Hierarchical Approach for No-Reference Consumer Video Quality Assessment

Jari Korhonen, *Member, IEEE*

***Abstract*—Nowadays, smartphones and other consumer devices capable for capturing video content and sharing it on social media in nearly real time are widely available for a reasonable cost. This is why there is also a growing need for No-Reference Video Quality Assessment (NR-VQA) of consumer produced video content, typically characterized by types of impairments that are fundamentally different from those observed in professionally produced video content. To date, most of the NR-VQA models in the prior art have been developed for assessing coding and transmission impairments, rather than capture artifacts. In addition, the most accurate known NR-VQA methods are computationally complex, making them impractical for many real life applications. In this paper, we proposed a new approach for learning-based video quality assessment, based on the idea of computing the features hierarchically so that low complexity features are computed for the full sequence first, and then high complexity features are extracted from a subset of representative video frames, selected by using the low complexity features. We have experimented the proposed approach against relevant benchmark methods using several recently published annotated public video quality databases, and our results show that the proposed approach outperforms the best performing benchmark methods in terms of quality prediction accuracy and complexity.**

***Index Terms*—Image quality, video signal processing, visual communications, quality management.**

I. Introduction

During the past few years, the rapid development of mobile camera technology and social media applications has turned regular users from passive consumers into active producers and distributors of visual information, including user generated video content. However, even though the digital camera technology has evolved substantially within only few years, the quality of user generated video content is still severely constrained by the technical limitations of the mobile consumer camera technology, as well as the users' limited knowledge and abilities in video content production. This is why user generated video content often suffers from types of quality impairments that are rarely observed in professional video content, such as camera shakiness, under- and overexposure, sensor noise, color distortions etc.

As the amount of user generated content in the Internet is growing, the number of advanced and technically savvy users and consumers is also growing. For example, there are a lot of videobloggers with relatively large bases of followers, making the distinction between professional content and user generated content more fluid. This growing segment of users tend to be devoted in continuously improving the content they produce, not limited to the interestingness, but also the technical quality of their videos. Since it is often challenging to assess technical video quality accurately directly by observing the preview version of the video, especially on a small screen of a smartphone or digital camera, those consumers would highly benefit from instant feedback on the quality produced by algorithm-based No-Reference Video Quality Metrics (NR-VQM). Another obvious example of a potential application for NR-VQM is the automatic quality labeling of consumer videos on content sharing platforms. Technical quality is directly related to the usability of several types of consumer content, such as instruction videos for cooking, repairing things yourself, athletic performances etc., and therefore such quality labeling would be a useful feature for many users.

Up to date, most of the related work in NR-VQA has focused on compression and transmission artifacts [1-12], and the target application of most of the practical NR-VQMs proposed in the prior art is quality monitoring in video streaming and digital broadcasting systems. According to the recent studies, the accuracy of the legacy NR-VQMs is not satisfactory when assessed using the recent consumer video quality datasets with natural distortions [13-16]. Better results have been achieved by applying state-of-the-art No-Reference Image Quality Metrics (NR-IQM) to consumer videos frame by frame; however, the computational complexity of the most accurate NR-IQMs is too high for many practical applications of NR-VQA on mobile consumer devices [15-16]. This is why there is an apparent need for new low complexity NR-VQMs that can accurately assess also typical capture artifacts in amateur videos, validated with datasets containing relevant user generated content.

In this paper, we present a proposed architecture for an NR-VQM to tackle the challenges of assessing consumer video quality. The proposed approach is based on the idea of computing the quality features hierarchically in several steps: only low complexity spatiotemporal features are computed for every frame, and more complex spatial features are only computed for a subset of representative frames. Our experimental results show that the proposed approach is able to

This paper was submitted for review on 31. July, 2018. This work was supported by the National Science Foundation of China under Grant 61772348.

J. Korhonen is with Shenzhen University, College of Computer Science and Software Engineering, Shenzhen City, Guangdong Province, 518060, P.R. China (e-mail: jari.t.korhonen@ieee.org).

predict the subjective video quality more accurately than the legacy NR-VQMs known from the prior art, with substantially lower computational complexity than the most accurate benchmark methods.

