https://doi.org/10.2352/EI.2022.34.9.IQSP-319

A continuous bitstream-based blind video quality assessment using multi-layer perceptron

Hugo Merly, Alexandre Ninassi and Chistophe Charrier - Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France

Abstract

We propose a continuous blind/no-reference video quality assessment (NR-VQA) algorithm based on features extracted from the bitstream, i.e., without decoding the video. The resulting algorithm requires minimal training and adopts a simple multi-layer perceptron for score prediction. The algorithm is computationally appealing. To assess the performance of the algorithm, both the Pearson Correlation Coefficient (PCC) and the Spearman Rank Ordered Correlation Coefficient (SROCC) are computed between the predicted values and the quality scores of two databases. The proposed approach is shown to have a high correlation with human visual perception of quality.

Introduction

Recording, streaming and accessing video content has never been easier and faster than today. As the *Human Visual System* (HVS) processes more and more digital imaging data, assess its quality becomes a major issue. To gather opinions from human observers on video quality, a guideline is provided by the Recommendation BT.500 from the ITU-R [6]. The subjective tests are conducted in a specific and controlled environment. As a result, a set of individual scores is used to compute the *Mean Opinion Score* (MOS). Since subjective experiments are not suitable in some cases, e.g. in real-time applications and streaming services, and time consuming, large efforts have been made to develop objective metrics. Based on the properties of the HVS and the analysis of the video content and distortions, they automatically score the quality of video. The aim is to judge the quality as humans do and get the highest correlation with MOS. Such metrics are divided into three categories, depending on the availability of an original video, which refers to reference: 1) the *Full-Reference* (FR) VQA, 2) the *Reduced-Reference* (RR) VQA and 3) the *No-Reference* or *blind* (NR) VQA approaches. While FR and RR schemes require an access to a reference or a set of representative features of it, a NR method assess quality without any information about the reference. Since a video can be parsed as a collection of pixels or as a formatted bitstream, two different approaches exist in each category. The first one is based on the extraction of features computed into any of the spatial domain or a transformed one. The second approach is to extract features directly from the bitstream, without having to decode the video. The latters are usually designed for a particular bitstream format, *e.g.*, following the H.264/AVC [4] or the H.265/HEVC [5] standards, in order to extract bitstream features, which makes it computational-friendly.

In a forensic context, and in order to facilitate the analysis of videos in a given time (around some ms), an automated solution that continuously scores the quality of videos is developed. In addition, due to the lack of the pristine reference, the proposed scheme is a no-reference bitstream-based VQA.

State of the art

In the field of NR VQA metrics, Shahid et al. [14] present a classification and a review of the last progresses. Most of the work has been made using pixel-based approaches since they are independent of the used technology. However, using a bitstream-based approach, Keimel at al. [7] developed a VQA model which extracted 64 features from an H.264 bitstream. They used a partial least square regression as a quality predictor to fit with the MOS. The initial database was composed of four 1080p reference videos, compressed following 8 levels of compression which finally lead to 32 distorted videos. Shahid et al. [15], based on Rosshlom and Lövström's work [11], used machine learning algorithms as a H.264 video quality estimator. In their last work on H.264 bitstreams [13], a support vector machine uses 18 features in order to predict the MOS and four FR VQA schemes. A database of 120 distorted videos, with CIF and QCIF resolution and 20-seconds long, is created. A further improvement is found by the same authors on H.265 bitstreams [12], where a larger features set and database are used and the prediction of the subjective MOS is also performed. Pandremmenou et al. [10] include the influence of packet loss in their estimation of video quality which made by the least absolute shrinkage and selection operator (LASSO). The various models presented in this section predict a single quality score per video, unlike our approach where we will a continuous rating of video quality.

