# OPTIMIZING FEATURE POOLING AND PREDICTION MODELS OF VQA ALGORITHMS

*Kongfeng Zhu*[⋆] *, Marcus Barkowsky*[†]*, Minmin Shen*[⋆] *, Patrick Le Callet*[†]*, Dietmar Saupe*[⋆]

[⋆]Department of Computer and Information Science, University of Konstanz, Germany
[†]LUNAM Universite de Nantes, IRCCyN UMR CNRS 6597, Nantes, France

## ABSTRACT

In this paper, we propose a strategy to optimize feature pooling and prediction models of video quality assessment (VQA) algorithms with a much smaller number of parameters than methods based on machine learning, such as neural networks. Based on optimization, the proposed mapping strategy is composed of a global linear model for pooling extracted features, a simple linear model for local alignment in which local factors depend on source videos, and a non-linear model for quality calibration. Also, a reduced-reference VQA algorithm is proposed to predict the local factors from the source video. In the IRCCyN/IVC video database of content influence and the LIVE mobile video database, the performance of VQA algorithms is improved significantly by local alignment. The proposed mapping strategy with prediction of local factors outperforms one no-reference VQA metric and is comparable to one full-reference VQA metric. Thus predicting the local factors in local alignment based on video content will be a promising new approach for VQA.

***Index Terms***— Video quality assessment, feature pooling, non-linear mapping, local alignment, reduce reference

## 1. INTRODUCTION

Given that increasingly knowledgeable users demand better quality image and video acquisition and display, it is highly desirable to automatically and accurately predict visual signal quality aligned with the perceived and reported quality by these users. Video quality assessment (VQA) aims to automatically, accurately and unbiasedly predict the visual quality of an image or a video in alignment with humans. With partial or no prior knowledge about the pristine video, reduced-reference (RR) or no-reference (NR) VQA is useful but can hardly predict visual quality as accurately as full-reference (FR) VQA. Thus it is of great importance to improve the performance of RR/NR-VQA algorithms.

An NR-VQA algorithm, in general, follows a two-stage framework, in which the two stages are distortion measurement and mapping to subjectively evaluated quality. The distortion measurement quantifies the features of distortion in distorted data by measuring artifacts [1, 2, 3], by analyzing parameters in the video bitstream [4], or by comparing the statistical properties with those of natural images/videos [5, 6]. At the second stage, a mapping transforms the extracted features of the distortion measurement to a single number, the predicted video quality. Neural networks are often used for non-linear mapping in such a way that a number of extracted features for distortion measurement are input to the neural network and its output is taken as the predicted quality of the videos or images [6, 7, 8].

Both stages are of great importance for the design of NR-VQA. However, it has been found that there has been less research focused on the mapping than on the distortion measurement. This is not surprising, because the interpretation of the content and quality of an image or a video by the human visual system (HVS) depends on high-level features, such as attentive vision, cognitive understanding, and prior experience viewing similar patterns, hence the relation of the distortion measurement to the perceptual quality is extremely complicated. Methods based on machine learning usually take the relation as a black box and provide a decent solution when the database for training contains a large number of videos with various video contents and all types of distortion. If the database is small, the method is prone to overfitting, resulting in poor performance in general [9]. Unfortunately, the existing public video databases are too small for training-based methods to obtain a robust non-linear function.

The focus of this paper is on optimizing feature pooling and the prediction model of a VQA algorithm. Rather than taking the non-linear mapping as a black box, we intend to decompose it into several linear or non-linear functions with only a few parameters, in order to circumvent the intractable problem of overfitting in methods based on machine learning. In the proposed mapping strategy, we study the mismatch of existing objective assessments to the subjective assessments, pool a number of extracted features for distortion measurement by a global linear function, then perform a local alignment for each video set[1] based on a local linear function, we finally apply a non-linear function for quality calibration. A RR-VQA algorithm is also proposed to predict the two local factors. With few parameters and one feature of the reference, the proposed RR model provides a good quality prediction.

