Video Quality Estimation for Mobile Streaming Applications with Neuronal Networks

Michal Ries, Jan Kubanek and Markus Rupp
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstasse. 25, A-1040 Vienna, Austria
(mries, mrupp)@nt.tuwien.ac.at

***Abstract*— The provision of mobile multimedia streaming applications becomes essential for emerging 3G networks. The crucial point of successful deployment of multimedia mobile services is the user satisfaction level, since the perceptual video quality for such low bit rates, frame rates and resolutions is limited. Depending on the content character of a video sequence, the compression and network settings, maximizing the subjective perceptual quality also differs. The complexity of quality estimation and maximizing perceptual quality for mobile streaming application is still high if only the most significant influence factors are taken into account. Aim of this work is to design an artificial neural network with low complexity for the estimation of visual perceptual quality, based on a combination of a possibly small set of the most important objective parameters (compression settings and content features). To achieve this, the neural network was trained with a set of objective and subjective parameters, obtained by an extensive survey. Moreover, estimations with neural networks do not require any knowledge about the original sequence. The achieved correlation with the data set is as good as if the more the complex human vision based estimation is applied.**

I. Introduction

The emergence of third-generation (3G) mobile networks offers new opportunities for the effective delivery of data with rich content, including multimedia messaging, video-streaming and video interactive service. The third generation of mobile systems enables a more convenient use of multimedia messaging and video-services by offering higher bandwidth and lower latency than its GSM and GPRS predecessors. In UMTS services, it is essential to provide required levels of customer satisfaction or equivalently provide required perceived video stream quality. The aim of our work is the prediction of perceptual quality for low resolution and low rate video streaming. The focus is to estimate the quality of mobile multimedia at the user-level and to find optimal codec settings for 3G streaming scenarios. To select optimal codec parameters, it is important to consider corresponding quality requirements based on human perception [1], [2]. The complexity of the task to determine perceived quality is rather high, because the human vision system is only a partly explored area [3]. There are three possible ways of perceptual quality estimation. The first method is to perform an extensive survey on selected test group of persons. The handicap of this approach is, that subjective evaluation is very expensive and time consuming. In this case, it cannot be realized very often. The second one is the evaluation with video quality metrics, based on objective parameters. It is problematic, due to its main drawbacks: generally objective parameters are decorrelated with subjective quality results [4], the original metrics require original sequences and significantly higher calculation effort [5], [6], and also most of the proposed metrics do not take into account network parameters, as well as content character [7], [5], [4], [8]. The third approach is using a neural networks [9], [10]. This methodology exploits the advantages of subjective and objective evaluation. The idea is to train the network by the set of input variables, i. e. objective video parameters, which mostly affect the quality. After the training the network to behave like a "real" human evaluating the video streams. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the biological nervous systems, such as the brain processes information [11], [12]. It is composed of a large number of highly interconnected processing elements (neurons), working in unison to solve specific problems. ANNs, like people, learn by example. The method consists of a procedure allowing to build a function $MOS = f(x_1; ...; x_n)$ where $x_1; ...; x_n$ are source and network parameters and MOS is a measure of quality (Mean Opinion Score). If the stream encounters source and network conditions such that the chosen parameters have values $x_1; ...; x_n$, then $f(x_1; ...; x_n)$ will be very close to the quality value an average human would rate it.

The paper is organized as follows: In Section 2 the sequences selected for evaluation are described as well as the setup of our survey, we performed to obtain the MOS values. Section 3 describes ANN design, training methods and generalization. The ANN performance on subjective video quality estimation and comparison with analytical video quality model describes Section 4. Section 5 contains conclusions and some final remarks.

The contribution of this paper is estimation of video quality for the most significant content types with utilization of ANN without original video sequences.

II. The test setup for video quality evaluation

For the tests we selected five video sequences each having ten-second duration and QCIF resolution. Screenshots of these sequences are depicted in Figure 1.

