Machine Learning approach for global no-reference video quality model generation

Ines Saidi, Lu Zhang, Vincent Barriac, Olivier Deforges

Ines Saidi, Lu Zhang, Vincent Barriac, Olivier Deforges, "Machine Learning approach for global no-reference video quality model generation," Proc. SPIE 10752, Applications of Digital Image Processing XLI, 1075212 (17 September 2018); doi: 10.1117/12.2320996

SPIE.

Event: SPIE Optical Engineering + Applications, 2018, San Diego, California, United States

Machine Learning approach for global no-reference video quality model generation

Ines Saidi[a,b], Lu Zhang[b], Vincent Barriac[a], and Olivier Deforges[b]

[a]Orange Labs Lannion, France
[b]IETR, CNRS UMR 6164, INSA Rennes, France

ABSTRACT

Offering the best Quality of Experience (QoE) is the challenge of all the video conference service providers. In this context it is essential to identify the representative metrics to monitor the video quality. In this paper, we present Machine Learning techniques for modeling the dependencies of different video impairments to the global video quality perception using subjective quality feedback. We investigate the possibility of combining no-reference single artifact metrics in a global video quality assessment model. The obtained model has an accuracy of 63% of correct prediction.

Keywords: QoE, video quality, video coding and transmission, no reference metric, Machine Learning

1. INTRODUCTION

With the rapid development of broadband telecommunication technologies and the expansion of mobility (3G, LTE, 5G and WIFI), video conferencing services have been created to complement face-to-face conversations. However, the Quality of Service (QoS) of these applications is usually not guaranteed. The IP-based networks, do not transmit the multimedia streams without errors. Many processes in the supply chain may degrade the perceptual quality. In this sense, there is a strong need for a measure of user perception of the quality in order to establish a trade-off between the user satisfaction and the available network resources. However, video quality evaluation is a complex task given the multiplicity of parameters impacting the perceived media. Many researches propose methods and automatic tools for objective video quality evaluation,[1–6] but there is no representative metric relevant for all degradation types and network conditions.

In our study context of videoconferencing services, we focus mostly on real time video quality assessment. In this case, we consider no-reference metrics which evaluate the quality of a video by measuring, from the video under test, the distortions that may have been caused by encoding, trans-coding and/or transmission. These metrics are the key indicators of audiovisual quality developed by the Department of Telecommunications in the AGH University of Science and Technology. This research work is a part of the MOAVI (Monitoring Of Audiovisual Quality by Key Indicators) project within the Video Quality Experts Group (VQEG).[7] Each of these metrics allow to measure the level of a single type of distortion that can occur for instance in a video conference call such as blur, noise, blockiness, flickering, block loss, freezing, slicing and interlacing.

However, the human perception of the quality does not distinguish between the types of distortion, but it gives a global appreciation of the quality. Our idea is then to try to combine all these MOAVI single artifact-based metrics into a global video quality model generated by Machine Learning (ML) algorithms. We distinguish two types of Machine Learning algorithms: unsupervised and supervised learning. The unsupervised algorithm consists in estimating the structure of an unlabeled data. The use case of an unsupervised algorithm is the classification of data into categories. On the other side, the supervised learning is used when the category structure of the database is already known. Thus, the supervised learning predicts a function or a model that maps the database to the predefined class labels. In our case we are considering supervised learning, and we are

Further author information. Send correspondence to:
E-mail: ines.saidi62@gmail.com, vincent.barriac@orange.com
E-mail: { olivier.deforges, lu.ge} @insa-rennes.fr

Applications of Digital Image Processing XLI, edited by Andrew G. Tescher, Proc. of SPIE Vol. 10752, 1075212 · © 2018 SPIE · CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2320996

interested in classification methods because of the discrete and labeled nature of our dataset and because our objective it to predict a variable.

The paper is organized as follows: Section 2 describes the database used to train the machine learning model. In Section 3 the video quality metrics applied on the database contents are presented. Section 4 introduces the used data mining tool. It is followed by Section 5, which shares the main findings and modeling results. Finally, Section 6 concludes the paper.

2. TRAINING DATABASE

The performance of each model generated by machine learning method depends directly on the used training database. The more database contains values representative of the final use cases and conditions, the more accurate the predictive model is. Thus, in our case we collected all the subjective databases available to us. We considered databases obtained in our previous subjective tests that has been already published in.[8,9] These tests was performed according to the ITU standards for subjective quality assessment of video services ITU-R BT.500[10] and ITU-T P.910.[11]

To broaden the spectrum of conditions, degradations and contents, we collect other bases of video sequences. From the literature we choose subjective databases with variety of the included impairment types: transmission error (packet loss, jitter, freezing, etc.), coding (variable bit rates), frame rate, different error concealment algorithms. These databases will come complete the sequences of our subjective tests and will be used to constitute the training database of the machine learning algorithm.

