

EVALUATION OF VIDEO QUALITY FLUCTUATIONS USING PATTERN CATEGORISATION

Clemens Horch, Julian Habigt, Christian Keimel and Klaus Diepold

Technische Universität München, Institute for Data Processing,
Arcisstr. 21, 80333 Munich, Germany
ch@tum.de, jh@tum.de, christian.keimel@tum.de, kldi@tum.de

ABSTRACT

Fluctuations of video quality over time can have a significant influence on the overall perceived quality as represented by the QoE. Existing methodologies for subjective video quality assessment, however, are often not suitable for the evaluation of these quality fluctuations, especially if they occur within very small time frames. In this contribution, we therefore propose a new method, VIQPAC, which addresses this shortcoming by using a pattern categorisation approach. Instead of requiring the subjects to provide a continuous quality evaluation, the subjects assess the overall quality impression and the strength of the quality fluctuation, combined with a categorisation of the encountered fluctuation pattern. This allows us to determine the fluctuation dependent temporal changes in the quality. The results show that VIQPAC is able to capture the pattern and strength of quality fluctuations, allowing for a proper description of the temporal quality changes within a video sequence.

Index Terms— Subjective video quality assessment, video quality, subjective testing, quality fluctuations, QoE, VIQPAC.

1. INTRODUCTION

The visual quality of video often fluctuates over time, influencing the overall Quality of Experience (QoE) of a video sequence. Temporal fluctuations can be caused by an encoder's rate control, but also during the streaming of video by changes in the available bitrate and subsequent switching between representations of different quality e.g. with MPEG DASH or, more generally, by time-variant transmission errors and the subsequent time-variant degradation of the received video due to fluctuating packet loss ratios. These fluctuations usually have a high frequency and do not occur over a longer time period i.e. the quality doesn't slowly change over a few minutes, but occurs rather quickly on a time scale of fractions of seconds.

QoE is commonly represented by the overall mean opinion score (MOS). The MOS, however, only provides a temporally pooled representation of the quality variation over time, when often it is necessary to also evaluate the temporal quality fluctuations, particularly to address the underlying causes of these fluctuations. For assessing such fluctuations, usually the only standardised (temporally) continuous quality assessment method, the Single Stimulus Continuous Quality Evaluation (SSCQE) defined in ITU-R BT.500 [1], is frequently used. Compared to non-continuous methods, the subjects do not rate the quality after watching the complete sequence, but use a slider to express their evaluation in real-time while watching the video sequence, with a typical sequence lasting between 20 and 30 minutes. Although this method allows to determine the temporal fluctuations of visual quality, the results are often more complicated

to analyse compared to non-continuous methods due to different reaction times or context effects [2]. Also Pinson and Wolf [3] noted that on average, the test subject need 6 s to adapt the slider position to a new quality level and therefore especially fast changes in quality might be difficult to obtain with SSCQE. Gauss et al. [4] used SSCQE for determining temporal quality changes of video sequences caused by packet loss by combining and repeating shorter video sequences with a length of 30 s in one 30 minute video, thus aiming to compensate for reaction time and context. Still they encountered several problems with respect to accuracy of the test results and only after an advanced selection process, resulting in the rejection of more than 50 % of the test subjects, the results were considered valid. Considering these shortcomings of SSCQE especially with respect to assessment of short term quality fluctuations, a method capable of describing these types of fluctuations is needed.

In this contribution we therefore introduce a new method, *Video Quality evaluation with PAttern Categorisation – VIQPAC*, aiming at addressing the shortcomings described above by using a pattern categorisation approach. Instead of requiring the subjects to provide a continuous quality evaluation as in SSCQE, the subjects assess the overall quality impression and the strength of the quality fluctuation, combined with a categorisation of the encountered fluctuation pattern. This allows us to determine the fluctuation dependent temporal changes in the quality even on a time scale below the reaction threshold of 6 s as observed by Pinson and Wolf [3].

This contribution is organised as follows: after presenting the design and rationale behind the proposed method, we briefly describe the setup of the subjective test conducted to verify the proposed method, before discussing the results of the verification test. We then compare the results from the verification test with the results gained with an objective video quality metric capable of detecting temporal quality fluctuations before concluding with a short summary.

2. EVALUATION OF VIDEO QUALITY FLUCTUATIONS USING PATTERN CATEGORISATION

In this section we describe the pattern categorisation approach behind the proposed VIQPAC method that allows the evaluation of video quality fluctuations for even rather short time frames.