Two of them ("akiyo", "foreman") are well-known professional test sequences obtained by a static camera. In the "akiyo" sequence a female moderator is reading news only by moving her lips and eyes. The "akiyo" sequence represents the news scenario. The "foreman" sequence contains a monologue

Fig. 1. Screenshots of the video test sequences used in the survey: ”akiyo”, ”foreman”, ”soccer”, ”panorama” and ”traffic”.

of a man moving his head dynamically and at the end of the sequence there is a rapid scene change. The ”foreman” sequence is a typical scenario for video calls. ”Soccer” and ”panorama” are both sequences with permanent camera movement. ”Soccer” is a professional wide angle sequence; the entire picture is moving uniformly. Additionally the players and the ball are moving in a fast way. ”Panorama” is a non-professional sequence, containing smooth and relatively slow movement of the whole scene. This is a typical scenario for weather cameras, wide angle survivalence and for tourists guides. The ”traffic” sequence is obtained by a static traffic camera. The camera is static and slowly moving cars can be observed. Each of the tested sequences represent a typical content offered by network providers nowadays.

All sequences were encoded with H.263 profile 3 and level 10. For subjective quality testing we used combinations of bit rates and frame rates shown in Table I. In total, there were 60 encoded test sequences, of which we excluded six combinations where the resulting video quality was clearly insufficient.

Frame Rate [frames/s]	Bit Rate [kbit/s]
5	18
5	44
5	80
7,5	18
7,5	44
7,5	80
10	18
10	44
10	80
15	18
15	44
15	80

TABLE I

TESTED COMBINATIONS OF FRAME RATES AND BIT RATES.

To obtain MOS values, we worked with 38 paid test persons. The chosen group ranged different ages (between 17 and 30), sex, education and experience with image processing.
The tests were performed according to the ITU-T Recommendation [14], using absolute category rating (ACR) method as it better imitates the real world streaming scenario. Thus, the subjects had not the original sequence as a reference, so their evaluation suffers from higher variance. People evaluated the video quality using a five grade MOS scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent). According to our experiences with previous psycho-visual experiments [4], [7], [8] the subjective results are slightly different if they are displayed on the UMTS handset or PC monitors. To emulate real conditions of the UMTS service, all the sequences were displayed on a UMTS handset Sony Ericsson Z1010. The viewing distance from the phone was not fixed, but selected by the test person. We have noticed that all subjects were comfortable to take the phone at a distance of 20-30 cm. At the beginning of the test session, three training sequences were presented to the test persons. Test sequences were presented in an arbitrary order, with the additional condition that the same sequence (even differently degraded) did not appear in succession. Three runs of each test were taken. In order to avoid the learning effect we made a break of half an hour between the first and the second run, and a break of two weeks between the second and the third run. However, there were no really noticeable differences between the first two runs and the third run, performed two weeks after. In the further processing of our results we have rejected the sequences which were evaluated with individual standard deviation higher than one. Following this rule, we excluded only 2,23% of the tests results.

III. NEURAL NETWORK DESIGN

The aim is to design ANN for the video quality estimation. We trained our ANN with a set of objective and subjective video parameters. By the term ”objective parameters” we understand both - the compression parameters and the content characteristics. We have chosen the three objective parameters, bit rate (BR), frame rate (FR) and f_{SI13}, which mainly affect the perceptual quality according to [5], like input parameters of our network. BR and FR are the basic codec parameters. The third parameter f_{SI13} is defined by ANSI [6], is a reference-free measure of overall spatial information, denoted as since images were preprocessed using the 13×13 Sobel filter masks. It is calculated as the standard deviation over an S-T region of $R(i,j,t)$ samples, i and j being the coordinates within the picture displayed in time t. The result is clipped at the perceptibility threshold P [6]:

$$\mathrm{f_{SI13}} = \{\mathrm{std}_{space}[R(i,j,t)]\}|_P : i,j,t \in \{\text{S-T region}\}, \quad (1)$$