The rest of the paper is organized as follows. In Section II, we discuss the related work concerning subjective video quality assessment, relevant video quality datasets, as well as the prior art for objective NR-VQMs. In Section III, we describe the proposed approach for NR-VQM in details, including feature extraction, temporal pooling of features, as well as learning-based regression for predicting MOS from the features. In Section III, we present the results of an extensive validation study on three public video quality datasets. In Section IV, we discuss the practical opportunities, limitations and possible future directions concerning the proposed approach. Finally, the concluding remarks are given in Section V.

II. Background and Related Work

Traditionally, research work on video quality assessment has comprised two clearly distinguished yet mutually related fields: subjective video quality assessment and objective video quality modeling. Since subjective video quality assessment typically requires substantial resources in terms of time and labor, there is a fundamental need for objective video quality models for many practical applications, such as real-time monitoring of video quality in streaming applications and automatic quality labeling in video content sharing platforms. However, subjective video quality assessment campaigns are still essential for creating the datasets with video sequences and the related subjective quality data that can be used as a ground truth for developing and validating the objective quality models.

A. Subjective Video Quality Assessment and Datasets

Subjective video quality assessment has been an active area of scientific research during the past twenty years, with several different methodologies developed and standardized [17-20]. The early work in the field focused mostly on assessing the severity of compression and transmission distortions [21-24]. In this case, it is possible to choose between single stimulus methods, where each test sequence is rated by test users directly, or double stimulus methods, where the quality impairment is rated in comparison with the unimpaired reference sequence. Different rating scales can be used: the most common methods are Absolute Category Rating (ACR), where users assign each test sequence to one of the pre-defined categories (for example, a five point ACR scale from "poor" to "excellent") or continuous scale, where users assign each test sequence a numerical score on a wider scale (for example, 1-100). When double stimulus methods are used, relative rating scales would be applied, respectively.

When consumer videos with capture artifacts are concerned, double stimulus methods cannot be used, since there are no unimpaired reference sequences available. The choice of methodology is therefore essentially the choice between different rating scales. In practical studies, both ACR and continuous scales have been used [21]. Another choice is between a lab-based study and a crowdsourcing study. During the past few years, crowdsourcing studies in the Internet have become popular, since it is easier and cheaper to recruit a large number of test users for an Internet-based study than for a lab-based study conducted in a specific fixed location [14,25]. On the other hand, crowdsourcing studies also have some specific problems. First, there is little control over the test environment, since every test person does the experiment using their own computer and Internet connection. Second, many users will not be fully focused on the task and some may be even intentionally cheating. Therefore, the results need to be processed more carefully than lab-based studies to detect and exclude unreliable users [23].

The first subjective video quality datasets were published around 2000 and they were focused on compression artifacts in digital television [21]. Since that, several video quality databases have been published, mostly containing compression noise, transmission distortion or a combination of them [22-24]. However, the first datasets including natural distortions, such as capture noise and camera shake, have been published relatively recently. The pioneering dataset with natural distortions was Consumer Video Database (CVD2014) [13], followed by KoNViD-1k [14] and LIVE-Qualcomm [15] datasets.

The main characteristics of CVD2014, KoNViD-1k and LIVE-Qualcomm datasets are summarized in Table I. Each database have their strengths and weaknesses. KoNViD-1k is by far the largest dataset, containing 1,200 video sequences with a substantial content diversity, ranging from user generated video clips to professionally produced video content, animation, screen content video and even time-lapse photography. CVD2014 and LIVE-Qualcomm datasets concern user generated video content more specifically. CVD2014 covers a wide range of capture devices with different resolutions and frame rates, whereas LIVE-Qualcomm dataset has been produced using a few different mobile phones, all with Full HD resolution (1920x1080 pixels) and similar frame rate. Subjective scores for the KoNViD-1k dataset have been collected in a crowdsourcing study, whereas CVD2014 and LIVE-Qualcomm datasets were produced in a traditional lab-based study.