[1] A video set refers to a group of videos in a video database for subjective assessment. They are generated from one source video with different levels of distortion.

The rest of this paper is organized as follows. In Sec. 2, we state the problem of mapping from distortion measurements to perceptual quality. An optimized mapping strategy and an algorithm that predicts the local factors are proposed in Sec. 3, and the experimental results are reported in Sec. 4. Finally we conclude in Sec. 5.

## 2. STATEMENT OF THE MAPPING PROBLEM

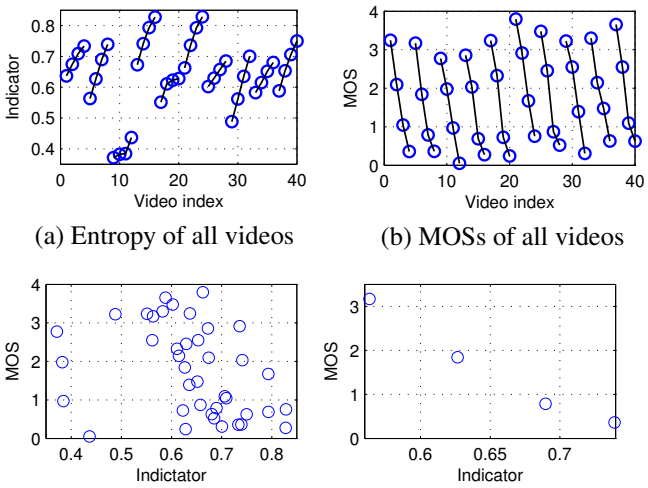
In practice, the mapping in this two-stage framework is designed to map/match the extracted features for a distortion measurement with the corresponding mean opinion score (MOS) of a subjective assessment. Understanding the reasons for any divergence between the subjective results and the objective predictions is necessary in order to improve the objective quality assessment. For the distortion measurement, we extract features as in an NR-VQA algorithm that was proposed based on Laplacian decomposition [6]. For each video sequence, we extract three intra-subband features: energy, entropy, and kurtosis; and also extract three inter-subband features: the Jensen–Shannon divergence, the structural similarity index between two subbands, and the smoothness.

Figures 1 (a) and (b) give an example of objective and subjective assessments in the LIVE mobile video database [10]. There are ten video sets with four compressed videos in each one. The $x$-axis is the video index and the $y$-axis is the distortion indicator, which is entropy in (a) and MOS in (b). The entropy of the videos in a video set differs significantly, depending on their content. In contrast to the entropy, the MOSs of all the video sets roughly start at the same value and are uniformly distributed in the same interval. The scatter plot of the entropy and the MOS in Fig. 1 (c) for all videos shows no linear correlation. However, a very clear linear correlation is observed in the scatter plot of the entropy and the MOS for individual video sets, as shown for the second video set in Fig. 1 (d). The same phenomenon is observed for all other features (indicators). Therefore, the poor performance of the indicator in Fig.1 (c) is related to the variance of the source content over all video sets.

The accuracy of quality prediction can be significantly improved if the entropy in each video set is aligned individually to the corresponding MOS via a simple linear function. Moreover, only an offset and a scale for one video set are required for this alignment. In this paper, we analyze the influence of specific video properties, such as the content, on the performance of a VQA algorithm during the parameter training or optimization. As this corresponds to local adaptations of the fitting function, we will call the associated fitting coefficients *local factors*.

## 3. THE PROPOSED MAPPING STRATEGY

Based on the study in Sec. 2, we decompose the stage of the mapping into linear combination, local alignment, and quality correction. We also propose to predict the local factors first, then the MOS in subjective assessment. The diagram of the proposed mapping strategy is illustrated in Fig. 2. The dashed box, which is considered as a black box in machine-learning based methods for pooling features and predicting MOSs with a large number of parameters, is the focus of this paper. The proposed mapping strategy with a few parameters and the prediction of local factors are presented in the following.