The proposed model

The synopsis of the proposed scheme is depicted in Fig. 1. Instead of assessing the quality of each image of the video, the scoring is performed on each *Group Of Pictures* (GOP). For each of them, a bitstream feature extraction is computed using a homemade bitstream parser. Then, a trained *Multi-Layer Perceptron* (MLP) performs quality prediction on each GOP, namely the *Quality Scores* (QS). The QS results are finally compiled in a continuous rating of the quality of the video. Since the most widely used codec for coding CCTV contents is H.264/AVC, the developed scheme is dedicated to this codec, event if an extension to H.265 codec can be designed.

Features extraction

A video is represented as a sequence of frames arranged in GOP structures. Each GOP consists of an intra-coded frame, namely the key frame, followed by inter-coded frames. Encoding the intra-frame (I-frame) uses the JPEG compression algorithm in order to reduce pixel information. In addition, to take advantage of temporal redundancies in a sequence of frames, inter-frames are encoded as a set of *Motions Vectors* (MV) which represents the displacement of the current pixel block relative to its corresponding block into the references frames. Predicted

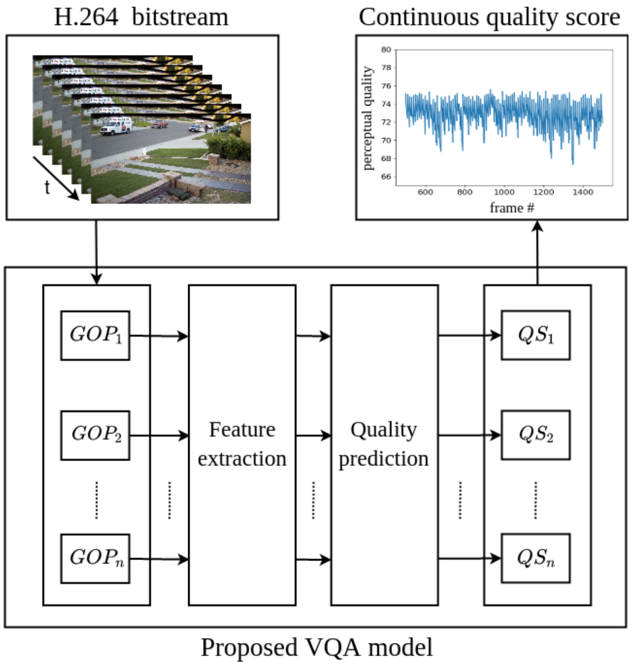

Figure 1. *Overview of the proposed model*

frames (P-frames) use only previous frames as references while Bi-directional frames (B-frame) use both previous and upcoming frames as references. In H.264/AVC, a frame is composed of a set of *slices*, themselves represented by a sequence of chroma and luma samples in the form of variable size blocks, namely the *MacroBlocks* (MB). According to the Annex B of H.264/AVC standard [4], a bitstream is represented by three different layers, as illustrated in Fig. 2. On the top, the *Network Abstraction Layer* (NAL) layer contains video data units: the *Video Coding Layer* (VLC) *Network Abstraction Layer* (NAL) units (NALUs). Each VLC NALU describes one slice for the current frame. Slice data are represented as a collection of MB including skip macroblock indicator, *i.e.*, the parts of the MB that have not been coded. At the bottom, the MB layer presents the features of the current MB.

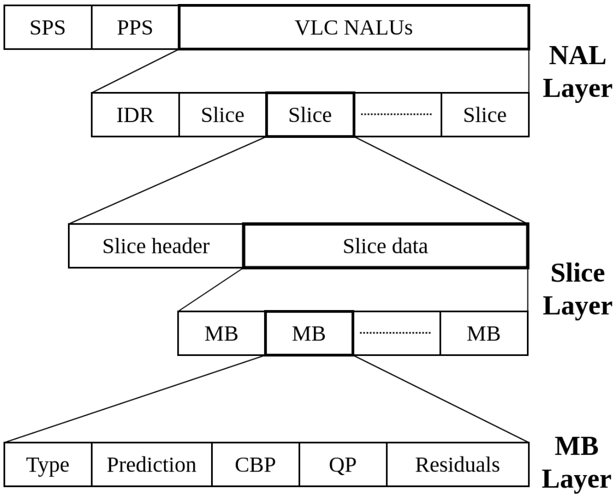

Figure 2. *Overview of the different layers in a H.264/AVC bitstream*

Inspired by the previous works on NR bitstream-based VQA and the works of the Video Quality Experts Group (VQEG) on an adjusted H.264 decoder which generates bitstream traces [2], the NALUs are parsed by a modified H.264 decoder, derived from the reference software [1]. This yields us to extract features at different levels of accuracy. The extracted parameters are the following:

