A FRAMEWORK FOR QUALITY PREDICTION OF VIDEO WITH UNKNOWN CONTENT

Silvio Borer *Jens Berger*

SwissQual AG
4528 Zuchwil, Switzerland
www.swissqual.com
silvio.borer@swissqual.com

ABSTRACT

In this paper a framework for objective quality prediction of video with unknown content is proposed. The framework consists of a general model for quality prediction together with a strategy to find optimal model parameters. The core of this strategy is a new error measure taking into account important requirements in the context of quality prediction. It is explained why standard least-square methods for model parameter estimation are not suitable.
An application to the estimation of quality of video sequences with different compression and transmission errors shows that the optimisation strategy overcomes the shortcomings of least-square methods.

Index Terms— video signal processing, optimization methods, quality control

1. INTRODUCTION

For many applications quality estimation of video sequences is indispensable. But the amount of data is often far too big to test quality by subjective viewing. In addition, operators testing the quality of their network often do not have control of the content of the transmitted video, e.g. in television broadcasting. Therefore, quality has to be estimated without knowledge of the video content. Todays quality estimation methods usually perform this task by detecting known and expected *degradations*, [1]. Types of degradations are for example blockiness as a result of a compression with a block-based codec. Codecs like ITU-T H.264 try to hide blockiness degradations by applying a smoothing filter on the decoder side, which can result in a blurriness degradation instead. There are degradations of temporal type, for example jerkiness, resulting from delays in the transmission of packets over the network, or frame rate reductions.

Recall some basic notions: Overall video quality is measured using a *mean opinion score (MOS)*. In a subjective test, subjects view video sequences sequentially, and rate their quality usually between $[1, 5]$, where $1 \equiv bad$, $2 \equiv poor$, $3 \equiv average$, $4 \equiv good$, $5 \equiv excellent$, see [2]. For each video sequence its MOS is the average rating given by the subjects. The goal is to estimate the MOS, the estimate is called the *predicted score*.

The paper is organised as follows: The next section explains the framework. In section 3 estimation of model parameters is discussed. Finally, after a toy example in section 4 explaining the ideas, the framework is applied to the prediction of video quality in section 5.

2. FRAMEWORK

To predict the score of a video sequence v a set of real valued functions $(f_k)_{k=1,..,m}$ is used, the set of *features*. Let $x_k = f_k(v)$ denote the *measured value* of feature f_k in sequence v. Example of a feature: count the number of times the difference between two subsequent frames is 0, i.e. count the number of repeated frames.
The predicted score s_{pred} is a function Φ of the feature values, called the *model*,

$$s_{pred}(v) = \Phi(f_1(v), ..., f_m(v)).$$

An example of the model Φ is a linear function of the measured feature values x_k.

Realistic assumption:

1. *The extracted features $f_1, ..., f_m$ are content dependent.*
2. *They do not allow to detect all types of degradations.*

Assumption 1, content dependency, means that two video sequences of different content showing the same type of degradation at the same perceived strength do not have the same measured feature values. As an example think of a feature detecting blockiness. Two video sequences compressed with the same codec could have the same visually perceived blockiness, but the measured value of the blockiness feature would be different.

An example of assumption 2 is the following: Suppose a video sequence is encoded and transmitted over a network. During transmission some parts of a frame are lost. The decoder has different error concealment strategies. One could

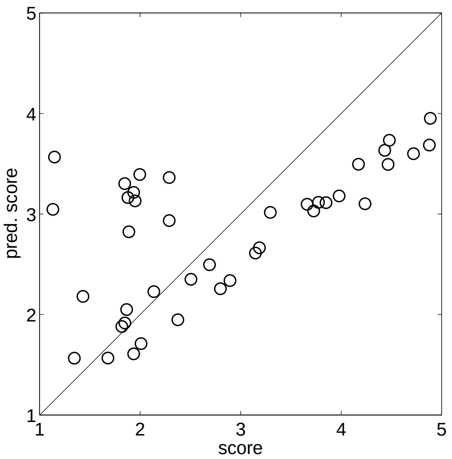

Fig. 1. Scatterplot of the target values of the toy data against the estimated values (○). Estimation is performed using a least square procedure. The line shows the identity function.

be to freeze the previous frames during the loss period, another could be to fill up only the lost parts of the frame with neighbouring information. The two strategies will result in very different types of degradations. One of the features could be tuned to detect packet loss by detecting freezing periods. Thus, the second type of degradation could not be detected.