The characteristics of all the databases described below are summarized in Table 1.

2.1 LIVE Mobile video quality assessment Database

The Live Mobile database is developed by the Laboratory for Image and Video Engineering at the University of Texas.[12] It's one of the most popular public VQA databases used by researchers to evaluate objective video quality assessment algorithms for wireless video transmission with regards to their efficacy in predicting visual quality. The importance of the LIVE Mobile VQA database is that it contains temporal distortions in addition to compression and packet loss distortion. In total, the distortion conditions consist of 4 conditions for H.264 compression impairments, 4 wireless-packet losses, 4 duration of frame freezes, 3 rates adapted and 5 temporal dynamics per reference. Details on these distortions are explained by authors in.[13]

- Compression impairments: encode source videos with H.264 Scalable Video Codec (SVC) at four bit rates ($R1 < R2 < R3 < R4$) between $0.7Mbps$ and $6Mbps$. 40 distorted videos are in this category.
- Frame freezes on stored video delivery and real time live video delivery: four conditions were simulated for each source video which leads to a total of 40 distorted videos.
- Rate adaptation: change the coding bit rate during the video. We have 30 rate adapted distorted videos.
- Temporal dynamics: simulate multiple switches of the coding rate yielding 50 distorted videos.
- Wireless channel packet loss. 40 distorted videos are generated.

2.2 EPFL-PoliMI video quality assessment Database

The EPFL-PoliMI (Ecole Polytechnique Fdrale de Lausanne and Politecnico di Milano) video quality assessment database is freely available for download on.[14] It was specifically designed for the evaluation of transmission over IP network impairments.[15,16] Packet loss distortions with different percentages (0.1%, 0.4%, 1%, 3%, 5%, 10%) are simulated in this video quality assessment test database. All sequences have been encoded with the H.264/AVC encoder adopting the High Profile.

2.3 SD ROI database

This database is developed by Boulos et al. in.[17] It contains videos of 6 different source contents with for each content, 14 H.264 coding conditions with or without error transmission simulations. The specificity of this database is that the spatial position of the transmission errors depends on the Region of Interest (RoI) in the video frames. The RoI are defined using an eyetracker algorithm. Then, some slice losses are introduced in the RoI and outside of it to test the impact of both the error propagation and the spatial location of the loss on the perceived quality. When the losses were outside the RoI, they occurred in the slices adjacent to the RoI. All losses were in a single I-picture to allow a longer temporal propagation.

2.4 SVC4QoE Replace Slice database

This database is developed by Y. Pitrey et al. in.[18,19] It is designed for the evaluation of mobile transmission quality. It contains 9 contents with for each content, the reference and 14 different impairment conditions. The sequences are coded with h264 and h264/SVC codecs with simulated transmission errors. Two error concealment algorithms were tested using the h264/SVC capability: frame level concealment and pixel level concealment.

2.5 SVC4QoE Temporal Switch database

Developed by Y. Pitrey et al.[20,21] this database is designed for evaluating the impact of network behavior and encoder configuration on the visual quality using SVC-based error concealment. It contains h264 and h264/SVC encoded sequences at different Quantization Parameter (QP) values. Several switching conditions were created between the QP values in order to test the impact of temporal quality switching on the perceived quality.

2.6 Target variable

For our model, the variable that we try to predict is the subjective MOS score that we consider as the "Target variable". Since the MOS is a numerical and continuous variable, it must be discretized in order to be considered by the ML algorithm. Thus, from MOS values we associate new variable that we call "Quality" having four values:

- Excellent: if $MOS \geq 4$
- Good: if $3 \leq MOS \geq 4$
- Fair: if $2 \leq MOS \geq 3$
- Bad: if $MOS \leq 2$

We fixed this division because it is the one that gives the best balanced values distribution as shown in Figure 1