- f_1 : the bitrate
- f_2 : the average Quantization Parameter (QP),
- f_3 : the delta QP, namely ΔQP
- f_4, f_5 : the average and maximum motion vector length
- f_6, f_7 the average and the maximum motion vector error length
- $f_8, \cdots, f_{10}$: the percentage of intra-, inter- and skip-coded MBs
- $f_{11}, \cdots, f_{13}$: the percentage of I-coded MBs with 16x16, 8x8, 4x4 partitioning modes
- $f_{14}, \cdots, f_{17}$: the percentage of P-coded MBs with 16x16, 16x8, 8x16 and 8x8 partitioning modes
- $f_{18}, \cdots, f_{20}$: the percentage of P-coded sub-MBs with 8x4, 4x8 and 4x4 partitioning modes
- $f_{21}, \cdots, f_{24}$: the percentage of B-coded MBs with 16x16, 16x8, 8x16 and 8x8 partitioning modes
- $f_{25}, \cdots, f_{27}$: the percentage of skipped P-coded, B-coded MBs and direct B-coded MBs

The feature f_1 is extracted directly from the slice layer while the remaining features $f_2, \ldots, f_{27}$ are computed from the macroblock layer.

The feature f_3 (ΔQP) is defined as:

$$\Delta\text{QP} = |\text{QP}_{slice} - \overline{\text{QP}_{mb}}| \tag{1}$$

where QP_{slice} is QP extracted at slice level, and $\overline{\text{QP}_{mb}}$ corresponds to the mean of QP extracted for each MB from a slice.

This features extraction provides information about the different aspects of the video. Video distortion at slice and MB levels are represented by the features f_1, f_2 and f_3 while motion content of the video is captured by the features f_4 to f_7. The encoder choices in macroblock coding are finally exhibited by the remaining features f_8 to f_{27}. The features f_8 to f_{24} are described by the tables 7-13 to 7-18 from the H.264/AVC standard [4]. For each slice of any frame of a GOP, these 27 features f_i are extracted.

Finally, a feature vector of each GOP k, namely V_{GOP_k}, is computed as follows

$$V_{GOP_k} = \left(\frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} f_{l,i,j} \right), \forall l \in [1, \ldots, 27] \tag{2}$$

where $f_{l,i,j}$ represents the l-*th* feature of the i-*th* slice of the j-*th* frame of the k-*th* GOP of the video.

Quality prediction

Once the feature vectors have been extracted from the bitstream, the quality score can be predicted for each GOP, in order to get a continuous quality prediction of the video. Among all available schemes, machine learning based approaches have demonstrated their ability to predict quality score with a high confidence and highest correlation with human judgments. From artificial neural networks schemes, the *Multi-Layer Perceptron* (MLP) offers a simple and handy architecture to tackle this regression problem and, furthermore, has been intensively used for pattern classification, recognition, prediction and approximation [13].