Despite these assumptions, scores should be predicted such that the following conditions are met:

Conditions:

1. *Good overall/average prediction.*
2. *High quality sequences should be above 4 (≡ good)*
3. *Possibility for confident interpretation of single sample prediction.*
4. *Possibility for future extension of the model by adding features.*

The standard way to proceed is the following: Choose a model Φ depending on parameters and use a least-square procedure to estimate model parameters. Then, condition 1 could be met. But under assumptions 1,2 the conditions 2, 3 will probably fail to be met. The reason is best seen in the toy example in section 4 below, and the corresponding figure 1.

Note that poor predictions resulting from assumption 1 can be avoided by considering *condition averages*, by predicting the average MOS of all video sequences of the same condition.

3. MODEL ESTIMATION

This section describes how model parameters can be estimated, such that the conditions 1-4 are met. The idea is to define a suitable error measure, often called *loss function*. We propose a loss function of the form

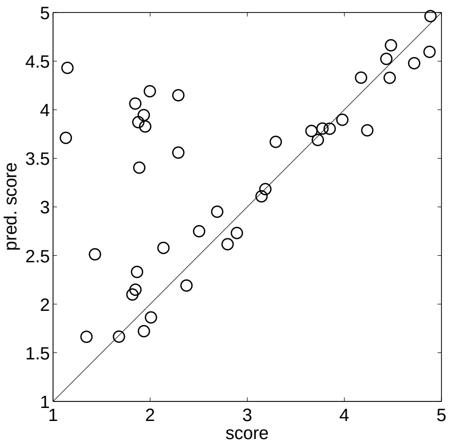

Fig. 2. Scatterplot of the target values of the toy data against the estimated values (○). Estimation is performed using a non-symmetric ϵ-insensitive loss function, compare with figure 1. The line shows the identity function.

$$F_{loss}(x) = \begin{cases} \frac{1}{\epsilon^2}(x-\epsilon)^2 & (x > \epsilon) \\ -(x+\epsilon) & (x < -\epsilon) \\ 0 & otherwise, \end{cases} \qquad (1)$$

where the parameter ϵ is a small constant. The loss function is non-symmetric and ϵ-*insensitive*, see [3] for extensive use of ϵ-insensitive loss functions for regression.

Suppose the model Φ depends on parameters $c = (c_1, ..., c_p)$. Optimise the parameters $c_1, ..., c_p$ by minimising the empirical error, i.e. the total loss of the data samples (v_i),

$$\text{minimise} \quad \sum_{i=1}^{n} F_{loss}(y_i - s_{pred,c}(v_i)), \qquad (2)$$

where y_i is the target value, the MOS, and $s_{pred,c}(v_i)$ is the predicted score. Note that the index c of the predicted score expresses the dependency on the parameters c. As the loss function is ϵ-insensitive, prediction errors smaller than ϵ do not contribute to the empirical error. Furthermore, the loss function is quadratic for values larger than ϵ, but linear for values smaller than $-\epsilon$. Therefore, large underestimates contribute stronger to the empirical error than large overestimates.

Note that care has to be taken by choosing the model Φ as minimising the empirical error might lead to overfitting [3].

The optimisation is performed using a stochastic gradient descent: At each iteration t there is an estimate c_t of the optimal parameters. Then, a sample video v_t is drawn, and based on this sample the empirical error is estimated as

$$E_{emp,t} = F_{loss}(y_t - s_{pred,c_t}(v_t)),$$

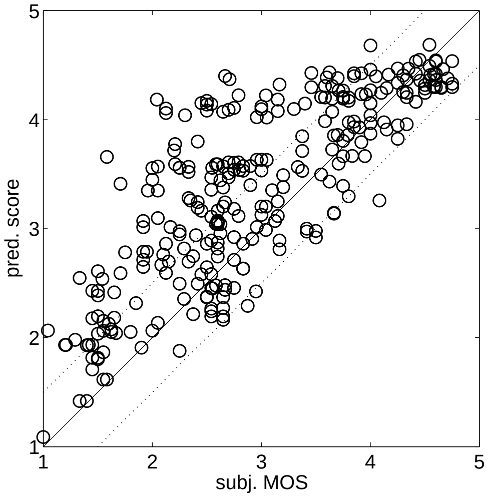

Fig. 3. Scatterplot of the target values against the estimated values (○). Estimation is performed using a a non-symmetric ϵ-insensitive loss function. The line shows the identity function, the dotted lines sit at a distance of ϵ.

