Feature Selection for Neural-Network Based No-Reference Video Quality Assessment

Dubravko Ćulibrk, Dragan Kukolj, Petar Vasiljević, Maja Pokrić, and Vladimir Zlokolica [⋆]

Faculty of Technical Sciences, University of Novi Sad,
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
culibrk@iis.ns.ac.yu,dragan.kukolj@rt-rk.com,petarv@uns.ac.rs,
{maja.pokric,vladimir.zlokolica}@rt-rk.com
http://www.ftn.uns.ac.rs

Abstract. Design of algorithms that are able to estimate video quality as perceived by human observers is of interest for a number of applications. Depending on the video content, the artifacts introduced by the coding process can be more or less pronounced and diversely affect the quality of videos, as estimated by humans. In this paper we propose a new scheme for quality assessment of coded video streams, based on suitably chosen set of objective quality measures driven by human perception. Specifically, the relation of large number of objective measure features related to video coding artifacts is examined. Standardized procedure has been used to calculate the Mean Opinion Score (MOS), based on experiments conducted with a group of non-expert observers viewing SD sequences. MOS measurements were taken for nine different standard definition (SD) sequences, coded using MPEG-2 at five different bit-rates. Eighteen different published approaches for measuring the amount of coding artifacts objectively were implemented. The results obtained were used to design a novel no-reference MOS estimation algorithm using a multi-layer perceptron neural-network.

Key words: Video quality assessment, no-reference approach, perceptual quality, neural-networks, multi-layer perceptron

1 Introduction

There is an increased need to measure and assess the quality of video sequences, as it is perceived by the multimedia content consumers. The quality greatly depends on the video codec, bit-rates required and the content of video material. User oriented video quality assessment (VQA) research is aimed at providing means to monitor the perceptual service quality.

It is well understood that the overall degradation in the quality of the sequence, due to encoder/decoder implementations as part of transport stream at

[⋆] This work was supported in part by Ministry of Science and Technological Development of Republic of Serbia, under Grant 161003.

various bit rates, is a compound effect of different coding artifacts. Three types of artifacts are typically considered, pertinent to pertinent to DCT block (JPEG and MPEG) coded data: blocking, ringing and blurring. Blocking appears in all block-based compression techniques due to coarse quantization of frequency components [1][2]. It can be observed as surface discontinuity (edge) at block boundaries. These edges are perceived as abnormal high frequency components in the spectrum. Ringing is observed as periodic pseudo edges around original edges [3]. It is due to improper truncation of high frequency components. This artifact is also known as the Gibbs phenomenon or Gibbs effect. In the worst case, the edges can be shifted far away from the original edge locations. This effect is observed as false edge. Blurring, which appears as edge smoothness or texture blur, is due to the loss of high frequency components when compared with the original image. Blurring causes the received image to be smoother than the original one [4].

There is a myriad of published papers that propose different measures of prominent artifacts which appear in coded images and video sequences [1]-[2]. The goal of each no-reference approach is to create an estimator based on the proposed features that would predict the Mean Opinion Score (MOS)[5] of human observes, without using the original (not-degraded) image or sequence data.

In the paper, the applicability of a large set of published features to the problem of MPEG coded video quality assessment is evaluated. An approach to the selection of the optimal set of measures is proposed, where a non-linear estimator is trained to predict MOS. The selection of a smaller subset of objective measures is performed by means of statistical analysis, resulting in a final set of five basic measures. Based on the selected features, a Multi-Layer Perceptron (MLP)[6] as a nonlinear estimator was trained to predict the MOS.

Section 2 provides an overview of the relevant published work. The methodology used is described in Section 3. Section 4 presents the experiments conducted to evaluate the proposed approach and results obtained. Conclusions and some directions for future work can be found in Section 5.

2 Background and related work

The work presented falls within the scope of no-reference methodologies [7]. No information regarding the original (not-coded) video is used to estimate video quality, as perceived by human observers. A subjective quality measure typically used is the mean opinion score (MOS), which is obtained by averaging scores from a number of human observers[8][1]. The correct procedure for conducting such experiments was derived from ITU-R BT.500-10 recommendations[5].

In the research presented here, 18 different measures of image and video quality have been evaluated. Since the goal of the research is to create a VQA approach able to achieve real-time processing, the measures have been selected both for their reported results and simplicity.