A MLP is a deep feedforward neural network which defines a mapping function $y = f(X, \theta)$ between the input vector of features X and the associated output y. The network learns the intrinsic parameters θ in order to create the most accurate approximation function. At least composed of two layers, the input and the output layers, the network may also contains additional so-called *hidden layers*. They act as filters representing different levels of information representation. Composed of bunches of *neurons* (or nodes), layers are characterised by a set of *weights* ω and a *bias* b, while layer outputs are shaped by a transfer function, the *activation function* ϕ. Since layers are fully connected in a MLP, the output y_j of the node j of the layer l_k is expressed as the sum of the weighted outputs of the layer l_{k-1} shaped by the activation function g :

$$y_j = g\left(b + \sum_{i=1}^{N} \omega_{ij} x_i\right) \quad (3)$$

where x_i is the output of the node i in the layer l_{k-1}, w_{ij} is the weight of the link between the node i in the layer l_{k-1} and the node j in the layer l_k and N is the width of the layer l_{k-1}. Historically, the activation function traduces the fire up conditions of a biological neuron as a step function. Continuous non-linear and piece-wise linear functions are now commonly used. Their differential properties are essential for applying *backpropagation algorithms*. During the learning phase, the vectors of training examples X_i passes through all layers of the network in order to predict the ground truth y_i in the output layer. An error function, namely the *loss function* J, is computed to quantify the average of differences between the predicted output $\hat{y}_i$ and the ground truth y_i. The *mean square error* (MSE) is employed as the loss function J to deal with regression problems. In order to reduce the loss function, the *gradient descent optimization algorithm* updates the network parameters ω and b, collectively denote as θ, in the opposite direction of the gradient ∇_θ of the loss function $J(\theta)$: $\theta = \theta - \alpha \cdot \nabla_\theta J(\theta)$, where α is the learning rate. During this step, the backpropagation algorithm is used to compute the gradient of the loss function with respect to the network parameters. It is common practice to use the *mini-batch gradient descent* to optimize the parameters update. The network parameters are iteratively updated after the processing of one random subset of training samples. With these various techniques, the network learns the best tuning of its intrinsic parameters in order to find the global minimum of the loss function. Once the model is trained, the testing phase consists of propagating a new set of samples through the network. This forward pass allows to evaluate the accuracy of the model is evaluated against an unknown testing set.

To assess quality prediction on an unknown H.264 bitstream, the best trade-off between complexity and performance is sought. The number and the size of the hidden layers have been determined as follows: the final network is composed of three hidden layers of 27 neurons each, the same size as the input feature vector. The *rectified linear unit* (ReLU) function is used as the activation function by each layer and the loss function is described by the MSE. The gradient descent optimizer and the backpropagation algorithm are used together to ensure the network learns the best possible parameter setting.

Searching for database

To feed the neural network describes in the previous part, a video database is required. Due to our field of investigations, the database must include videos from CCTV cameras and the associated continuous MOS, *i.e.*, the continuous ground truth which can be viewed as a quality score for each frame. To the best of our knowledge, such CCTV video databases with continuous MOS are not available. Moreover, no unspecific video database with a continuous MOS has been published yet. By the way, we have to face a twofold challenge: 1) build a CCTV database and 2) provide the associated ground truth.

Database design

In order to build a consistent database, 20 CCTV video sequences have been selected for their potential interest. They represent various situations and localization. CCTV cameras are located above roads and highways, but also above parking lots, busy streets and industrial and residential areas. They can be fixed, rotating and they can zoom on areas of interest. Video sequences are taken at different times of the day (day, night, sunset, sunrise) taking account of variations in lighting conditions. The video sequences have a resolution of 1280x720 pixels (HD) and last between 30 seconds and 5 minutes. A snapshot of the database is presented Fig. 3. After collecting the initial 20 videos, the aim is to build the final database containing each video at different compression levels. The x264 library, included in the well-known FFmpeg software, is used to encode each video with 11 different levels of compression. Set by the *QP* parameter, the compression ratio of encoded videos fluctuates between low ($QP = 20$) and high ($QP = 51$) levels.

The final database, namely the G-CCTV database, is composed of 20 references videos and its 11 distorted versions, for a total of 240 videos. From a GOP point of view, the G-CCTV database contains 67 892 samples, which will be fed by the network, a part of them for the training phase, the remaining for test one.

Ground truth generation

For each GOP of the G-CCTV database, a vector of 27 features is computed which is used as the input of the designed MLP. Yet, to train the proposed scheme, a ground-truth is needed to adjust hyper-parameters values.