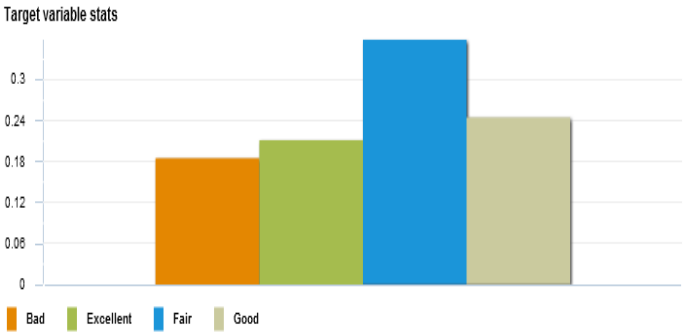
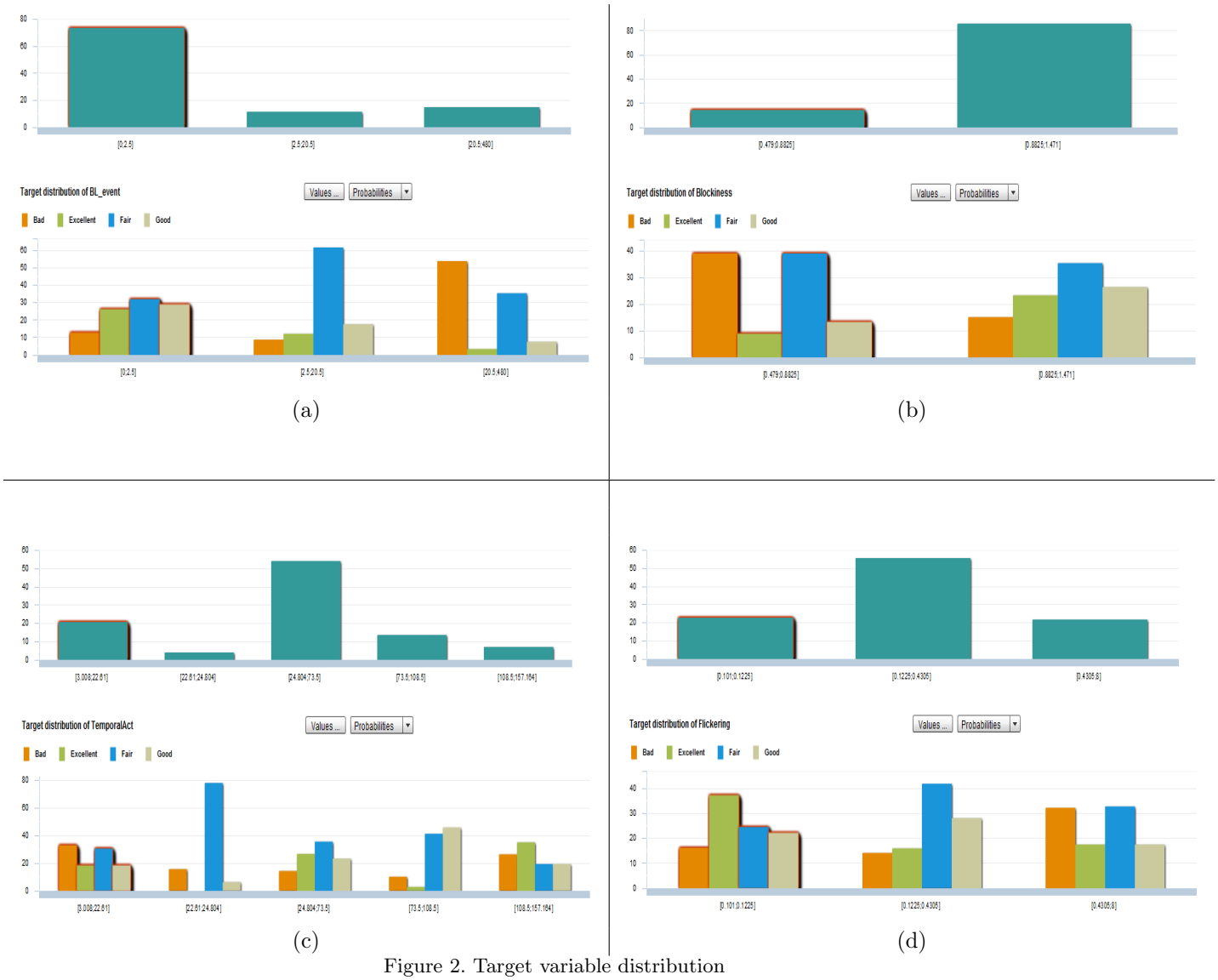


Figure 1. Target variable distribution

2.7 Outliers treatments

We consider a value as outlier when it stand out too much from the values generally observed on a variable. The process carried out concerned all the variables to remove the clearly incoherent values that are outside the range of 95% of the confidence interval. Visualization of the distribution and of the evolution of the mean of the variables, made it possible to judge the relevance of keeping or not these values in the sample. This treatment removed outliers that accounted for only less than 1% of the sample.

Before processing to the training of the ML model it is essential to have an idea on the distribution of the variables according to the Target as shown in Figure 2 (in these representations we draw our attention on the fact that the presentation of the four levels of the target variable does not follow the order of evolution of the quality.)