where y_t denotes the MOS of sample v_t. Update the estimate of c_t by gradient descent

$$c_{t+1} \longleftarrow c_t - \eta \partial_c(E_{emp,t}),$$

where the constant $\eta << 1$ is the learning rate. Iterate these steps until convergence. The reader is referred to [4] for more details.

4. TOY EXAMPLE

In this section the framework is applied to toy data as an illustration. Consider the following toy example: the model is a linear prediction of the score based on one feature f. MOS values (y_i) and measured values (x_i) corresponding to the feature f are generated artificially in the following way: Suppose there are two types of degradations. First, one that is detected by feature f and second, one that is not.

For the first type of degradation the MOS values (y_i) are sampled from a uniform distribution in $[1, 5]$. The measured values (x_i) equal to the target values up to Gaussian noise σ, $x_i = y_i + \sigma$.

For the second type of degradation the MOS values (y_i) are sampled from a normal distribution with mean 1.7. The measured values are given by normal samples around the value 4. The interpretation is that the feature can almost not detect this type of degradation.

First, for comparison, model parameters are estimated using a least-square fit. The result is shown in figure 1. Second, the model parameters are estimated by using loss function (1), and minimising the empirical error (2). The result is shown in figure 2.

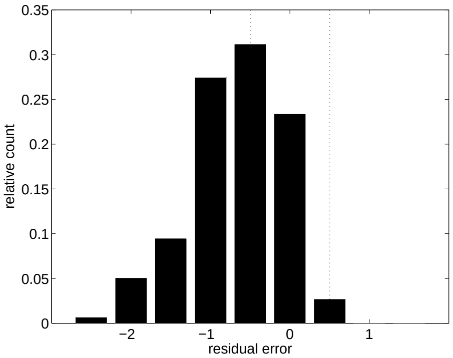

Fig. 4. Score prediction: distribution of residual errors of condition averages. The dotted lines show the value of the insensitivity parameter ϵ.

The main differences between the two results are, first, the estimates using least-square fit span a considerably smaller range. Samples with a MOS score above 4 fall between 3 and 4. On the other hand, our proposed method produces estimates, which are close to the target value for samples belonging to the first type of degradation. Samples with a high target value have a high estimated value. Second, with the proposed method large residuals are always overestimates, contrary to the least-square fit. Therefore, looking at the estimated score of a single sample there is a high confidence that the MOS value is not much larger than the estimate.

5. VIDEO QUALITY ESTIMATION

In this section the framework is applied to the prediction of the quality of video sequences in the following database:

5.1. Database of video sequences

A database consisting of 15 source video sequences of QCIF format of 10 seconds length, and 281 processed samples is used. Different *conditions* were defined to generate from the 15 source sequences the processed sequences. The conditions were freezing, performing frame rate reduction, encoding/decoding the sequences with either of a h263, h263+, h264 encoder/decoder, by blurring the frames, by simulating packet loss, or by a combination of these. For each condition at least 5 samples of different content were generated. All the video sequences were rated by 12 experts and averaged to yield the mean opinion scores (MOS).

5.2. Features and result

There are two features: one is a temporal feature measuring the jerkiness of a video sequence. Here, freezing is interpreted as a strong jerkiness. It is mainly a temporal feature.