Usually, psycho-visual tests with many people are used to generate such a ground-truth, *i.e.*, the MOS. Specific and controlled associated environments for those tests are defined by the ITU-R BT.500-14 recommendation [3]. Yet, such a test suffers many drawbacks: the selection of people involved in the test, the process is time consuming (many hours over several month), access to a dedicated room, and so on.

In this paper, we investigate how it is possible to use scores provided by VQA algorithms. In other words, can we estimate the amount of bias we might induces if the scoring is performed using existing VQA algorithms, and are observed differences between subjective scores and predicted scores statistically significant?

Among all existing VQA algortihms, four FR I/VQA

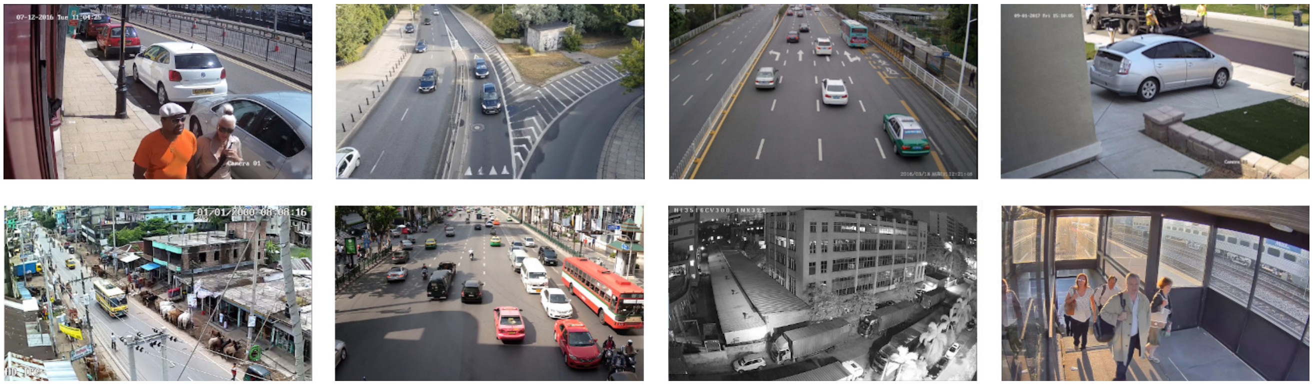

Figure 3. *A snapshot of the reference videos of the G-CCTV database*

schemes have been selected: Structural Similarity (SSIM) Index [21], the Multi-Scale Structural Similarity (MS-SSIM) Index [20], the Visual Information Fidelity (VIF) [16] and the Video Multi-method Assessment Fusion (VMAF) [9]. All those schemes provide the highest correlation with human judgments and the implementations of the algorithms were either publicly available on the Internet or obtained from the authors. A brief description of the trail VQA algorithms are hereafter:

1. SSIM is a widely used structural-based IQA. It computes the structural similarities of two images by comparing three features as the luminance, the contrast and the structure. The similarities are calculated locally, and the computations are performed by a sliding window, before averaging over the entire image.
2. MS-SSIM is an iterative IQA as an extension of the SSIM Index. Similarities are computed at different stages. Between each stage, the images are low-pass filtered and down-sample. The image quality is finally assessed on the last stage.
3. VIF is a statistical IQA which uses Natural Scenes Statistics (NSS) to model natural inputs images. The source image, namely the reference, is modeled by a Gaussian Scale Mixture (GSM). The tested image is modeled by the distortion model. It is obtained by attenuating and adding Gaussian noise to the source model. VIF is finally computed between the two images by comparing their VIF criterion.
4. VMAF is an open-source, learning-based FR VQA model. It fuses two FR IQA models, DLM [8] and VIF [16] and a temporal feature as three inputs of a Support Vector Machine (SVM) algorithm. The machine learning algorithm is trained on Netflix databases to match with MOS scores.