In terms of content diversity, KoNViD-1k is by far the largest dataset, since the content is collected from 480 unique authors from a publicly available video database. On the other hand, one can argue that many content types present in the dataset, e.g. animation and screen content videos, are not highly relevant for practical applications of consumer video quality models. Comparing CVD2014 and LIVE-Qualcomm datasets, the variety of devices and impairment types is wider in the CVD2014 dataset, whereas the number of scenes is larger in the LIVE-Qualcomm dataset. On the other hand, many of the scenes in LIVE-Qualcomm dataset are relatively similar, showing outdoor activities in daylight. All the three databases have somewhat different emphasis, and their usability is therefore dependent on the target application and context.

TABLE I
PUBLIC CONSUMER VIDEO QUALITY DATABASES COMPARED: CVD2014, KONVID-1K AND LIVE-QUALCOMM

Dataset characteristic	CVD2014 [13]	KoNViD-1k [14]	LIVE-Qualcomm [15]
Number of test videos	234	1200	208
Video resolution	640x480, 1280x720	960x540	1920x1080
Video frame rate	9-30 frames/sec	23-29 frames/sec	30 frames/sec
Video length	11-28 sec	8 sec	15 sec
Number of scenes	5	1200	54
Number of devices	78	>164	8
Test methodology	Lab-based	Crowdsourcing	Lab-based
Number of test subjects	27-33 (six different experiments)	642 (min. 50 per video)	39
Rating scale	Continuous 1-100	Absolute Category Rating (1-5)	Continuous 1-100
Audio track included	Yes	Yes	No
Main strength	Realistic consumer content with a large number of devices and impairment types.	Very wide diversity of contents and distortion types. Large number of test users.	Realistic consumer content with smartphones. Uses Full HD resolution.
Main weakness	Some methodological inconsistencies between experiments. Small number of different scenes.	Some contents and distortions have little practical relevance to NR-VQA. Test methodology prone to unreliable individual scores.	Large number of scenes, but different scene types not very well balanced. Only smartphones used as camera.
Remarks	Six different experiments. In some of the experiments, other information collected aside of video quality (audio quality, contrast, blurriness etc.)	Some contents in the database are clipped from the original.	Additional information collected concerning the dominating distortion type of each test sequence.

B. Objective Video Quality Models

Since subjective video quality assessment requires substantial resources, there is a significant interest in algorithm-based objective video quality models. Traditionally, objective video quality models have been classified into Full Reference (FR), Reduced Reference (RR) and No Reference (NR) models [26]. FR models use the undistorted version of the video signal as a reference, and this is why they are best suited for assessing the performance of video compression methods. In video streaming and digital broadcasting applications, the receiver does not have the reference video available, and this is why FR methods cannot be used for monitoring the video quality at the receiver side. RR models use features extracted from the original video as a reference signal. Since the reference features can be transmitted as a side information along the video signal with a small overhead, RR models are more appropriate for networking applications than FR models. NR models rely solely on the features extracted from the analyzed test video sequence. When capture artifacts are concerned, only NR models are applicable, since the unimpaired reference sequence does not exist in the first place.

FR and RR video quality models have reached a mature state already [27-30], and the main challenge in the objective video quality assessment lies in the research and development of NR models. In the recent literature, a large number of different NR-VQA models have been proposed. NR-VQA models can be divided into bitstream-based, pixel-based and hybrid models [26]. Bitstream-based quality models use the features extracted directly from the encoded video sequence, such as quantization parameter, coding bitrate, motion vectors etc. [1-3,6]. In a networked video scenario, bitstream can include also information about transport stream, such as packet losses. Pixel-based models use the raw video signal solely as an input [4-5,7-10,12], and hybrid models combine features from the bitstream with features extracted from the decoded video signal [11,26]. Bitstream-based models and hybrid models are useful for monitoring video quality in video streaming and digital broadcasting scenarios, but for assessing wild video impairments, such as capture artifacts, only pixel-based NR-VQA models are applicable. This is why we focus on pixel-based NR-VQA models in the rest of this paper.