2.1. Overall Test Design

The basic idea behind VIQPAC is to divide the quality assessment task into three separate tasks: in the first task, the participants provide their overall quality impression on a continuous scale using a slider, similar to the SSCQE scale in ITU-R BT.500 [1]. To avoid

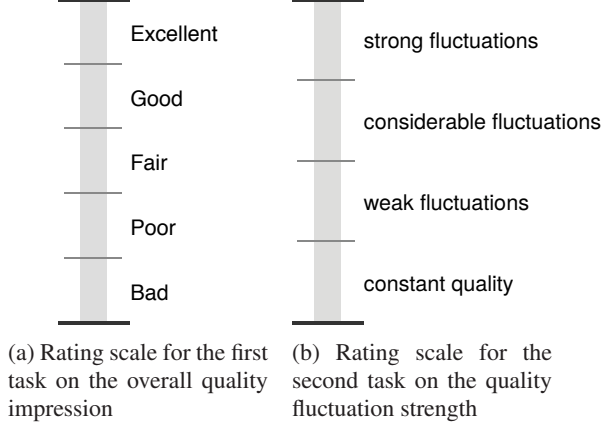


Fig. 1: Rating scales for temporal quality assessment

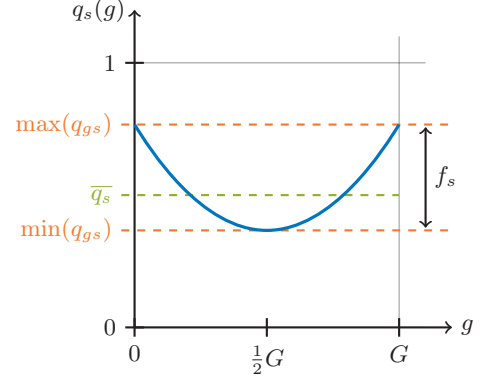
any further complexity, a single stimulus presentation of the video sequences was chosen in this contribution, but a double stimulus presentation could also be utilised. Similarly, the continuous scale can be replaced with discrete scale. In the second task, the subjects rate the strength of the quality fluctuation during the video sequence on a continuous scale from 'constant quality' to 'strong changes' as illustrated in Fig. 1. In the third task, the test subjects identify the overall quality fluctuation pattern during the sequence: each of the six categories is represented by a small qualitative plot as shown in the right column of Table 1. The first five categories were designed using polynomial functions with increasing degrees from 0 to 2 and the sixth category is representative for sequences with fast, but noticeable quality fluctuations that do not fit into any of the first five categories. The results from these three evaluation tasks can then be combined, providing not only an indication of the overall fluctuation pattern, but also the magnitude of the fluctuation pattern, thus categorising the quality pattern.

The major advantage of this approach is that the subjects can categorise the temporal quality fluctuations even for short video sequences. One drawback is, however, that it is not possible to provide arbitrary patterns for any fluctuation pattern that may be encountered and the longer the sequences get, the more complex the quality progression pattern may become. In general, more patterns would allow a better representation of possible fluctuations, but would lead to a more cognitive demanding assessment task for the subjects as they would be required to distinguish between a large number of different fluctuation patterns. The six patterns proposed in Table 1 are therefore a reasonable compromise, leading to good results as we will see in Section 4.

Since the assessment now consists of three simultaneous tasks and the subjects therefore face a higher cognitive workload, the test subjects can repeat each videos as often as required, similar to the interactive approach taken in SAMVIQ [5]. Also this requires a more thorough training phase before the actual test compared to traditional video quality assessment methods, as the additional tasks for identifying the quality fluctuation pattern and corresponding fluctuation strength need to be explained properly to the subjects in order to gain valid results.

2.2. Pattern Categorisation

In order to evaluate the temporal quality fluctuations, the fluctuation is categorised using the overall quality impression, the fluctuation



(a) Pattern 4: parabola, open at top

Fig. 2: Example of the function representing the categorised fourth pattern and the relationship between the function and the two constraints.

strength and the selected pattern chosen by test subjects in the first, second and third task, respectively. The categorised pattern is then expressed as a function that allows us to provide a continuous quality rating, incorporating the overall nature of the temporal quality variation within the given time frame i.e. the complete video subsequence.