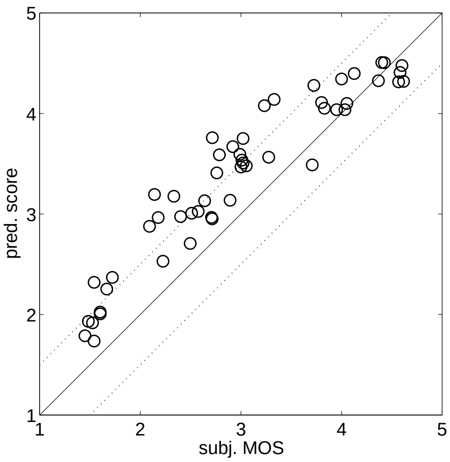

Fig. 5. Scatterplot of the target values against the estimated values (○) of the condition averages of the same model as in figure 3. The line shows the identity function, the dotted lines sit at a distance of ϵ.

The second feature is mainly spatial, composed of a blockiness, a blurriness, and a packet loss measure. The features are scaled to take values in $[0, 1]$.

The model is a multiplicative model of the form

$$\Phi(x_1, x_2) = \tau(x_1^{c_1} \cdot x_2^{c_2}), \tag{3}$$

where c_1, c_2 are the model parameters, and τ rescales the interval $[0, 1]$ linearly to the interval $[1, 5]$. Hence score prediction takes the form

$$s_{pred}(v) = \tau\left(f_1(v)^{c_1} \cdot f_2(v)^{c_2}\right). \tag{4}$$

A value of $\epsilon = 0.5$ is chosen for the insensitive region of the loss function (1). The estimation is performed in the log-domain, where the model is linear. The estimated parameters are $c_1 = 0.39$, and $c_2 = 0.97$. Figure 3 shows a scatterplot of the mean opinion scores against the estimated scores. Recall the desired conditions 1-4 of section 2. The correlation coefficient is 0.8, the model has a good overall performance, keeping in mind that the model has only two parameters. For condition averages, the correlation coefficient increases to 0.95. The samples with MOS above 4 have a predicted score above 4, too, in accordance with condition 2. There are very few estimates lower than ϵ compared to the MOS value. Hence, for each sample there is high confidence that the predicted score is not more than ϵ below the MOS value, as desired by condition 3. The last condition 4, extensibility, is discussed in section 6.

Figure 4 shows the residual error distribution. Note that the mean of the estimated scores is not equal to the mean of the MOS scores.

Figure 5 shows the scatterplot of MOS values against predicted scores for condition averages. As a result of the strategy to avoid estimates much smaller than the MOS value, samples with low MOS value are systematically somewhat too high. Figure 6 shows the distribution of residuals. Residual errors are considerably smaller than in figure 4.

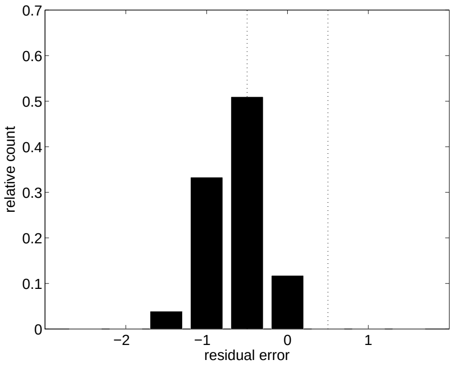

Fig. 6. Score prediction: distribution of residual errors of condition averages. The dotted lines show the value of the insensitivity parameter ϵ.

6. DISCUSSION

In this paper a framework for quality estimation of video with unknown content is proposed. The advantage of this concept with respect to the initial assumptions and conditions is demonstrated on real world data. Of course, the framework itself could be used for other similar problems like quality estimation of video with known content, prediction of audio or audio-visual quality.

One possible advantage of our framework, which is not tested in the current paper is extensibility. Suppose there would be a type of degradation, not detected by our features. Corresponding samples would be largely overestimated. If an additional feature could be designed detecting only this type of degradation, the model could be extended to include the additional feature. Ideally, the extended model would give similar predictions except for the before undetected type of degraded samples, where performance could be improved. Testing this idea is an open point for future work.

7. REFERENCES

[1] Stefan Winkler, “Video quality and beyond,” in *Proceedings of the EUSIPCO*, Poznan, Poland, September 2007.

[2] ITU-T, “Subjective video quality assessment methods for multimedia applications,” *ITU-T Recommendation P.910*, 1999.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 1995.

[4] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge MA, 2002.