This feature is sensitive to changes in the overall amount of spatial activity within a given S-T region. For instance, localized blurring produce a reduction in the amount of spatial activity, whereas noise produces an increase of it. The output parameters of our ANN are the MOS values obtained by extensive survey.
In multi-layer networks, with any of a wide variety of continuous nonlinear hidden-layer activation functions, one hidden layer with an arbitrarily large number of units suffices for the ”universal approximation” property [11], [12]. According to this knowledge we designed the network with three layers - input, one hidden and output layer (Figure 2). But there is no theory yet to determine, how many hidden units are needed to approximate any given function. The best

number of hidden units depends in a complex way on: the numbers of input and output units, the number of training cases, the amount of noise in the targets, the complexity of the function or classification to be learned, the architecture, the type of hidden unit activation function, the training algorithm and the regularization [13]. In most situations, there is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each. If there are too few hidden units, high training errors and high generalization errors are obtained due to underfitting and high statistical bias. If there are too many hidden units, low training errors but still have high generalization errors are obtained due to overfitting and high variance [9]. The error on the training set is driven to a very small value, but when new data is presented to the network the error increases. The network has memorized the training examples, but it has not learned to generalize new situations. Assume for example, if there is one continuous input X that takes values on the interval (0, 100) and if there is one continuous target $Y = \sin(X)$. In order to obtain a good approximation to Y, about 20 to 25 hidden units with tangents-hyperbolic function are required. Although, one hidden unit with sine function would do the job [16]. A possible way how to improve generalization is to have a network that is just large enough to provide an adequate fit, because it can approximate a more complex function. If we use a sufficiently small network, it will not have enough power to overfit the data. But it is difficult to know beforehand how large a network should be for a specific application. Another way is to have much more points in a training data set than network parameters, avoiding chance of overfitting.