Without loss of generality, we assume in this contribution a fixed sampling interval of the quality fluctuation patterns which is equal to the length of one group of pictures (GOP). Usually a GOP in most (broadcast-related) H.264/AVC applications lasts about 0.5 s, providing a sufficiently granular sampling rate. Hence the temporal direction of the video is quantised into a number of equally sized intervals, dependent on the size of the GOP.

The quality rating of the g -th GOP of the s -th test subject q_{gs} is calculated in dependency of the overall quality impression \bar{q}_s determined in the first task and corresponding to the MOS, the perceived strength of the quality fluctuation f_s as determined in the second task and the fluctuation pattern selected in the third task. The categorised pattern functions q_{gs} for the different basic patterns as illustrated in Table 1 are constructed with respect to the following two constraints: firstly, the distance between the maximum and minimum q_{gs} equals the fluctuation strength f_s

$$f_s = \max(q_{gs}) - \min(q_{gs}) \quad (1)$$

and secondly, the average of all q_{gs} per sequence equals the overall quality rating \bar{q}_s of the s -th subject







$$\bar{q}_s = \frac{1}{G} \sum_{g=0}^G q_{gs}, \quad (2)$$

where G denotes the number of GOPs in the video sequence and S the number of test subjects. These two constraints on the chosen functions are illustrated for the fourth fluctuation pattern in Fig. 2. Hence we obtain for each subject s and GOP g a quality rating q_{gs} . We then average over all subjects to gain one quality rating per GOP q_g as

$$q_g = \frac{1}{S} \sum_{s=0}^S q_{gs}. \quad (3)$$

All G q_g of the G GOPs are then combined into the overall categorised quality pattern $q(g)$, representing the quasi-continuous video

Table 1: Fluctuation patterns and corresponding functions to determine the categorised fluctuation patterns

Pattern	Function of categorised pattern	Icon
1 constant	$q_{gs} = \overline{q_s}$	
2 linear increasing	$q_{gs} = \overline{q_s} - \frac{f_s}{2} + \frac{f_s}{G} \cdot g$	
3 linear decreasing	$q_{gs} = \overline{q_s} + \frac{f_s}{2} - \frac{f_s}{G} \cdot g$	
4 parabola, open at top	$q_{gs} = \frac{4f_s}{G^2} \cdot g^2 - \frac{4f_s}{G} \cdot g + \overline{q_s} + \frac{2f_s}{3}$	
5 parabola, open at bottom	$q_{gs} = -\frac{4f_s}{G^2} \cdot g^2 + \frac{4f_s}{G} \cdot g + \overline{q_s} - \frac{2f_s}{3}$	
6 oscillating	$q_{gs} = \frac{f_s}{2} \cos(g) + \overline{q_s}$	

quality evaluation of a video sequence. Fig. 4 shows an example of how the individual categorised quality patterns $q_s(g)$ of each subject result in the overall quality pattern $q(g)$.

3. VERIFICATION TEST SETUP

In order to verify the proposed assessment method, we conducted a subjective test with the VIQPAC method described in the previous section.

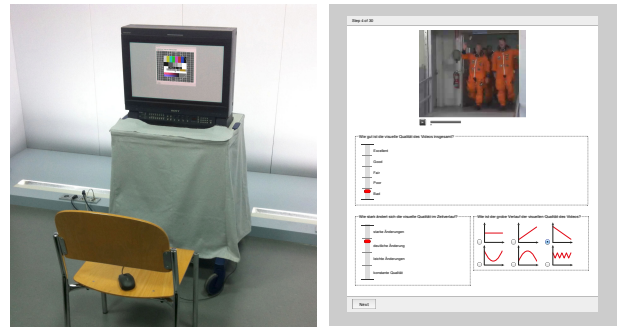
3.1. Test Conditions

The subjective test was conducted in the ITU-R BT.500 [1] compliant video quality assessment laboratory at the Institute for Data Processing (LDV) at the Technische Universität München (TUM) shown in Fig. 3a. The video sequences were presented using a *Sony BVM-L230* reference display (screen diagonal 23") at a visible height of 8 cm and with a viewing distance of about 60 cm. $S = 21$ persons between 14 to 28 participated in the test, most of them students, but not experts in video processing.