In order to evaluate the bias we introduced exploiting FR VQA values instead of subjective ones, we applied a hypothesis test between the subjective scores and the ratings provided by the trial VQA algorithms. This test is based on the Student t-test which determines whether two population means are equal or not.

The t-test is performed on the dataset provided by the LIVE Wild Compressed Video Quality Database [22]. The database is composed of 55 reference videos from the LIVE Video Quality Challenge (VQC) Database [17, 18, 19]. This latter will be described in the following section. Each reference is down-scaled and compressed by the H.264 codec at four different compression levels, for a total of 220 distorted videos. For this database, the authors computed the MOS on the whole database, leading to 275 subjective scores. The four schemes presented above are full reference metrics, so it makes sense to perform the t-test between the Differential MOS (DMOS) and the four FR VQA values. The DMOS of a distorted video y is computed from the MOS of the reference video x and the MOS of the distorted video y as :

$$DMOS(y) = MOS(x) - MOS(y) \tag{4}$$

To use the same range as the MOS, from 0 (bad) to 100 (excellent), the DMOS is normalized by the min-max scaler and multiply by a factor 100. The final nDMOS of the distorted video y is expressed as :

$$nDMOS(y) = \frac{DMOS(y) - min(DMOS)}{max(DMOS) - min(DMOS)} \cdot 10^2 \tag{5}$$

The results of the two-sided t-test, from the quality scores generated on the 220 compressed videos, are presented in tab. 1. The significant level in order to reject the null hypothesis is $\alpha = 0.05$. Rejection of the null hypothesis, *i.e.*, the t-test shows that two populations are statistically different, is represented by a 1 when $p - value \leq \alpha$. Affirmation of the null hypothesis, *i.e.*, the t-test shows that two populations are not statistically different, is represented by 0 when $p - value \geq \alpha$. As we can observe, VMAF is the only VQA that does not present statistical differences between the predicted quality score and the nDMOS, as the null hypothesis is not rejected.

Thus, we assume that VMAF can be used to provide Quality Scores (QS) that can serve as ground truth without introducing statistically significant bias.

SSIM, MS-SSIM and VIF quality scores will serve as ground truths thereafter for performance comparisons.

Results

The results are divided in two parts. Firstly, the model is trained and tested on the G-CCTV database and the performances are computed using trail VQA algorithms as ground-truth.

Secondly, the generalization capability of the proposed scheme is investigated. The model is trained on the entire G-CCTV database, and it is tested on another database, namely the LIVE VQC Database [17, 18, 19], and vice-versa. In order to

T-test results between nDMOS values and the predicted scores of the four FR VQA schemes on the LIVE Wild Compressed Video Quality Database. For the last column, the symbol gives the result of the hypothesis test on the database: '1' means that the mean of the scores provided by the algorithm for the row is statistically different than the mean of dMOS, '0' means it is not statistically different.

Metrics	T-value	P-value	nDMOS
SSIM	33.85	$2.66 \cdot 10^{-121}$	1
MS-SSIM	33.85	$2.60 \cdot 10^{-121}$	1
VIF	34.03	$4.98 \cdot 10^{-122}$	1
VMAF	0.33	0.74	0

quantify the model performances, the *Pearson Correlation Coefficient* (PCC) and the *Spearman Rank Order Correlation Coefficient* (SROCC) are computed between the predicted QS by the proposed scheme and the ground truth values.

Model performance

In order to properly evaluate the performance of the model, the G-CCTV database is divided into two distinct subsets: one dedicated to the training phase and one to the testing phase. The first one represents 70 % of the database, composed of videos from 14 randomly selected reference videos, and the second one is composed of the remaining 30 %, composed of videos from the 6 remaining reference videos. The prediction of the QS, obtained by the model, is repeated 1000 times, with for each iteration a new random splitting of the training set and the test set. The Fig. 4 shows the performance of the model represented by the PCC and the SROCC between the expected QS values (provided by the four trail algorithms) and the QS predicted by the model, with their 99% confidence interval. The highest correlation rate is obtained when VMAF scores are used to train the model, with a score of 0.99. Considering VIF, the correlation rates are approximately 0.98. Since VIF is an intrinsic part of VMAF, it is not surprising to find similar correlation scores. The same holds for SSIM and MS-SSIM where the correlation scores although lower are very close, up to 0.96. The 27 bitstream features are more correlated to QS generated by VMAF or VIF models than SSIM or MS-SSIM models. These results consolidate our previous results, which show the relevance of using VMAF instead of DMOS as a ground truth.