(a) (b)

(c) (d)

Figure 2. Target variable distribution

Table 1. Properties of subjective VQA databases

	Orange	**Live Mobile**	**EPFL**	**SD RoI**	**SVC4QoE Replace Slice**	**SVC4QoE Temporal Switch**
Nbr. of sequences	268	170	78	84	140	390
Resolution	VGA, HD 720p	HD 720p	CIF	SD (720×576)	VGA	VGA
Duration	10s to 3 min	10 s	8 to 10 s	10 s	10 s	10 s
Frame rate	30fps	30 fps	30 fps	20 fps	30 fps	30 fps
Distortion types	packet loss Jitter H.264 encoding H.265 encoding	H.264 encoding wireless packet loss frame freezes rate adaptation temporal dynamic	packet loss	Packet loss	H.264 encoding H.264/SVC encoding transmission errors	H.264 encoding
Encoder	H.264 AVC	H.264 AVC	H.264 AVC	H.264/AVC	H.264	H.264

Our training database consists in a total of 1130 data lines. Each line consists in a video sequence on which all the no reference metrics are applied and a subjective Mean Opinion Score (MOS) is associated.

3. NO REFERENCE VIDEO QUALITY METRICS

The main focus of our research work is automatic assessment of video quality in real time conversational services. In this context, it is necessary to detect a large set of distortion types. We consider the single artifact no reference MOAVI metrics. The proposed metrics estimate the presence of different video quality impairments such as Blockiness, Block loss, Blur, Noise, Flickering, etc. We judge these metrics representative of the type of degradation that may infect a video conference or a video-telephony cal. By exploring end to end transmission of a video content in a multimedia conversation stream, the artifact Key Performance Indicators (KPI) can be grouped into four categories[?](Fig. 3).

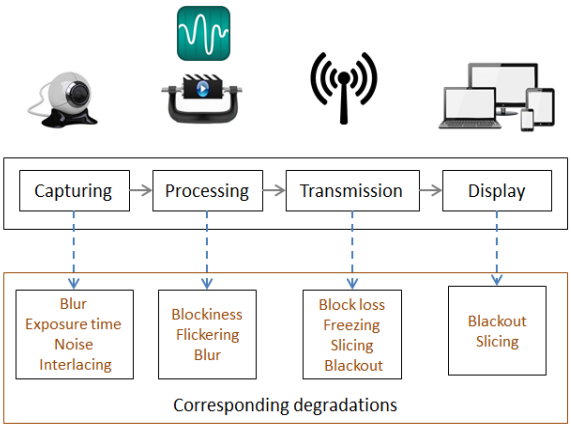


Figure 3. End-to-end transmission chain with the generated impairments

- Capturing : blur, exposure time, noise, interlacing.
- Processing: blockiness, flickering, blur.
- Transmission: blockloss, freezing, slicing, blackout.
- Display: blackout, slicing.

An evaluation of these metrics to identify the conditions under which they can be used effectively and in line with human perception in the use case for video-conferencing is performed. The results are presented in.[22]

In this study we consider also a global no reference video quality metric in order to compare its performance with the one of the single artifact based MOAVI metrics. We chose the completely blind Video Integrity Oracle VIIDEO metric[23] because its a video based metric (take into account the temporal aspect of the video) unlike other metrics that are image quality based.

4. DATA MINING TOOL

For the machine learning and data mining studies we used a software called "Khiops".[24] It integrates the work done at Orange Labs on data preparation, automatic variable construction for multi-table databases and large-scale modeling. Khiops allows to quickly perform the descriptive and explanatory phases in a Data Mining project. The database must be formatted according to a text file format, with a line per record, one header line containing the variable names and a field separator.

The first step is the specification of the data dictionary, which is the choice of the variable types (Categorical, Numerical, Date, Time or Time stamp) in the database to analyze. This dictionary is automatically built by Khiops owing to a parsing of the database file. The built dictionary is saved in a dictionary file, which basic syntax allows easy modifications. The Data Miner must then validate the variable types in the built dictionary, and eventually specify which variables to ignore in the analysis or construct new variables owing the derivation rule language.

The second step checks the correctness of the database file. In this step, Khiops parses the database file and completely checks formatting or variable type errors.