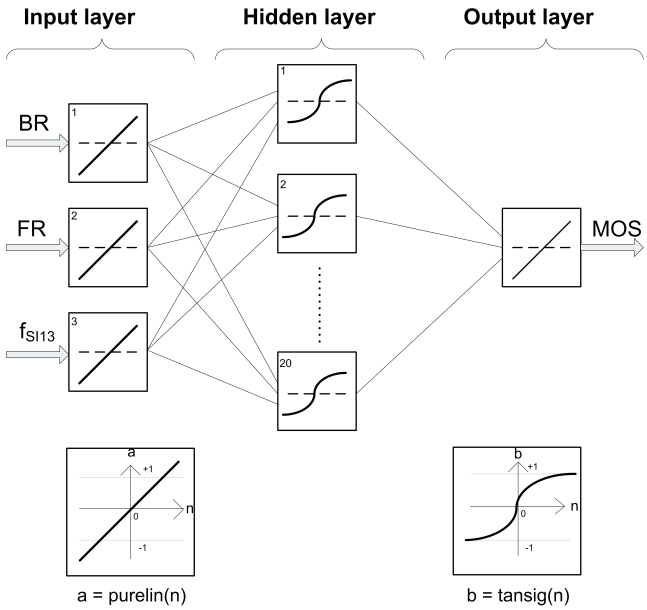

Fig. 2. Architecture of proposed three-layer feedforward ANN

A typical recommendation is that the number of weights should be not more than 1/30 of the number of training cases [16]. Such rules are only concerned with overfitting and are at best crude approximations. Also, these rules do not apply when regularization is used. It is true that without regularization, if the number of training cases is much larger (but no one knows exactly how much larger) than the number of weights, overfitting or underfitting appears more often. For a noise-free quantitative target variable, twice as many training cases as weights may be more than enough to avoid overfitting [17].
The lack of training data in our case requires to improve on the generalization. We tested a few training methods (Variable learning rate, Resilient backpropagation, Quasi-Newton algorithm), but generalization was insufficient. Finally we applied one of the methods improving the generalization. This method is called Automated regularization [21], that is a combination of Bayesian regularization [18], [19], and Levenberg-Marquardt training [22]. The weights and biases of the network are assumed to be random variables with specified distributions. The regularization parameters are related to the unknown variances associated with these distributions. One feature of this algorithm is that it provides a measure of how many network parameters (weights and biases) are being effectively used by the network. We scale inputs and targets so that they fall in the range [-1,1], because this algorithm generally works best when the network inputs and targets are scaled so that they fall approximately in the range. The outputs were converted back into the same units that were used for the original targets.
Once the network weights and biases have been initialized, the network is ready for training. During training the weights and biases of the network are iteratively adjusted to minimize the squared error between the network outputs and the target outputs.
Finally, we proceeded to design several ANNs, in order to find a trade off between ANN's minimal number of neurons in hidden layer and accuracy. We present the training data to the network by 54 vectors with dimension of four (see Figure 2, BR, FR, f_{SI13} and MOS), with three input values and one target value. Each vector consists of these values: BR, FR, f_{SI13} and MOS. We propose the three layered ANN architecture with three linear units in input layer and one linear unit in the output layer. The minimal training and generalization error was obtained for hidden layer which consists of 20 tangents-signoid neurons [23].

IV. Evaluation and performance of proposed ANN

We investigated and compared the network response in more details. The performance of a trained network can be measured to some extent by the errors on the training, validation and test sets.

A. Fitting performance of proposed ANN

We performed a linear regression analysis (see Figure 3) between the network response and the corresponding targets. This relationship between estimated ($\mathbf{y}$) and target data ($\mathbf{x}$) (or between predicted and subjective MOS) is represented in the form:

$$\mathbf{y} = m\mathbf{x} + b, \tag{2}$$

where m corresponds to the slope and b to the y-intercept of the best linear regression relating targets to network outputs. If

there is a perfect fit (outputs are exactly equal to targets), the slope would be 1, and the y-intercept would be 0. In our case we obtain results clearly close to optimum m =0,80215 and b=0,59462. Furthermore the correlation factor (Equation: (3)) between estimated and target data (or between predicted and subjective MOS) of the proposed ANN is 90,4%.

$$r = \frac{(\mathbf{x} - \overline{x})^T(\mathbf{y} - \overline{y})}{\sqrt{((\mathbf{x} - \overline{x})^T(\mathbf{x} - \overline{x}))((\mathbf{y} - \overline{y})^T(\mathbf{y} - \overline{y}))}}, \qquad (3)$$

These results are comparable with the performance of the analytical model for video quality estimation. The fitting performance of Reference-Free Video Quality Metrics (RFM) [5], for all content classes, is 96%. Alternative comparison between ANN and RFM allows us the mean square error (MSE). MSE (Equation: (4)) is an accepted measure of goodnes of fit. The performance of ANN and RFM is compared and summarized in Table II.

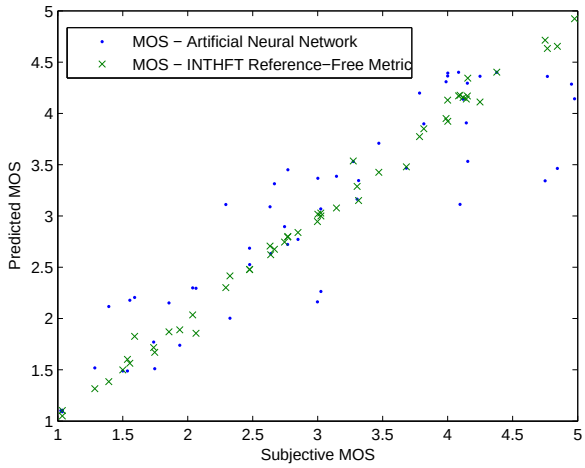

Fig. 3. ANN and RFM mapping on subjective data

We calculated MSE values for individual content classes and for all sequences. For the MSE and the correlation coefficient the vector $\mathbf{x}$ corresponds to **average** MOS values (averaged over test runs of all subjective evaluations for particular test sequence) for all tested encoded sequences. Vector $\mathbf{y}$ corresponds to the prediction made by the proposed ANN application and RFM metric. The dimension of $\mathbf{x}$ and $\mathbf{y}$ refers to N.

$$\text{MSE} = \tfrac{1}{N}(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y}), \qquad (4)$$

The performance by MSE of RFM is significantly better. The reason is following, the RFM metrics take strongly content character into account . The sequences are separated according to content character [5] as can be seen in Table I to five content classes and each class has its own coefficients. So the RFM metrics are not general but dedicated to certain content classes. On the other hand, the ANN model is general for all content classes and training performance is sufficient video quality estimation. Also we cannot expect better fitting performance of the general model compared to the content dependent model.