Most perceived blockiness measures are based on the notion that the block-edge-related effects can be masked by high spatial activity in the image itself, and that the blockiness cannot be observed in very bright and very dark regions. Wang *et al.*[1] proposed a no-reference approach to quality assessment in JPEG coded images. His final measure is derived as a non-linear combination of a blockiness, local activity and a so-called zero-crossing measure. The combination is supposed to provide information regarding both blockiness and blurring in JPEG coded images.

Recently, Babu *et al.* [8] proposed a blockiness measure for use in VQA, which takes effects along each edge of the block into account separately. Thus, they derive a measure surpassing the Wang *et al.* approach.

Kusuma and Zepernick [7] describe three additional measures focusing on image-activity and contrast. They propose using two different image activity measures edge and gradient activity, as a way to detect and measure ringing and lost blocks.

Spatial activity of the images and video frames in general has a profound effect on the quality of video coding. Within the work presented here, additional measures related to texture have been used to ensure a better description of the spatial activity within the frames of the sequence. These are based on the work Idrissi *et al.* [9].

Kim and Davis [10] proposed a noise and blur measure, aimed at evaluating the quality of video within the framework of automatic surveillance. They show their local-variance-based measure, dubbed fine-structure, able to describe video degradation well, in terms of noise and blur. In order to arrive at a single measure for the quality of a video sequence, based on the values of their proposed measure obtained for the inspected frames of the sequence, they used median as a statistic robust to outliers.

Kirenko [3] proposed simple measures for ringing effects detection, allowing for efficient real-time implementation.

In addition to spatial activity, the coded video quality depends on the temporal dynamics of the sequence. In order to be able to capture the characteristics of video material two motion intensity measures have been devised to describe the average magnitude of motion in a frame: (i) global motion intensity, calculated from the global motion field, and (ii) object motion intensity, calculated by subtracting the global motion from the MPEG motion vectors [2].

In 2005, Babu and Perkis proposed using their proposed quality measures to train a MLP estimator of MOS [11], when JPEG coded images are concerned. MLP has not, to the best of our knowledge, been used for VQA.

3 The Proposed Method for Video Quality Assesment

An set of 18 different features has been evaluated based on the VQEG sequences [12]. The features,with their respective references, are listed in Table 1. To make for an efficient VQA approach the set of features has been reduced to 5 features deemed to describe the quality best. These five features have subsequently been used to train a multi-layer perceptron neural-network, as an estimator for the MOS of new sequences.

#	Feature	Reference
1	Two field difference	[13]
2	Variance ratio	[10]
3	Blockiness	[8]
4	Ringing	[3]
5	Ringing 2	[3]
6	Global motion vector intensity	[2]
7	Activity	[1]
8	Blocking effect	[1]
9	Zero-crossing rate	[1]
10	Z score	[1]
11	Gradient activity	[7]
12	Edge activity	[7]
13	Contrast	[7]
14	Correlation	[9]
15	Energy	[9]
16	Homogeneity	[9]
17	Variance	[9]
18	Contrast	[9]

Table 1. List of measures evaluated with pertinent references.

3.1 Creating the Training Set

The training set used is based on nine SD sequences made available by Video Quality Experts Group (VQEG) for purposes of testing the quality of video codecs. Each sequence has been encoded using five different bit-rate settings (0.5Mb, 1Mb, 2Mb, 3Mb, 4Mb). Values of the features have been calculated for 110 frames of the sequences, i.e. half of the frames of the sequence, distributed uniformly. The mean opinion score (MOS), which is a subjective quality measure obtained by averaging scores from a number of human observers, is derived from tests created according to ITU-R BT.500-10 [5] recommendations. Double Stimulus Continuous Quality Scale (DSCQS) method was used, where pairs of sequences were presented to the viewer. The first one being an original sequence and the other the processed impaired sequence. The final test video was formed by pairing original and degraded video sequence and the observers were asked to evaluate the quality of overall impaired sequences using a five-point grading scale, from 1 to 5, according to perceived quality. Number of viewers had to be at least 20 for each test run to be able to obtain statistically meaningful results, and the test run was kept to maximum of 30 minutes in order to maintain viewer attention. The final MOS value for a sequence is the average score over for all observers for the sequence at a specific bit rate. The MOS values obtained for the sequences are shown in Table 2.