(a) Entropy of all videos (b) MOSs of all videos

(c) Scatter plot of all videos (d) Scatter plot of the 2nd set

**Fig. 1**. Example of mismatch between objective and subjective assessment in LIVE mobile video database

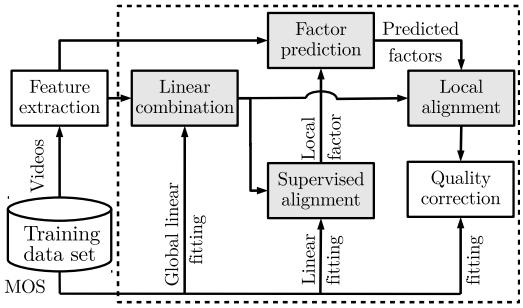



**Fig. 2**. Diagram of the proposed mapping strategy

### 3.1. Mapping model

For the $n$th video in the $m$th video set, let us denote by $\hat{y}(m,n)$ its predicted quality. The six extracted video-level features are denoted by $f_i(m,n)$, $i = 1,\ldots,6$, where $m = 1,\ldots,M, n = 1,\ldots,N$, and $M$ and $N$ are, respectively, the total number of video sets in the database, and the number of videos in one set. The proposed non-linear mapping strategy includes a global linear function, a local alignment, and a quality calibration, described as follows.

**Global linear function** combines the six video-level features with a linear function

$$y'(m,n) = \Sigma_i w_i f_i(m,n), \tag{1}$$

where $w_i$ is the weight of the $i$th feature. The linear mapping is applied to the videos, so we consider it as a *global* factor.

**Local alignment** is a local linear function within one video set. For a set of videos with the same content but different levels of distortion, the locally aligned objective video quality is

$$y''(m,n) = s(m)y'(m,n) + o(m), \tag{2}$$

where $s(m)$ and $o(m)$ are the scale and offset of the $m$th set of videos, respectively. Note that both $s(m)$ and $o(m)$ are determined separately for each video set, thus they are *local* factors, depending on the video content.

**Quality calibration** uses a non-linear function to align the objective quality measurement with the subjective quality. It is known that subjective results are not linear for the whole quality range, e.g., subjective ratings saturate before they reach the extremes [9]. For this reason, a non-linear regression is proposed to align the prediction. The common logistic function $g(\cdot)$ transforms the aligned objective video quality $y''$ to the predicted quality [11],

$$\hat{y}(m,n) = g(y''(m,n)), \tag{3}$$

$$g(x) = \frac{\beta_1 - \beta_2}{1 + \exp(-(x-\beta_3)/|\beta_4|)} + \beta_2. \tag{4}$$

### 3.2. Prediction of local factors

A supervised alignment is applied to obtain the two parameters $s(m), o(m)$ of each video set. We propose to predict the two local parameters of each video set rather than predicting the corresponding MOS directly. As shown above, the two local parameters for a video set are determined by the content of the source video, hence analyzing the source video is the first option. Based on the study in video databases, we find the local parameters of a source video are strongly correlated with its structural complexity, namely, entropy. Thus, we propose the following algorithm to predict the two local parameters [2].

1) Prediction of the offset. A linear relation exists between offsets and scales, obtained by supervised alignment (see Fig. 4), so the offset can be predicted from the scale by

$$\hat{o}(m) = a_1 s(m) + a_0, \quad m = 1, \ldots, M. \tag{5}$$

2) Prediction of the scale. The video-level feature $f_0 = H_0(X)/H_3(X)$, called the entropy ratio of the source video, is found to be highly related with the scale. $H_0$ and $H_3$ are the entropies of two Laplacian subbands, averaged over all frames of the video [6]. The relation is modeled by a third-order polynomial, and the predicted scale value is

$$\hat{s}(m) = q\left(f_0(m)\right), \quad m = 1, \ldots, M, \tag{6}$$

$$q(x) = \alpha_3 x^3 + \alpha_2 x^2 + \alpha_1 x + \alpha_0. \tag{7}$$

[2] A more detailed discussion of the relation between the offsets and the scales, and the reason for their linear relation here, will be discussed in more detail in a later paper.