Sequence	ANN	RFM
Akiyo	0,66378	0,01149
Foreman	0,30840	0,00295
Soccer	0,01619	0,00213
Panorama	0,04551	0,00039
Traffic	0,08068	0,02450
All sequences	0,23278	0,00860

TABLE II
PERFORMANCE OF ANN AND RFM BY MSE

B. *Validation of proposed ANN*

Finally we build an application (see Figures 4 and 5) for video quality estimation, which is based on our ANN network.

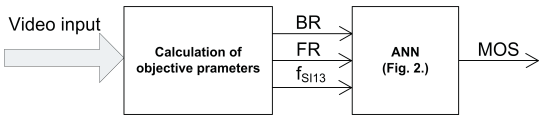

Fig. 4. Functional block of proposed application

Fig. 5. GUI of application for video quality estimation

We perform with this application further validation of our proposed ANN. The source material for the validation is composed of two test sequences (see Figure 6) with different sequence character. The test sequences were approximately ten-seconds long, in order to keep scene integrity. The first sequence was a soccer sequence shot with wide camera angle and rapid ball and players movement. The camera was continuously tracing the ball movement. This sequence was extremely dynamic and fast. The second sequence was a talk show by a well known TV celebrity. This was quite to the expectations almost static, the presenter was moving just slightly his head and hands. The sequences were encoded

Fig. 6. Validation sequences: soccer (left), talk show (right)

with the H.263 codec profile 3 and level 10. The encoder settings combinations are listed in Table I. The subjective tests were executed exactly in the same way as is described in Section 2. The only difference was, compared to our previous experiences, to use two test runs rather than three. The test group consists of 10 new test persons, without any previous experiences with subjective video quality evaluation. The amount of 10 test persons was sufficient for a validation purpose. The aim of this survey was to obtain completely new video quality survey results, which were absolutely independent from the training data set. Due to the independent subjective video quality results and the new test sequences we can validate our ANN application and confirm the right choice of objective parameters and ANN's design.

Sequence	MSE	Correlation Coef. (r)
Soccer	0,6585	0,9061
Talk show	0,8415	0,8608
Both sequences	0,7369	0,8896

TABLE III
VALIDATION RESULTS OF ANN APPLICATION BY MSE AND CORRELATION COEFFICIENT

The quality of ANN prediction on independent data set is characterized by MSE (Equation: (4)) and Correlation Coefficient (Equation: (3)), where the vector $\mathbf{x}$ corresponds to **average** MOS values (averaged over two runs of all 10 subjective evaluations for particular test sequence) for all tested encoded sequences. Vector $\mathbf{y}$ corresponds to the prediction made by the proposed ANN application. The results shown in Table III confirm, that the ANN application is a suitable video quality estimation for the most significant video streaming content types.

V. CONCLUSION

In this paper we design an application for video quality estimation of 3G video streaming services based on ANNs. Our result demonstrates that it is possible to predict the video quality for the 3G video streaming scenario with low complexity video parameters, if we choose parameters that most significantly influence the subjective quality according to a multivariate analysis [5]. We successfully applied this knowledge to the proposed ANN. Furthermore, our ANNs were able to generalize this for all investigated content types. Due to the generalization property we designed a universal application for video quality prediction.
The fitting performance of RFM is better, than that of ANNs, because the RFMs metrics were proposed for certain content types. On the other hand, our ANN is a universal video quality predictor, the lower ANN accuracy is the prize for its universality. The validation results for significantly different content types show, that our ANN takes also content character into account. Finally, our proposed ANN is a reliable prediction tool for video quality prediction of low resolution and low bit rate, frame rates encoded sequences, that a tool is proper for video streaming over 3G networks.