	Bit rate [Mb/s]				
Test sequence	0.5	1	2	3	4
”Parade”	1.800	1.200	2.900	3.850	4.300
”Harp”	1.150	2.100	2.850	4.200	4.450
”Ant”	1.077	2.038	3.269	3.538	4.500
”Kayak”	1.100	1.850	3.300	3.950	4.700
”Formula”	1.885	2.385	3.308	4.192	4.231
”Food court”	1.150	2.150	3.550	4.400	4.800
”Scrolling titles”	1.450	2.800	3.650	3.950	4.400
”Football”	1.200	1.800	3.150	3.800	4.700
”Train”	1.962	1.615	3.231	4.154	4.654

Table 2. MOS for the training sequences.

3.2 Feature Ranking and Selection

To evaluate the predictive capability of each feature (measure), when MOS estimation is concerned, a wrapper methodology for attribute selection has been used [14]. Each feature was evaluated separately by providing it as input of a simple MLP, whose output was the MOS prediction. A simple Multi-layer perceptron (MLP) neural-network estimator has been trained based on a single measure. The estimators contained 3 nodes in a single hidden layer and were trained using 50% of our data, 25% was used for validation and another 25% for testing. A set of statistics was collected for the performance of each estimator, including: root mean square error (RMSE), Pearson correlation, Spearman correlation, maximum absolute prediction error (MAPE) and outlier ratio (OR). The features were than ranked according to the performance of the estimators. The ranking of measures determined through evaluation conducted on the VQEG sequences is shown in Table 3. The values of the statistics are listed along with the feature number corresponding to numbers in Tables 1 and 4. Table 4 provides the descriptions of the top-ranking features.

As the tables show, the highest ranking feature is the combined measure of Wang *et al.* (the Z-score). However, since the two out of three constituents of this measure ranked high (blocking effect and zero-crossing rate), the Z-score was not selected for the final set of features. The rationale for this being the fact that the MLP should be able to combine the constituents in a more informed way and achieve better performance. Thus, the final set of features selected includes: the blockiness measure of Babu *et al.*, the blocking effect measure and the zero-crossing rate of Wang *et al.*, the edge activity measure by Kusuma and Zepernick and the second ringing metric proposed by Kirenko. These are indicated in bold print in Table 4.

Forward selection has been explored as an alternative to the proposed approach, where features have been added to the selected set, using progressively more complex MLP estimators to rank the growing feature sets. Selecting the

RMSE	#	MAPE	#	Spearman	#	Pearson	#	OR	#
1.0264	10	0.2853	10	0.4443	10	0.5344	10	0.0576	10
1.1320	3	0.3198	3	0.3192	8	0.3560	9	0.0365	8
1.1349	9	0.3330	8	0.2964	9	0.3534	8	0.0239	3
1.1357	8	0.3331	5	0.2635	3	0.3506	3	0.0179	6
1.1528	5	0.3377	9	0.2156	6	0.2947	5	0.0135	12
1.1684	12	0.3424	18	0.2048	5	0.2535	4	0.0100	16
1.1714	11	0.3433	12	0.1962	11	0.2495	16	0.0095	5
1.1746	4	0.3445	6	0.1833	12	0.2466	12	0.0077	4
1.1748	16	0.3446	11	0.1773	16	0.2464	7	0.0069	15
1.1755	15	0.3454	15	0.1657	7	0.2460	11	0.0064	18
1.1768	7	0.3454	4	0.1539	1	0.2460	6	0.0063	9
1.1769	6	0.3456	16	0.1362	13	0.2440	15	0.0056	11
1.1790	18	0.3471	7	0.1351	15	0.2392	18	0.0048	13
1.1945	13	0.3507	1	0.1318	18	0.1767	13	0.0043	1
1.1976	1	0.3508	13	0.0989	4	0.1638	1	0.0021	7
1.2059	17	0.3529	17	0.0661	2	0.1170	17	0.0020	17
1.2102	14	0.3532	14	0.0405	17	0.0836	14	0.0014	14
1.2154	2	0.3547	2	0.0345	14	0.0451	2	0.0002	2

Table 3. Feature ranking.