The third step, the most important one, is to analyze the predictive value of the explanatory variables or pairs of variables. In supervised analysis, when a target variable is specified, Khiops evaluates the predictive importance of any numerical or categorical explanatory variable, and of any pair of explanatory variables. Two reports, for uni-variate and bi-variate analysis, are produced at the end of the data analysis, based on the train data set. They summarize the information contained in each analyzed variable or pair of variables. In the case of supervised tasks, a scoring model is computed as well, based on a Selective Naive Bayes predictor. A modeling report summarizes the features of the built classifier or regressor. Two evaluation reports, based on the train and test data, evaluate the performance of the scoring model. New dictionaries and scoring dictionary, are produced, allowing a deployment of the scoring model.

The fourth step is the deployment step. This is done by applying the new dictionary or the scoring dictionary on new data, in order to compute score variables. This functionality can also be used to construct any new variable, described using the derivation rule language.

5. SELECTIVE NAIVE BAYES MODEL: OBTAINING A GLOBAL VIDEO QUALITY SCORE

In our case, the variable to predict is the "Quality" metric defined above. Given the categorical nature of this target variable, we have the choice between a number of ML prediction methods, such as decision trees, random forests, and so on. Our choice is the Selective Naive Bayes (SNB) method because of:

- its simplicity,
- it is adapted to large volumes of data,
- its good performance often rented in publications,[25]
- it is implemented in the software (Khiops) that we used.

As input for the ML algorithm we consider all the MOAVI metrics. The SNB algorithm defines the variables that are the most related to the MOS scores through an indicator called "Level". The level represents the evaluation of the predictive importance of the variable. It is a value between 0 (variable without predictive interest) and 1 (variable with optimal predictive importance). Figure 4 shows the distribution of the level values of our variables. The most correlated variable with subjective scores in our database is clearly Block loss event.

In Khiops data-mining tool we fixed 70% of the database used for training and 30% for testing. The samples are chosen randomly by the algorithm. The table presented in Figure 2 shows the predictor evaluation on the test and training samples. The SNB classifier is evaluated using the following criteria:

- Accuracy: evaluates the proportion of correct prediction.
- Compression: evaluates the predicted target probabilities using a negative log likelihood approach and is normalized (between 0 and 1) using the baseline predictor.
- AUC: area under the ROC curve (AUC) which evaluates the ordering of the predicted scores per target value.

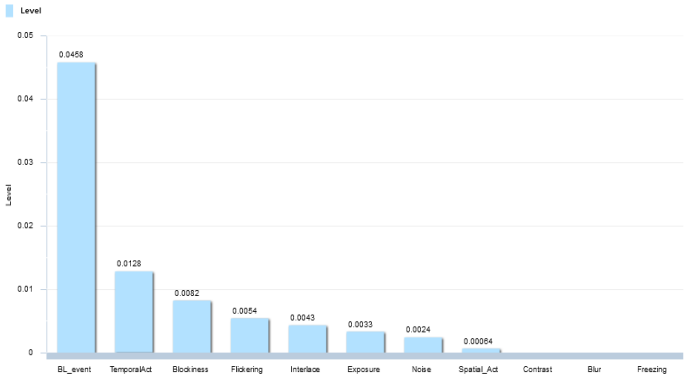


Figure 4. Level distribution

Table 2. Predictor evaluation

Name	Type	AUC	Compression	Accuracy
Selective Naive Bayes	Train	0.7077	0.1341	0.5082
Optimal	Train	1	1	1
Selective Naive Bayes	Test	0.6915	0.0917	0.4438
Optimal	Test	1	1	1

For our generated model we have 0.44 of accuracy, 0.09 for compression and 0.69 AUC which corresponds not to a fine prediction. According to these evaluation indicators, the generated model is not accurate for video quality assessment.

A confusion matrix is reported for the classifier, to compare the predicted values (prefixed by $) and the actual values ones. As shown in table 3, for Bad and Good values, the model gives a correct prediction in 70% of the cases. However, for Fair and Excellent the model gives a correct prediction in less than 50% of the cases. This can be explained by the fact that Fair and Good classes are close to each other and over-represented in our database.

Table 3. Confusion matrix

Target	%Bad	%Excellent	%Fair	%Good
%Bad	70	0	17.5	12.5
%Excellent	14.5	42.4	21.74	21.2
%Fair	13.7	17	48	21
%Good	0	16.42	13.4	70.1

Moreover, the cumulative gain curve, drawn in Figure 5, evaluates the quality of the model. The green curve corresponds to the results of the SNB model applied on the test sample database. The purple one corresponds to an optimal model. The black curve corresponds to the worst model, that is to say the one that is equivalent to a random choice of the class.