**Algorithm 1** Reduced-reference video quality prediction

**Require:** $f_0$, $a_1$,$w_1, \ldots, w_5$, $\alpha_0, \ldots, \alpha_3$, $\beta_0, \ldots, \beta_3$

1: $a_0 = 0$ and $w_6 = 1 - \Sigma_{i=1}^{5} w_i$
2: compute $f_1, \ldots, f_6$
3: $s = q(f_0) \Leftarrow \alpha_0, \ldots, \alpha_3$
4: $y = s \cdot \left(\Sigma_{i=1}^{6} w_i f_i + a_1\right) + a_0 \Leftarrow a_1, w_1, \ldots, w_5$
5: $\hat{y} = p(y) \Leftarrow \beta_1, \ldots, \beta_4$

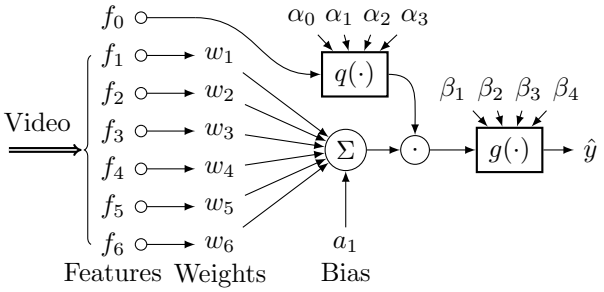



**Fig. 3**. Flow chart of the RR-VQA algorithm

All the parameters are determined after the optimization. We set $w_6 = 1 - \Sigma_{i=1}^{5} w_i$ and $a_0 = 0$ to reduce the redundancy of offsets in Eq. 1 to Eq. 7. The RR-VQA algorithm based on the proposed mapping strategy and the prediction of local parameters are summarized in Algorithm 1, and the corresponding flow chart in Fig. 3. Only one scalar feature $f_0$ (8 bits) of the source video is required. Its bits are so few that it can be easily embedded in the video by watermarking.

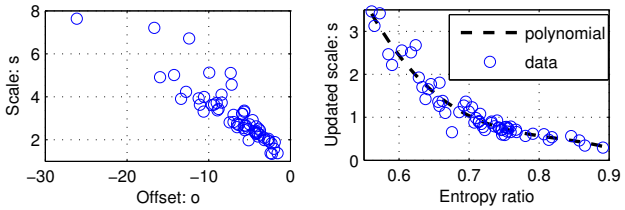
## 4. EXPERIMENTAL RESULTS

In this section, we demonstrated how the local alignment of Sec. 3 improves the performance of a VQA algorithm. We tested the proposed process of an optimization strategy, and predicted the local factors in the local alignment. We carried out experiments on the IRCCyN/IVC video database (available in [12]) with various video contents, in which there are $M = 60$ video sets with $N = 5$ videos in each one, and the associated subjective results [13]. The experiment was designed as follows.

Step 1: Feature extraction. We extract the six features mentioned in Sec. 2 for each video sequence as in [6].

Step 2: Supervised local alignment. The linear least squares fitting technique is first used to create the global linear model of MOSs to the six extracted video-level features in Eq. 1. Then the local linear model of MOSs within each video set is created, see Eq. 2. The two local parameters $s$ and $o$ are plotted in Fig. 4(a). In Fig. 5, we compare the scatter plots of the predicted MOSs and the MOSs before and after the local alignment. The performance is improved significantly after the local alignment.