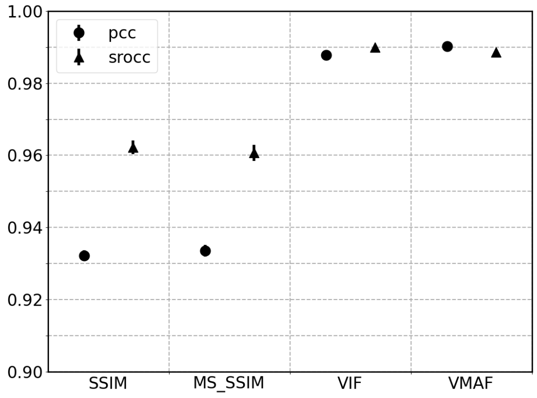

Figure 4. *Performance of the proposed model using each of the four FR-VQA methods as ground truth. The PCC and SROCC values are shown with their 99% confidence intervals.*

Database Independence

The generated G-CCTV database contains only specific types of videos. In order to not limit the model to CCTV videos, we have investigated the independence of the model with respect to the content of the training database. For this, the LIVE VQC database is used for the richness of its content. It is composed of 585 videos of different resolutions, captured on the fly by 80 users with a various set of smartphones. Each video content comes from natural scenes such as concerts, sport events, travel videos and so on.

As for the G-CCTV database, a new database, derived from the LIVE VQC database, of 6435 compressed videos and the 585 associated references is created by modulating the QP value during the encoding process. It produces 107 484 samples, corresponding to each GOP of the database, on which the ground truth is provided by VMAF.

Performance of the model using different databases for training and testing, with a ground truth generated by VMAF.

	PCC		SROCC	
Train set Test set	G-CCTV	LIVE-VQC	G-CCTV	LIVE-VQC
G-CCTV	-	0.957	-	0.948
LIVE-VQC	0.948	-	0.938	-

To evaluate the model performances over the two databases, the model is firstly trained on the whole G-CCTV database and then tested on the whole derived LIVE VQC database. The reverse operation is performed, *i.e.*, the model is trained on the derived LIVE VQC database and then tested on the G-CCTV database.

As previously, a replication strategy of 1000 replicates is applied to compute the quality scores. Both PCC and SROCC are used to evaluate the performance of the proposed approach.

The results are provided in tab. 2. The correlation rates achieved in both cases remain consistent, despite the use of different training and test databases. The obtained values are around 0.95 and are very similar in both cases, with slightly higher results when the model is trained on the LIVE VQC database. As this database is richer in content types, it is not surprising to observe this trend. These results show that the used database to train the model has very limited influence on the performance of the system. Thus, the developed model can be used to evaluate the quality of any video database while maintaining high performance.

Conclusion

We have described a bitstream-based approach to the no-reference/blind VQA problem. The proposed algorithm continuously scores the quality of the video, instead of providing only one score. The new NR-VQA model uses a small number of computationally convenient features directly extracted from the bitstream. The algorithm can be easily trained to achieve excellent predictive performance using a MLP model. The method correlates highly with human visual judgments of quality.

We have shown that when no DMOS are available, it is possible to use predicted scores to train the model. Investigations allow us to hypothesize that scores obtained using VMAF can be used as ground-truth to train the model. In addition, it as been shown that no bias is introduced in the results.

References

[1] H.264 Reference Software. http://iphome.hhi.de/suehring/tml/.