#	Feature	Reference
3	**Blockiness**	[8]
5	**Ringing 2**	[3]
8	**Blocking effect**	[1]
9	**Zero-crossing rate**	[1]
12	**Edge activity measure**	[7]
10	Z score	[1]

Table 4. Description of top ranking features with pertinent references.

best feature set after each iteration, yielded exactly the same ranking as the independent analysis.

3.3 VQA estimator

A block diagram of the proposed video-quality estimator is shown in Fig. 1. Based on the selected set of features a MLP neural network is trained. The network contains 5 input nodes, 7 nodes in the hidden layer and a single output node corresponding to the MOS. No significant gain in prediction performance has been observed when increasing the number of nodes in the hidden layer.

The video quality assessment is conducted by calculating the five selected features for half of the frames of the sequence, uniformly distributed (i.e. the frame rate was halved to make the approach more efficient). The features obtained for

Test sequence	RMSE train	RMSE test	RMSE test stddev
"Parade"	0.6631	0.5142	0.0747
"Harp"	0.7424	0.9697	0.0825
"Ant"	1.1410	1.1460	0.1494
"Kayak"	0.7741	0.9475	0.0428
"Formula"	0.8178	0.6938	0.1487
"Food court"	0.8263	0.9351	0.0793
"Scrolling titles"	0.7852	0.6113	0.1204
"Football"	0.6941	0.7898	0.0218
"Train"	0.8127	0.7052	0.1223

Table 5. Cross-validation results.

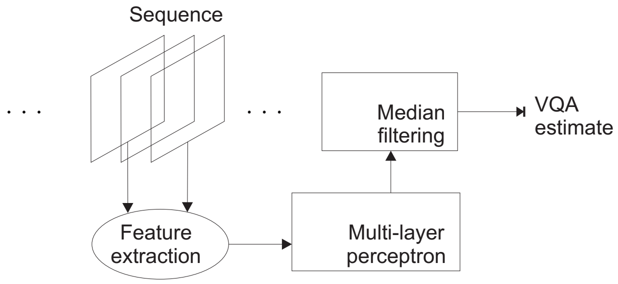


Fig. 1. Block diagram of the proposed approach.

each evaluated frame were fed into the neural network and the measure of the quality for that frame obtained.

Since the standard deviation of the estimator RMSE (RMSE test stddev) over the frames of a single sequence is relatively high, robust statistics should be used to arrive at the final single measure of sequence video quality. Kim and Davis [10] suggest using the median of the quality values across the frames to achieve this. We followed their recommendation and adopted the median of values across the evaluated frames of the sequence as the final measure of sequence quality. Median is known to be a measure robust to the outliers, which commonly occurred in the experiments performed.

4 Results and Discussion

Two different approaches to the testing of the proposed approach were taken: using a part of the data as a separate test set and cross-validation.

Based on a test set comprised of 25% of data available, the proposed estimator achieved the RMSE value of 0.6364 averaged over 20 trial runs, with a standard deviation of 0.0241. The best published quality measure evaluated (Z-score of Wang *et al.*) achieved significantly higher RMSE (1.0264), suggesting that the proposed approach benefited from additional features introduced. The plots of of the test set results achieved per test case (sequence coded at a specific bit rate) are shown in Fig. 2, for both the proposed approach (MLP) and Wang *et al.*. Estimate for a specific test case is the median value of quality estimates across all the evaluated frames of the sequence. As the figure shows, the proposed approach is able to achieve significantly better prediction than that of Wang *(*et al.) approach.

The error on the training set comprised of 50% of the data (another 25% was used for validation), was 0.6268, indicating that there was no over-fitting.

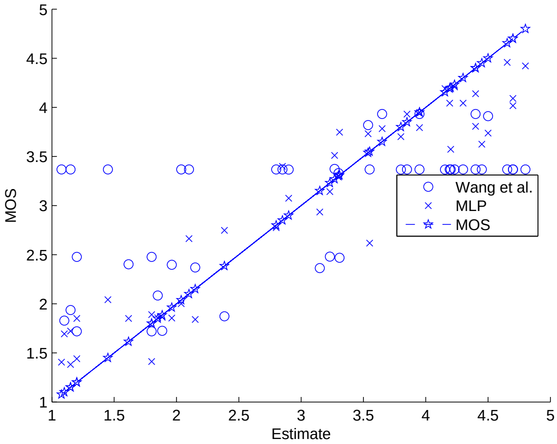
To evaluate the applicability of the proposed approach to a more realistic scenario, where quality evaluation is to be done for new sequences the likes of which may not be present in the training set, cross-validation was performed. This was done in a supervised way, by excluding all data pertinent to a single sequence. The results of this nine-fold cross-validation are shown in Table 5. While the estimator maintained a good RMSE, the results indicate that the training set is not diverse enough to allow for balanced performance when whole sequences are excluded. This suggests that the training set should be extended, and possibly that specialized estimators should be constructed based on the sequence characteristics and/or content.