For presenting the video and recording the subjects' ratings, we used the *QualityCrowd2* [6] framework to provide an interactive user interface for all three assessment tasks in the test. Fig. 3b shows a screenshot of the user interface: a player including a button to replay the video, the overall quality impression scale for task one, the quality fluctuation strength scale for task two and lastly the selection of the perceived overall quality fluctuations patterns for task three. For better comprehensibility, the quality fluctuation strength scale was labelled in German, as the test subjects were mostly native German speakers. On average a subject finished the test in 15.9 minutes: 4.1 minutes for the training phase, 1.2 minutes for a (hidden) stabilization stabilisation phase, and 10.7 minutes for the test phase itself.

3.2. Video Sequences

As the aim of VIQPAC is to enable the assessment of quality fluctuations, an important requirement on the video sequences to be used in the method's verification are visible quality changes over time. Several datasets were examined by a group of expert viewers and it was decided to use the IT-IST dataset by Brandão and Queluz [7] as it exhibited clearly visible quality fluctuations. The videos in this dataset consist of sequences in CIF resolution 352×288 pixels at 30 fps encoded with H.264/AVC using a fixed GOP length of 15 frames. We removed the first GOP and last GOP as both were incomplete i.e. had less than 15 frames, resulting in 240 frames per sequence and $G = 16$ GOPs per sequence. We selected a subset of the five sequences *City*, *Football*, *Foreman*, *Table* and *Tempete*, at four different bit rates each, covering a wide range of MOSs and quality fluctuation, resulting in $N = 20$ test conditions.



(a) Test environment at TUM (b) Screenshot QualityCrowd2

Fig. 3: ITU-R BT.500 compliant test environment (left) and interface of the *QualityCrowd2* software used in the test (right)

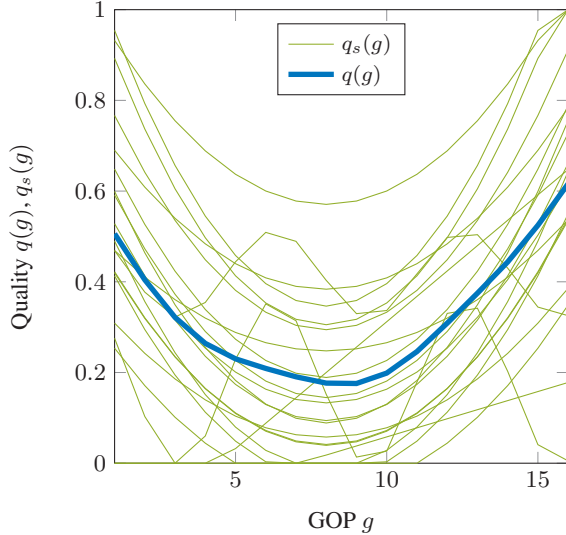


Fig. 4: Categorized quality patterns for the sequence *Football* at 256 kBit/s: both for individual subjects represented by $q_s(g)$ and the average for all subjects $q(g)$

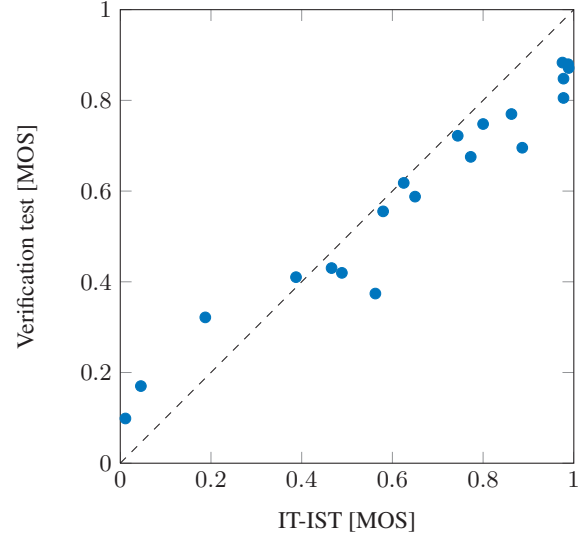


Fig. 5: Scatter plot of the overall quality impression \bar{q}_s from the IT-IST data set from [7] compared to the results from the verification test

4. RESULTS OF THE VERIFICATION TEST

Before discussing the overall results represented by the categorised quality pattern, we discuss the results of the three evaluation tasks that are necessary to gain the overall categorised quality pattern.