Pixel-based NR-VQA models can be further classified into analytical and learning-based models. Analytical models are typically based on a limited set of distortion specific features, and they are ready to use without training [3-4,10]. Learning-based models use more generic set of features that is combined into a predicted quality score by using some machine learning algorithm for regression [5,7-9,12]. Learning-based NR-VQA models are more versatile than analytical models, but they need to be trained using video sequences with relevant quality impairments and the related MOS data as ground truth. The recently published NR-VQA models are almost exclusively learning-based NR-VQA models. Their main differences lay in the used features and regression models.

No-Reference Image Quality Assessment (NR-IQA) has been a subject for intensive research during the past few years, and many NR-VQA models are based on features used originally for NR-IQA. The best known learning-based general purpose NR-VQA method, Video-BLIINDS (V-BLIINDS) [5], uses a combination of video specific temporal features and spatial features originally designed for NR-IQA metric NIQE [31]. The temporal features are extracted using block-based motion estimation and Discrete Cosine Transform (DCT) coefficients computed from frame differences. For regression, V-BLIINDS uses Support Vector Machine (SVM). In a similar fashion, NR-VQM method in [9] uses spatial features used for

NR-IQA scheme BRISQUE [32], together with spatial and temporal activity indices and codec specific features to detect MPEG-2 and H.264/AVC artifacts. Since the method is designed for assessing compressed video sequences, it is not suitable for modeling wild video artifacts.

In [8], another method using six different frame based features was introduced. The frame level features are based on DCT coefficients, and they represent different artifacts, such as sharpness and blockiness. The frame level features are combined into video level features by using simple Minkowski pooling for each feature. Then, neural network is trained for nonlinear regression of the video level features to predict the video quality score. The model shows reasonably good performance on the traditional datasets, but the authors make a note that their model is distortion specific and limited to compression artifacts, and therefore it is not highly relevant for wild video impairments either.

For SACONVA [7], 3D shearlet transform is used for feature extraction. The use of three dimensional transform allows capturing the spatial and temporal features with one transform. Convolutional Neural Network (CNN) is then used to evolve the features, and at the final stage, logistic regression is applied to obtain the quality scores. SACONVA has been reported to outperform V-BLIINDS and several FR video quality metrics [7]; however, it has been only tested with traditional video quality databases with compression and packet loss artifacts, so its performance on wild video impairments is not known.

As we have discussed above, several NR-VQA models have been proposed in the prior art, but they are mostly designed for assessing compression and packet loss artifacts, and they have not been validated with any of the recently published consumer video quality datasets. In addition, implementations for most of those methods are not publicly available, which limits their use for benchmarking. A notable exception is V-BLIINDS, as it is publicly available and it has been benchmarked with both CVD2014, KoNViD-1k and LIVE-Qualcomm datasets [13,15,33]. Interestingly, the results from those studies show that NR-IQA models CORNIA [34,35] and FRIQUEE [36] perform better than V-BLIINDS on consumer videos. Both NR-IQA models have been applied to the video sequences frame by frame, and then the resulting features were computed by simple averaging for temporal pooling. Therefore, it seems that the state-of-the-art NR-IQA models are capable to outperform all the publicly available NR-VQA models also in video quality assessment.

Unfortunately, the most advanced NR-IQA models are computationally too complex for many practical applications of NR-VQA. Apparently, there is a need for a low complexity NR-VQA method that performs on par with the NR-IQA methods in terms of accuracy. This was our motivation for developing a low complexity hierarchical NR-VQA model proposed in this paper.

III. Proposed Method

In order to reduce the complexity of feature extraction, we have chosen a hierarchical approach for feature computation. First, low complexity features are computed using the complete sequence of video frames. Then, we use the low complexity features to determine a set of representative subset of frames for computing high complexity features. Finally, low and high complexity features are pooled and merged as a single feature vector representing the whole video sequence. In this Section, we will describe the general methodology and the features at the level of details that is appropriate for a research paper. Readers interested in full details may consult the public implementation of the method, available in jarikorhonen.github.io.