VI. ACKNOWLEDGMENT

The authors would like to thank mobilkom austria AG&Co KG for supporting their research. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG&Co KG.

REFERENCES

[1] M. H. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Transactions on broadcasting, Vol. 50, Issue: 3, pp 312-322, Sept, 2004.
[2] S. Winkler, F. Dufaux, "Video Quality Evaluation for Mobile Applications," Proc. of SPIE Conference on Visual Communications and Image Processing, Vol. 5150, pp. 593-603, Lugano, Switzerland, July 8-11, 2003.
[3] H. H. Bulthoff, S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," Proc. of the National Academy of Sciences of the USA, Vol. 89, Issue 1, pp. 60-64, 1992.
[4] O. Nemethova, M. Ries, E. Siffel, M. Rupp, "Quality Assessment for H.264 Coded Low-Rate and low-Resolution Video Sequences," Proc. of Conf. on Internet and Inf. Technologies (CIIT), pp. 136-140, St. Thomas, US Virgin Islands, 2004.
[5] M. Ries, O. Nemethova, M. Rupp. "Reference-Free Video Quality Metric for Mobile Streaming Applications," Proc. of the DSPCS 05 & WITSP 05, pp. 98-103, Sunshine Coast, Australia, December, 2005.
[6] ANSI T1.801.03, "American National Standard for Telecommunications - Digital transport of one-way video signals. Parameters for objective performance assessment," American National Standars Institute, 2003.
[7] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, M. Rupp, "Audiovisual Estimation for Mobile Streaming Services," Proc. of International Symposium on Wireless Communication Systems IEEE Ed., Siena, Italy, Sept, 2005.
[8] M. Ries, O. Nemethova, B. Badic, M. Rupp, "Assessment of H.264 Coded Panorama Sequences," Proc. of the First International Conference on Multimedia Services and Access Networks, Orlando, Florida, June, 2005.
[9] G. Rubino, M. Varela, "A new approach for the prediction of end-to-end performance of multimedia streams," Proc. of International Conference on Quantitative Evaluation of Systems (QEST'04), IEEE CS Press, University of Twente, Enschede, The Netherlands, September, 2004.
[10] S. Mohamed, G. Rubino, "A study of real-time packet video quality using random neural networks" IEEE Transactions On Circuits and Systems for Video Technology, 12(12):1071-1083, December, 2002
[11] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[12] K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, Vol. 2, pp. 359-366, 1989.
[13] F. Girosi, M. Jones, T. Poggio, "Regularization Theory and Neural Networks Architectures," Neural Computatio, Vol. 7, pp. 219-269, 1995
[14] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," September 1999.
[15] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," 2000, available at http://www.vqeg.org/.
[16] A. Weigend, "On overfitting and the effective number of hidden units," Proceedings of the 1993 Connectionist Models Summer School, pp. 335-342, 1993.
[17] S. Geman, E. Bienenstock, R. Doursat, "Neural Networks and the Bias/Variance Dilemma," Neural Computation, Vol. 4, pp. 1-58, 1992.
[18] R. M. Neal, Bayesian Learning for Neural Networks, New York: Springer-Verlag, 1996.
[19] J. M. Bernardo, A. F. M. Smith, Bayesian Theory, New York: John Wiley, 1994.
[20] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning representations by back-propagating errors," Nature, Vol. 323, pp. 533-536, 1986.
[21] F. D. Foresee, M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization," Proceedings of the 1997 International Joint Conference on Neural Networks, pp. 1930-1935, 1997.
[22] Hagan, M. T., and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, no. 6, pp. 989-993, 1994.
[23] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, D. L. Alkon, "Accelerating the convergence of the backpropagation method," Biological Cybernetics, Vol. 59, pp. 257-263, 1988.