4.1. Overall Quality Impression

In the first step, we compared the overall quality impression expressed by \bar{q}_s to the MOS provided in [7] by determining the Pearson correlation, Spearman rank order correlation and RMSE between the corresponding ratings for each test condition as shown in 2 and Fig. 5. For all video sequences, the Pearson correlation exceeds 0.95, the threshold proposed by VQEG [8] as required inter-test correlation for considering the results from two separate tests equivalent. This result not only shows that the setup of the verification test allows us to obtain valid results for the overall video quality, but also indicates that both the interactive nature of the test setup and the two additional tasks do not distract the subjects from the judgement of the overall video quality.

Table 2: Comparison between the results from the verification test and the results provided in [7]

Sequence	Pearson	Spearman	RMSE
City	0.970	1.000	0.112
Football	0.996	1.000	0.091
Foreman	0.997	1.000	0.098
Table	0.997	1.000	0.144
Tempete	0.998	1.000	0.072
all	0.976	0.959	0.106

4.2. Quality Fluctuation Strength

For the second task, the results can not be validated as easily as for the first task as the results in the IT-IST data set do not contain continuous quality ratings. However, we can still assess the plausibility of the quality fluctuation strength f_s . In Fig. 6, the average rating of f_s per sequence is shown and we can notice that the fluctuation strength is strongly correlated with the subjective quality with a Pearson correlation of 0.79 i.e. the higher the quality, the lower the strength of the observed fluctuations. Fig. 6, however, also shows a high variance of the fluctuation strength assessment. This may be explained by the fact that this task was perceived by many subjects to be the most complex of the three tasks. In particular the subjects tended to start the assessment of each test case by first performing the first task, the overall quality assessment, and the third task, the identification of the quality fluctuation pattern, before assessing the fluctuation strength, often only after a repeated viewing of the video.

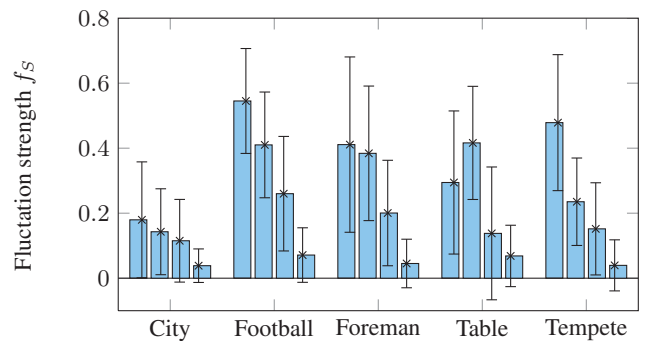








Fig. 6: Average and standard deviation of the quality fluctuation strength for all sequences. Each bar corresponds to one of four different rate points, from the lowest (left) to the highest (right) bitrate

Table 3: Percentages of the results for the third task: quality fluctuation patterns. Rate point (RP) 1 corresponds to the lowest bit rate, RP 4 to the highest. Depending on the sequence and bit rate, clearly dominating quality patterns emerge.

Sequence	RP						
City	1	71.4	4.8				23.8
	2	66.7	14.3	4.8			14.3
	3	66.7	9.5	4.8			19.0
	4	100.0					
Football	1		9.5		81.0		9.5
	2	4.8	9.5		66.7	9.5	9.5
	3	23.8	9.5		66.7		
	4	76.2		4.8	19.0		
Foreman	1	19.0		76.2	4.8		
	2	14.3	9.5	14.3	42.9	4.8	14.3
	3	47.6	4.8	14.3	28.6		4.8
	4	90.5		4.8	4.8		
Table	1	33.3		33.3	4.8		28.6
	2	4.8	57.1		23.8		14.3
	3	71.4	9.5		9.5		9.5
	4	81.0	9.5		9.5		
Tempete	1	4.8		4.8			90.5
	2	33.3	4.8	4.8			57.1
	3	66.7	4.8	14.3			14.3
	4	90.5			4.8		4.8

4.3. Quality Fluctuation Pattern

Similar to the second task, the results from the third task can only be checked for plausibility. Table 3 shows the pattern selected most often per sequence and the corresponding percentage. One can see in this table that in 17 out of 20 cases, more than half of all subjects agreed on one pattern; in 15 out of these 20 cases it was even more than two thirds. This indicates that the six patterns provided in VIQPAC were a reasonable choice, as the test subjects were able to express the perceived quality fluctuations quite well.