[2] Modified JM H.264/AVC Codec. https://vqeg.github.io/software-tools/encoding/modified-avc-codec/.

[3] Methodologies for the subjective assessment of the quality of television images. Recommendation ITU-R BT.500-14, ITU, 10/2019.

[4] Advanced video coding for generic audiovisual services. Recommendation H.264, ITU-T, 2003-2021.

[5] High efficiency video coding. Recommendation H.265, ITU-T, 2013-2021.

[6] Methodologies for the subjective assessment of the quality of television images. Recommendation BT.500-14, ITU-R, 2019.

[7] Christian Keimel, Manuel Klimpke, Julian Habigt, and Klaus Diepold. No-reference video quality metric for hdtv based on H.264/AVC bitstream features. *2011 18th IEEE International Conference on Image Processing*, pages 3325–3328, 2011.

[8] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, 2011.

[9] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[10] Katerina Pandremmenou, Muhammad Shahid, Lisimachos Paul Kondi, and Benny Lövström. A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses. In *Electronic Imaging*, 2015.

[11] Andreas Rossholm and Benny Lovstroem. A new low complex reference free video quality predictor. In *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 765–768, 2008.

[12] Muhammad Shahid, Joanna Panasiuk, Glenn Van Wallendael, Marcus Barkowsky, and Benny Lövström. Predicting full-reference video quality measures using HEVC bitstream-based no-reference features. *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–2, 2015.

[13] Muhammad Shahid, Andreas Rossholm, and Benny Lövström. A no-reference machine learning based video quality predictor. *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 176–181, 2013.

[14] Muhammad Shahid, Andreas Rossholm, Benny Lövström, and Hans-Jürgen Zepernick. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on Image and Video Processing*, 2014:1–32, 2014.

[15] Muhammad Shahid, Andreas Rossholm, and Benny Lövström. A reduced complexity no-reference artificial neural network based video quality predictor. In *2011 4th International Congress on Image and Signal Processing*, volume 1, pages 517–521, 2011.

[16] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[17] Zeina Sinno and Alan C. Bovik. Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2018.

[18] Zeina Sinno and Alan C. Bovik. Live video quality challenge database. http://live.ece.utexas.edu/research/LIVEVQC/index.html, 2018.

[19] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019.

[20] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[21] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity, 2004.

[22] Xiangxu Yu, Neil Birkbeck, Yilin Wang, Christos G. Bampis, Balu Adsumilli, and Alan C. Bovik. Predicting the quality of compressed videos with pre-existing distortions. *IEEE Transactions on Image Processing*, 30:7511–7526, 2021.

Author Biography

Hugo Merly received the engineering degree in electronics from the University of Science and Technology, Nantes, France, in 2019. For the past year, he has been working as a research engineer with the GREYC laboratory at the University of Caen, France, which is centred on the field of digital sciences. His current research interests include video quality assessment, face recognition and artificial intelligence.

Alexandre Ninassi graduated from engineering school of University of Nantes in 2005. He obtained his PhD in Information Technology from the University of Nantes in 2009. In 2010, he joined the ENSICAEN school of engineering in Caen as research engineer. He belongs to the SAFE (Security, Architecture, Forensics, biomEtrics) unit in the GREYC research lab. His research focuses on biometrics and design of image and video quality metrics.

Christophe Charrier (M'10) received the M.S. degree from the Nantes University of Science and Technology, Nantes, France, in 1993, and the Ph.D. degree from the University Jean Monnet, Saint-Etienne, France, in 1998. He has been an Associate Professor with the GREYC laboratory at the University of Caen, France, since 2001. In 2008, he was a Visiting Scholar with the Laboratory for Image and Video Engineering, University of Texas, Austin. From 2009 to 2011, he was an Invited Professor with the Computer Department, University of Sherbrooke, Sherbrooke, QC, Canada. He is now the head of the SAFE research team. His current research interests include digital image and video coding, quality assessment, computational vision, biometrics and forensics.