Step 3: Prediction of local parameters. The scatter plot of scale and offset and scatter plot of scale and entropy for all video set are illustrated in Fig. 4. Obviously, there is a clear linear relation between the scale and the offset, and




(a)Scales vs. offset (b) Scales vs. entropy ratio

**Fig. 4**. Prediction of local parameters

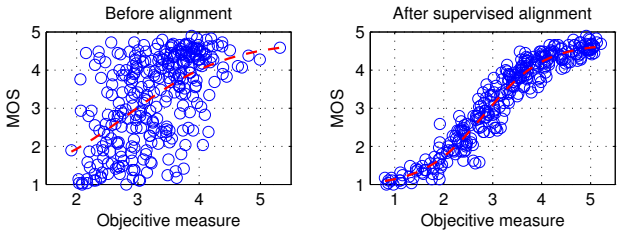



**Fig. 5**. Scatter plots of MOS vs. predicted MOS

a non-linear relation between the scale and the entropy ratio. The linear least squares fitting technique is again used to create a linear model of the scale to the offset in Eq. 5. Then the unconstrained optimization based on the basic quasi-Newton method [14] determines the parameters of the third-order polynomial in Eq. 7.

Step 4: One-step optimization. The unconstrained non-linear optimization is executed according to Eqs. 1–7 in a combined step optimizing all parameters at once. The initial values of all parameters are set to the values we obtained in the above steps, otherwise, it might become trapped in a local minimum. The number of parameters in the mapping model is 14, which is less than 5% of the number of videos in the database, thus overfitting is circumvented in the proposed model, in contrast to the multilayer neural network adopted in [6] with more than 150 parameters. The optimized parameters for the IRCCyN/IVC video database are listed in Table 1. They are suitable for VQA using Alg. 1 for videos of the same resolution and degradation type as in the training database.

To numerically evaluate and compare the performance of the proposed algorithm with other VQA metrics, we employed four statistical indexes: Pearson's correlation coefficient (LCC), Spearman's rank ordered correlation coefficient (SROCC), the root mean squared error (RMSE), and the mean absolute error (MAE), between the predicted MOS and the MOS. The performance comparison was done for the

**Table 1**. Optimized parameters for IRCCyN/IVC VD

| | | | | |
|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ |
| 0.2068 | 0.6474 | 0.0108 | -0.0237 | 0.0974 |
| $a_1$ | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
| -0.1939 | 53.608 | -81.354 | 17.499 | 18.903 |
| $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | |
| 4.7432 | 1.3946 | 3.3246 | 0.1373 | |

**Table 2**. Performance comparison in IRCCyN/IVC VD

| Metrics | LCC | SROCC | RMSE | MAE |
|---|---|---|---|---|
| NR-VQA | 0.4372 | 0.4207 | 1.1052 | 0.8571 |
| Proposed | **0.8474** | **0.8148** | **0.6138** | **0.4771** |

**Table 3**. Performance comparison in LIVE MVD

| Metrics | LCC | SROCC | RMSE | MAE |
|---|---|---|---|---|
| MOVIE | 0.8103 | 0.7738 | 0.6674 | — |
| NR-VQA | 0.5722 | 0.5794 | 1.0765 | 0.8748 |
| Proposed | **0.7960** | **0.7955** | **0.7049** | **0.5572** |

IRCCyN/IVC video database (IRCCyN/IVC VD) and LIVE mobile video database (LIVE MVD) [10].

We performed 200 runs for IRCCyN/IVC VD. For each run, we randomly chose half of the video sets for training and the other half for testing. The median values of the four indexes of all tests are given in Table 2. In the LIVE mobile video database, a similar strategy was applied to 40 compressed videos (10 sets). In each run, five video sets were for training and the other five sets for testing. The process was repeated $\binom{10}{5}$ times on $\binom{10}{5}$ train-test pairs. Table 3 illustrates the performance comparison with NR-VQA [6] and one FR-VQA metric [15]. The results of MOVIE [15] were taken from [10].

## 5. CONCLUSION

A non-linear mapping strategy was proposed in the paper to circumvent the overfitting problem in machine-learning based methods, while maintaining accuracy, efficiency and consistency. After studying the mismatch between distortion measurements and the perceptual quality, the extracted features for the distortion measurement were first combined via a linear function, then a local alignment within each individual video set was designed to improve the performance. Finally, the 4-parameter logistic function was adopted for the quality calibration. The local parameters in the local alignment were predicted by analyzing the content of the source video, for example, the structural complexity.