4.4. Continuous Quality Rating

Lastly, we need to evaluate whether the continuous quality ratings represented by the quality ratings per GOP q_g and the resulting overall categorised quality fluctuation pattern $q(g)$ is valid. The quality rating q_g is the result of the evaluation of the functions described in

Table 4: Correlation coefficients and RMSE between categorised quality pattern $q(g)$ and predicted quality $\hat{q}(g)$ for all G GOPs

Sequence	Pearson	Spearman	RMSE
City	0.989	0.963	0.142
Football	0.871	0.877	0.149
Foreman	0.942	0.944	0.126
Table	0.963	0.963	0.085
Tempete	0.947	0.965	0.103
all	0.893	0.898	0.124

Table 1 with the overall quality impression \bar{q}_s , the quality fluctuation strength f_s and the selected quality fluctuation pattern, followed by the averaging over all subjects according to (3).

Due to the fact that existing subjective video quality evaluation methods are not able to assess the visual quality fluctuation on the required small time scale, we decided to use the visual quality predictions from the no-reference H.264/AVC bitstream-based video quality metric proposed by Horch et al. [9]. It is based on the multi-way data analysis approach suggest by Keimel et al. [10], but extends it to a more granular prediction. In particular, it is able to provide not only an overall quality prediction for a complete video sequence, but also a quality prediction per GOP.

Fig. 8 shows the categorised quality fluctuation patterns for two rate points of the *Football* and *Tempete* sequences as examples. We can see that the quality rating $q(g)$ and the quality predicted by the video quality metric $\hat{q}(g)$ share the same overall categorised quality fluctuation pattern. Clearly, the quality predictions $\hat{q}(g)$ provides more details as the metric is not limited to one of the six patterns as the test subjects are in the third task. Nevertheless, the overall shape is very similar.

Comparing the quality ratings gained with VIQPAC and the quality prediction gained with no-reference metric from [9] for all sequences and each GOP with the Pearson correlation, the Spearman rank order correlation and RMSE, we can see in Table 4 and the corresponding scatter plot in Fig. 7 that the results are very similar. In particular, except for the sequence *Football* all correlation

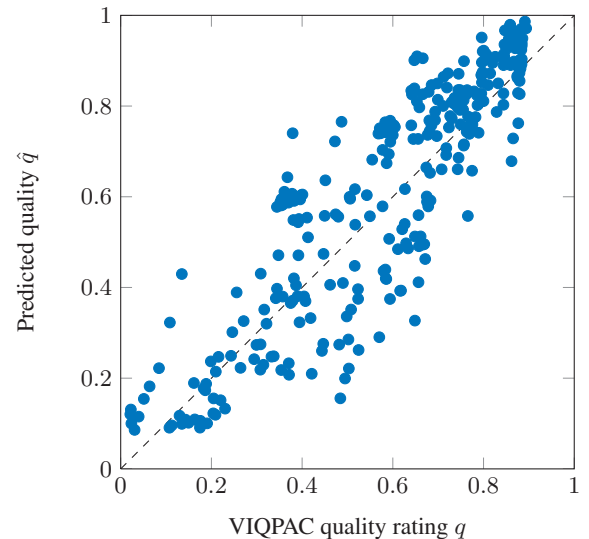


Fig. 7: Scatter plot of quality ratings gained with VIQPAC and the quality prediction gained with the no-reference metric from [9] for each GOP

coefficients are well above 0.9 and the RMSE for all sequences is below 0.15. The comparably worse performance of VIQPAC for the sequence *Football* is due the fast motion present in the sequence, resulting in strong bitrate and consequently quality fluctuations. Still, even considering all sequences, the correlation coefficients and the RMSE are still quite high.