Based only on MOAVI single artifact based metrics it is shown that the ML approach generate a model that is not accurate in predicting the global video quality. Thus, we have the idea to add another no-reference metric to the training variables which is VIIDEO. This metric will bring information on the global quality of

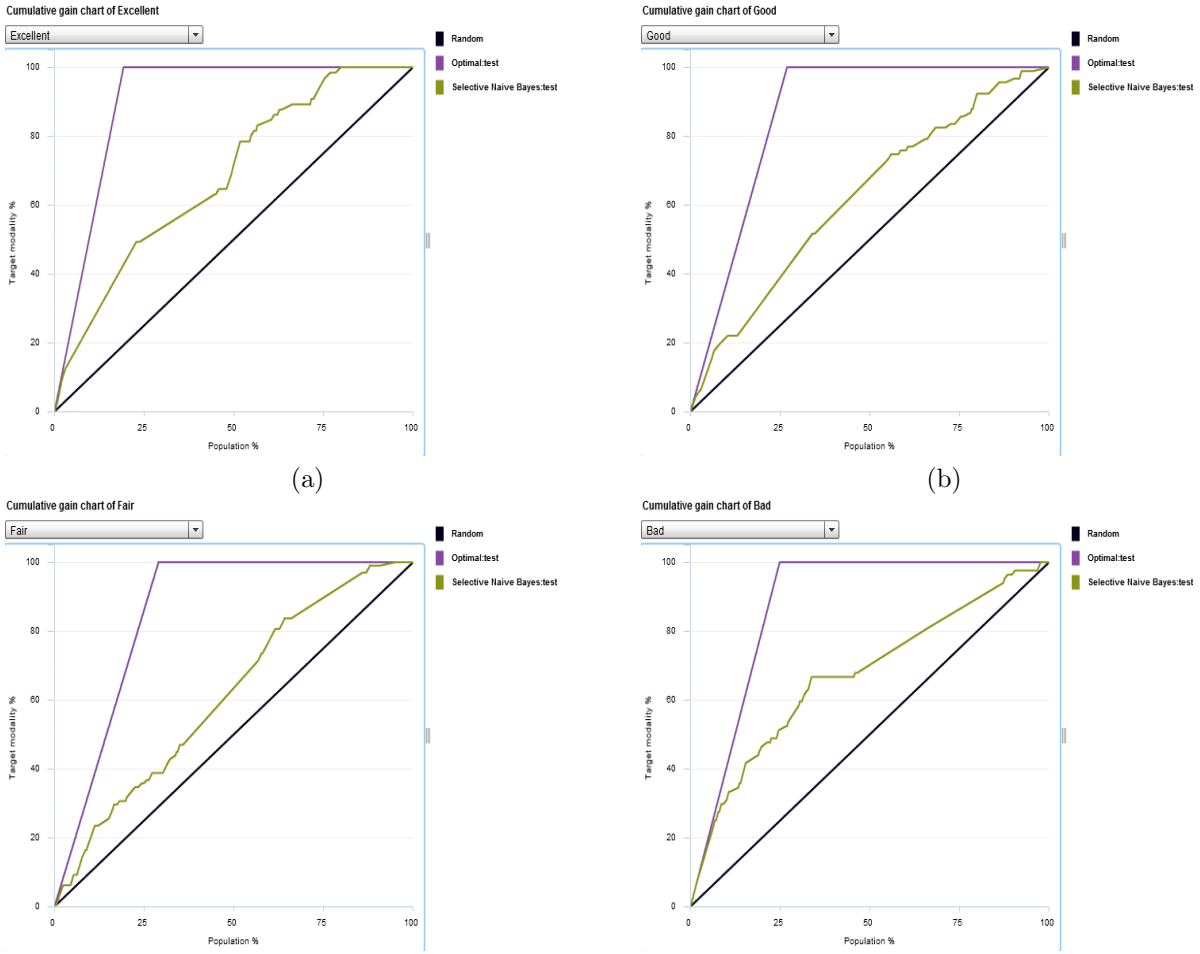


Figure 5. Cumulative gain curve for Excellent (a), Good (b), Fair (c) and Bad (d) quality classes

the sequence (not dedicated for a specific distortion) that could enhance the performance of the prediction model.

We apply the same methodology as described above, we add the VIIDEO metric values for all our 1130 sequences and we re-run the training and testing processes. The evaluation of the generated prediction model presented in Table 4 shows clearly that the accuracy of the model is improved 0.618.