The experimental results on the IRCCyN/IVC influence content video database indicated that this local alignment improved the performance of the NR-VQA method significantly. Only 14 parameters and one scalar feature (8 bits) of the source video were required in the proposed RR-VQA algorithm. It is comparable to one FR-VQA metric in one database and outperforms one NR-VQA in two databases, with many fewer parameters. In conclusion, we believe that objectively identifying missing indicators by searching for (local) influence factors, such as content, may be a promising approach for training NR-VQA models. The prediction of local factors without reference is the subject of further research. This approach can also be extended by identifying the type of perceptually distinguishable artifacts.

## 6. REFERENCES


[1] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi, "A no-reference perceptual blur metric," in *Proceedings of 2002 International Conference on Image Processing*, 2002, vol. 3, pp. 57–60.

[2] Zhou Wang, Alan Conrad Bovik, and Brian L. Evans, "Blind measurement of blocking artifacts in images," in *International Conference on Image Processing*, Vancouver, BC, Canada, 2000, vol. 3, pp. 981–984.

[3] Rémi Barland and Abdelhakim Saadane, "A new reference free approach for the quality assessment of MPEG coded videos," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, Antwerp, Belgium, Sept. 2005, vol. 3708, pp. 364–371.

[4] Tomás Brandão and Maria Paula Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1437–1447, Nov. 2010.

[5] Rajiv Soundararajan and Alan Conrad Bovik, "Survey of information theory in visual quality assessment," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 391–401, May 2013.

[6] Kongfeng Zhu, Keigo Hirakawa, Vijayan Asari, and Dietmar Saupe, "A no-reference video quality assessment based on Laplacian pyramid," in *IEEE International Conference on Image Processing*, Melbourne,Australia, Sept. 2013.

[7] Dubravko Ćulibrk, Dragan Kukolj, Petar Vasiljević, Maja Pokrić, and Vladimir Zlokolica, "Feature selection for neural-network based no-reference video quality assessment," in *International Conference on Artificial Neural Networks*, Limmassol, Cyprus, Sept. 2009, pp. 633–642.

[8] Kongfeng Zhu, Vijayan Asari, and Dietmar Saupe, "No-reference quality assessment of H.264/AVC encoded video based on natural scene features," in *Mobile MultimediaImage Processing, Security, and Applications, SPIE Defense, Security, and Sensing*, Baltimore, Maryland, US, May 2013, vol. 8755(4).

[9] Christian Keimel, Tobias Oelbaum, and Klaus Diepold, "Improving the verification process of video quality metrics," in *International Workshop on Quality of Multimedia Experience*, San Diego, CA, US, July 2009, pp. 121 – 126.

[10] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[11] Ann Marie Rohaly, Philip Corriveau andJohn Libert, Arthur Webster, Vittorio Baroncini, John Beerends, Jean-Louis Blin, Laura Contin, Takahiro Hamada, David Harrison, Andries Hekstra, Jeffrey Lubin, Yukihiro Nishida, Ricardo Nishihara, John Pearson, Antonio Franca Pessoa, Neil Pickford, Alexander Schertz, Massimo Visca, Andrew Watson, and Stefan Winkler, "Video quality experts group: Current results and future directions," in *Proc. SPIE 4067, Visual Communications and Image Processing 2000*, Perth, Australia, May 2000, pp. 742–753.

[12] "IRCCyN/IVC video database of content influence," `http://www.irccyn.ec-nantes.fr/spip.php?article771&lang=en`.

[13] Yohann Pitrey, Marcus Barkowsky, Romuald Pépion, Patrick Le Callet, and Helmut Hlavacs, "Influence of the source content and encoding configuration on the preceived quality for scalable video coding," in *SPIE Human Vision and Electronic Imaging XVII*, Jan. 2012.

[14] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering. Springer, second edition, 2006.

[15] Kalpana Seshadrinathan and Alan Conrad Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.