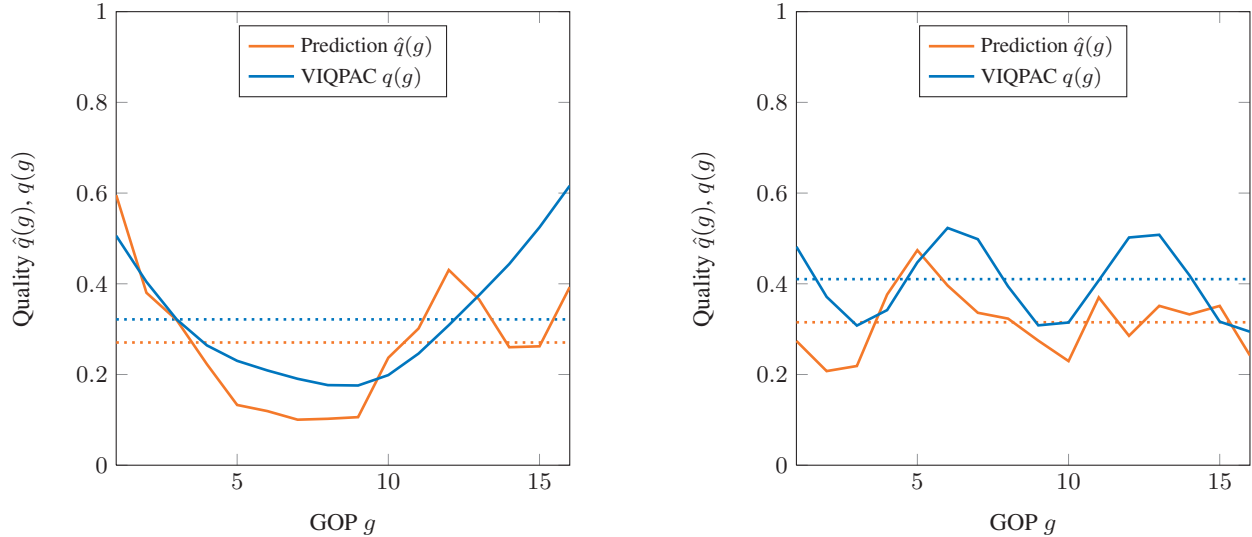


Fig. 8: Categorized quality patterns determined using VIQPAC and quality predictions gained with the no-reference metric from [9] for the sequence *Football* (left) and *Tempeste* (right) from the IT-IST dataset. The dotted lines represent the corresponding means

5. CONCLUSION

In this contribution we presented VIQPAC, a new subjective quality assessment method that allows the continuous video quality fluctuation assessment even for comparably short video sequences. Instead of requiring subjects to directly assess the quality fluctuations in real-time, we split up the overall evaluation into three separate tasks that can be executed consecutively, resulting in a categorized quality pattern, representative of the temporal quality changes.

The results of the verification test show that not only is the proposed VIQPAC method able to reproduce the overall MOS of a sequence gained with other subjective testing methodologies, but also to reproduce continuous quality ratings for comparably short time frames and with a high sampling rate, so far not assessable with existing testing methodologies. In future work, we intend to verify the VIQPAC with other data sets, discrete rating scales and double stimulus presentation of the test cases.

6. REFERENCES

- [1] ITU-R, *Recommendation BT.500: Methodology for the subjective assessment of the quality of television pictures*, International Telecommunications Union, Radiocommunication Sector, Jan. 2012.
- [2] S. Winkler, *Digital Video Quality – Vision Models and Metrics*, Wiley, 2005.
- [3] M. H. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies,” in *Proceedings of the SPIE Conference on Video Communications and Image Processing*, June 2003, vol. 5150, pp. 573–582.
- [4] S. Gauss, T. Muller, J. Wuenschmann, and A. Rothermel, “Continuous subjective quality evaluation of terrestrial broadcast video,” in *Proceedings of the 2011 IEEE International Conference on Consumer Electronics (ICCE 2011)*, Sept. 2011, pp. 356–360.
- [5] ITU-R, *ITU-R BT.1788: Methodology for the subjective assessment of video quality in multimedia applications*, Jan. 2007.
- [6] C. Keimel, J. Habigt, C. Horch, and K. Diepold, “Qualitycrowd – a framework for crowd-based quality evaluation,” in *Proceedings of the 2012 Picture Coding Symposium (PCS 2012)*, May 2012, pp. 245–248.
- [7] T. Brandão and M.P. Queluz, “No-reference quality assessment of H.264/AVC encoded video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437–1447, Sept. 2010.
- [8] Video Quality Experts Group (VQEG), “Report on the validation of video quality models for high definition video content,” Tech. Rep., June 2010.
- [9] C. Horch, C. Keimel, J. Habigt, and K. Diepold, “Length-independent refinement of video quality metrics based on multiway data analysis,” in *Proceedings of the 2013 IEEE International Conference on Image Processing*, Sept. 2013, pp. 44–48.
- [10] C. Keimel, M. Rothbucher, Hao Shen, and K. Diepold, “Video is a cube,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 41–49, Sept. 2011.