Table 4. Predictor evaluation after adding VIIDEO metric

Name	Type	AUC	Compression	Accuracy
Selective Naive Bayes	Train	0.8038	0.3739	0.6350
Optimal	Train	1	1	1
Selective Naive Bayes	Test	0.8097	0.3597	0.618
Optimal	Test	1	1	1

Table 5. Confusion matrix

Target	%Bad	%Excellent	%Fair	%Good
%Bad	91.72	1.27	5.73	1.27
%Excellent	0	52.56	23.72	23.7
%Fair	0.48	18.90	54.78	25.84
%Good	0	6.90	13.79	79.31

The new confusion matrix presented in Table 5 shows that the new model makes less error in quality prediction compared to the one trained only on MOAVI metrics. For Fair and Excellent classes the model gives more that 50% of correct prediction. For the Good class it gives 79.31% of correct prediction. For the Bad class it reaches 91% of correct prediction which is particularly interesting because for a monitoring and diagnostic tool it is important to detect a Bad quality when there is a problem more than to detect a good quality.

The good performance of the model is confirmed by the cumulative gain curves shown in Figure 6. The green curves corresponding to the SNB predictor are close to the optimal predictor.

6. CONCLUSION

In this chapter we investigate the possibility of combining no-reference single artifact metrics taken from MOAVI in a global video quality assessment model. The obtained model has an accuracy of only 0.44 which is not enough for a good model. After adding no reference VIIDEO metric to the training variables of the ML algorithm, the model is enhanced and reach 0.63 of accuracy. This result is encouraging because we consider that even if our database contains only 1130 sequences, this volume allowed to generate a promising prediction model. We recommend to collect more databases with more diversified conditions.

REFERENCES

[1] Moorthy, A. K., Choi, L. K., Bovik, A. C., and De Veciana, G., "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing* **6**(6), 652–671 (2012).

[2] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing* **19**(6), 1427–1441 (2010).

