] C. Lee, S. Cho, J. Choe, T. Jeong, W. Ahn, and E. Lee, "Objective video quality assessment," *SPIE Optical Engineering*, vol. 45, p. 7004, Jan. 2006.

[5] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[6] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[7] Z. Wang, A. Bovik, and E. Simoncelli, *Structural Approaches to Image Quality Assessment.* Academic Press, 2005, ch. 8.3.

[8] Y. Fu-zheng, W. Xin-dai, C. Yi-lin, and W. Shuai, "A no-reference video quality assessment method based on digital watermark," in *Proc. 14th IEEE Personal, Indoor and Mobile Radio Communications 2003*, vol. 3, Sep. 2003, pp. 2707–2710.

[9] M. Carli, M. Farais, E. Drelie Gelasca, R. Tedesco, and A. Neri, "Quality assessment using data hiding on perceptually important areas," in *Proc. IEEE International Conference on Image Processing 2005, (ICIP2005)*, vol. 3, Sep. 2005, pp. 1200–3.

[10] Z. Wang, G. Wu, H. R.Sheikh, E. P.Simoncelli, E. Yang, and A. C.Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[11] P. L. Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316 – 1327, Sep. 2006.

[12] A. C. Bovik and S. Liu, "DCT-domain blind measurement of blocking artifacts in DCT-coded images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 1139–1149, Dec. 2002.

[13] F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 3, Sep. 2002, pp. 57–60.

[14] H. Martens and T. Naes, *Mutltivariate Calibration*. Wiley & Sons, 1992.

[15] *ITU-T J.144. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, ITU-T Std., Mar. 2004.

[16] *Question ITU-R 211/11 ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 11, 2002.