5 Conclusion

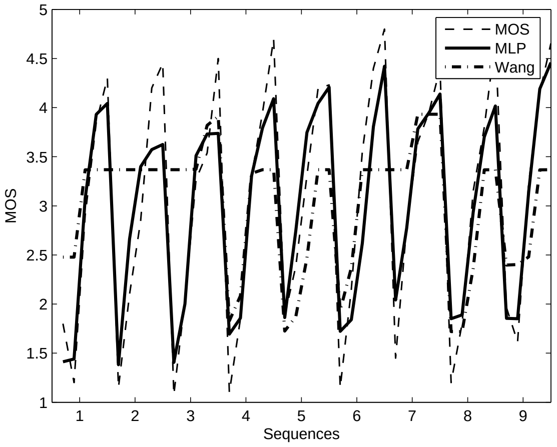
A large number of features designed to detect and measure the coding artifacts introduced by DCT block coding algorithms, has been evaluated in terms of applicability to video-quality assessment of MPEG2 coded video sequences. An approach to determining the correct reduced set of features, based on the training data available has been described. A multi-layer perceptron based estimator of MOS has been trained using the five selected features. The proposed estimator achieved results superior to those of the single features evaluated, in terms of RMSE, when compared on frame-by-frame basis. The results of the experiments conducted suggest that a larger set of sequences should be used for MLP training in order to improve performance in a general case. In addition, the sequences could be separated into similar groups and specialized estimators constructed for each cluster, in order to improve the performance even further.

References

1. Wang, Z., Sheikh, H.R., Bovik, A.C.: No-reference perceptual quality assessment of jpeg compressed images. (2002) 477–480
2. Warwick, G., Thong, N.: Signal Processing for Telecommunications and Multimedia, Chapter 6: Classification of Video Sequences in MPEG Domain. Springer (2004)
3. Kirenko, I.: Reduction of coding artifacts using chrominance and luminance spatial analysis. Consumer Electronics, 2006. ICCE '06. 2006 Digest of Technical Papers. International Conference on (Jan. 2006) 209–210
4. Ferzli, R., Karam, L.: A no-reference objective image sharpness metric based on just-noticeable blur and probability summation. Image Processing, 2007. ICIP 2007. IEEE International Conference on **3** (16 2007-Oct. 19 2007) III –445–III –448
5. BT.500, I.R.: Methodology for the Subjective Assessment of the Quality of Television Pictures. (2002)
6. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)
7. Kusuma, T., Caldera, M., Zepernick, H.: Utilising objective perceptual image quality metrics for implicit link adaptation. (2004) IV: 2319–2322
8. Venkatesh Babu, R., Perkis, A., Hillestad, O.I.: Evaluation and monitoring of video quality for uma enabled video streaming systems. Multimedia Tools Appl. **37**(2) (2008) 211–231
9. Idrissi, N., Martinez, J., Aboutajdine, D.: Selecting a discriminant subset of co-occurrence matrix features for texture-based image retrieval. (2005) 696–703
10. Kim, K., Davis, L.: A fine-structure image/video quality measure using local statistics. (2004) V: 3535–3538
11. Babu, R., Perkis, A.: An hvs-based no-reference perceptual quality assessment of jpeg coded images using neural networks. (2005)
12. `ftp://ftp.crc.ca/crc/vqeg/TestSequences/Reference/`
13. Wolf, S., Pinson, M.: Ntia report 02-392: Video quality measurement techniques. Technical report, Institute for Telecommunication Sciences
14. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco (2005)



(a) Estimate scatter plot: proposed approach (MLP), Wang *et al.* (Wang) and true MOS values (MOS).



(b) Estimate over the sequences: proposed approach (MLP), Wang *et al.* (Wang) and true MOS values (MOS).

Fig. 2. Test results for the test set containing 25% of data.