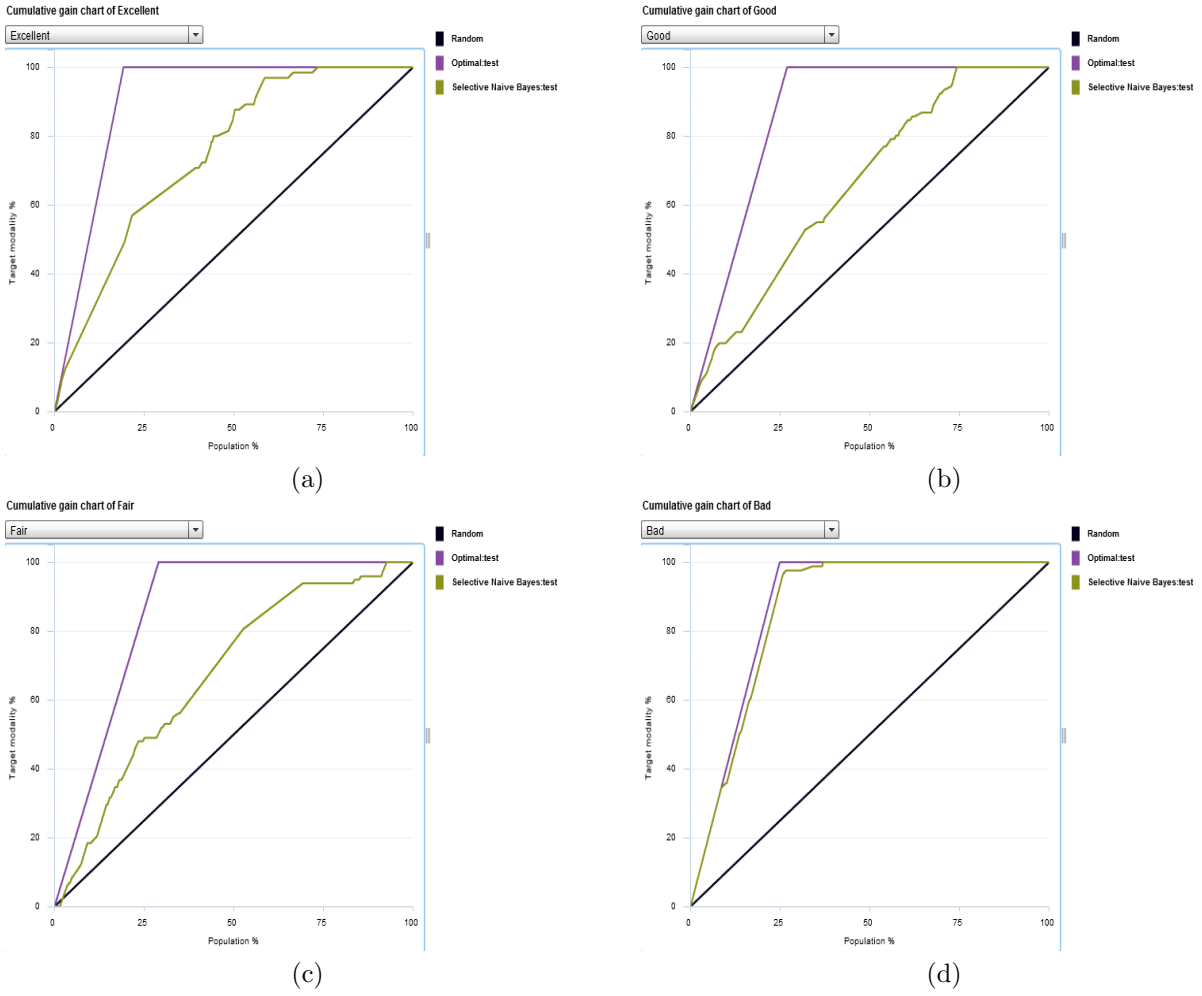


Figure 6. Cumulative gain curve for Excellent (a), Good (b), Fair (c) and Bad (d) quality classes

[3] Martínez-Rach, M. O., Piñol, P., López, O. M., Perez Malumbres, M., Oliver, J., and Calafate, C. T., "On the performance of video quality assessment metrics under different compression and packet loss scenarios," *The Scientific World Journal* **2014** (2014).
[4] Winkler, S., [*Digital video quality: vision models and metrics*], John Wiley & Sons (2005).
[5] Wu, H. R. and Rao, K. R., [*Digital video image quality and perceptual coding*], CRC press (2005).
[6] Zlokolica, V., Kukolj, D., Lukic, N., and Temerinac, M., "Evaluation on the selection of video quality metrics for overall visual perception," in [*Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*], 23–28, IEEE (2010).
[7] "No reference metrics." http://vq.kt.agh.edu.pl/metrics.html. [Online; accessed 17-January-2017].
[8] Saidi, I., Zhang, L., Barriac, V., and Deforges, O., "Interactive vs. non-interactive subjective evaluation of ip network impairments on audiovisual quality in videoconferencing context," in [*Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*], 1–6, IEEE (2016).
[9] Saidi, I., Zhang, L., Barriac, V., and Deforges, O., "Audiovisual quality study for videotelephony on ip networks," *Proc. MMSP* (2016).
[10] BT-500, I.-T. R., "Methodology for the subjective assessment of the quality of television pictures," (2012).
[11] P.910, I.-T. R., "Subjective video quality assessment methods for multimedia applications," (April 2008).
[12] "Laboratory for image and video engineering." http://live.ece.utexas.edu/.
[13] Moorthy, A. K., Choi, L. K., De Veciana, G., and Bovik, A. C., "Mobile video quality assessment database," *IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks* (2012).
[14] "Epfl database." http://vqa.como.polimi.it/.
[15] De Simone, F., Tagliasacchi, M., Naccari, M., Tubaro, S., and Ebrahimi, T., "A h. 264/avc video database for the evaluation of quality metrics," in [*Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*], 2430–2433, IEEE (2010).
[16] De Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., and Ebrahimi, T., "Subjective assessment of h. 264/avc video sequences transmitted over a noisy channel," in [*Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*], 204–209, IEEE (2009).
[17] Boulos, F., Chen, W., Parrein, B., and Le Callet, P., "Region-of-interest intra prediction for H.264/AVC error resilience," in [*Image Processing (ICIP), 2009 16th IEEE International Conference on*], 3109–3112, IEEE (2009).
[18] Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., and Le Callet, P., "Aligning subjective tests using a low cost common set," in [*Euro ITV*], irccyn contribution (June 2011).
[19] Pitrey, Y., Barkowsky, M., Le Callet, P., and Pépion, R., "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," in [*SPIE Optical Engineering*], (Aug. 2010).
[20] Pitrey, Y., Engelke, U., Le Callet, P., Barkowsky, M., and Pépion, R., "Subjective quality of svc-coded videos with different error-patterns concealed using spatial scalability," in [*Third European Workshop on Visual Information Processing (EUVIP)*], paper number 67 (July 2011).
[21] Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., and Le Callet, P., "Aligning subjective tests using a low cost common set," in [*Euro ITV*], irccyn contribution (June 2011).
[22] Saidi, I., Zhang, L., Barriac, V., and Deforges, O., "Evaluation of single-artifact based video quality metrics in video communication context," in [*Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*], 1–3, IEEE (2017).
[23] Mittal, A., Saad, M. A., and Bovik, A. C., "A completely blind video integrity oracle," *IEEE Transactions on Image Processing* **25**(1), 289–300 (2016).
[24] "Khiops data mining tool web site." https://khiops.predicsis.com/.
[25] Domingos, P. and Pazzani, M., "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine learning* **29**(2), 103–130 (1997).