A. General Description

First, we divide the tested video sequence into segments. Processing of one segment is illustrated in Fig. 1. We use segment length that corresponds to one second of video; if the frame rate is 30 frames per second, one segment contains 30 frames (in Fig. 1, there are ten frames for a segment). Then, we compute the Low Complexity Features (LCF) using triplets of

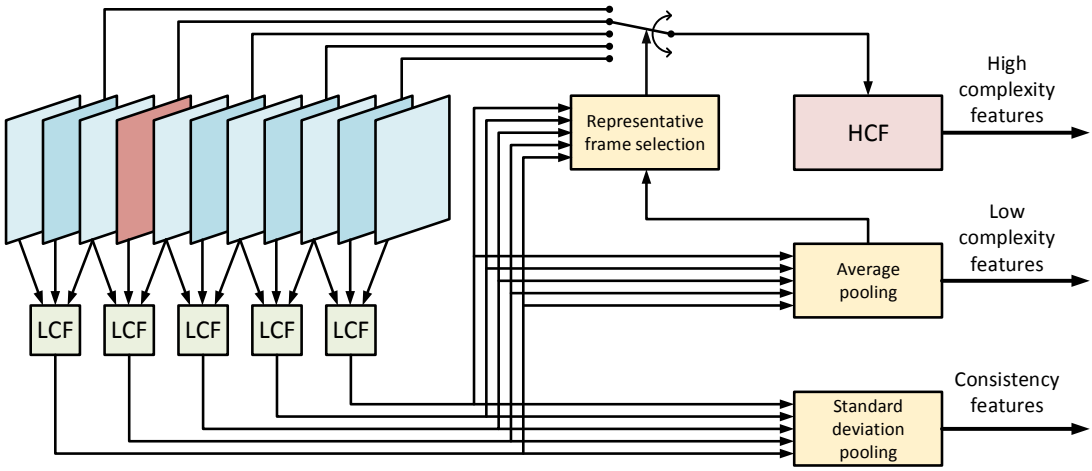


Fig. 1. Processing of one segment of video. LCF denotes computation of low complexity features and HCF high complexity features. The representative frame used for HCF is shown in red color.

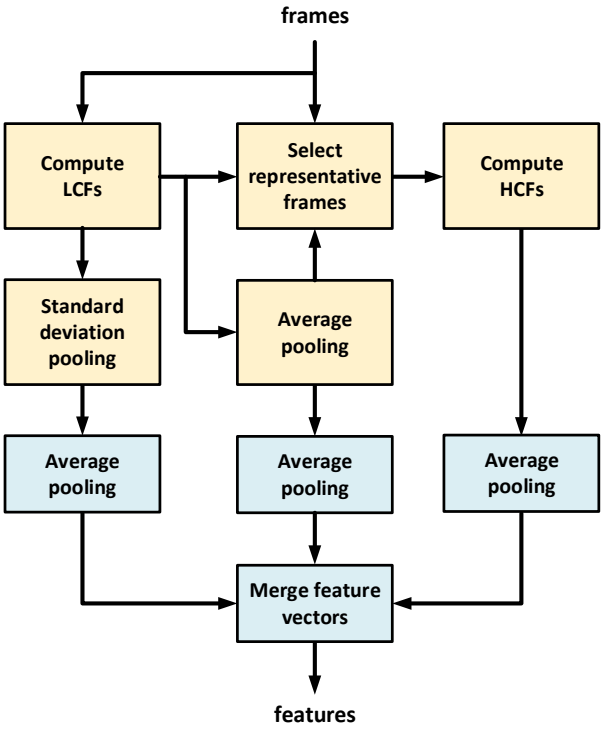


Fig. 2. Block diagram of the proposed method for sequence level feature vector computation. Yellow blocks show the segment level processing (see also Fig. 1), and blue blocks show the sequence level processing.

frames so that one frame is the reference frame (the dark colored frames in Fig. 1) and the neighboring frames are target frames for motion estimation (the light colored frames in Fig. 1). The LCFs are then pooled by using average pooling. We also compute a set of consistency features by computing standard deviations for the low complexity features. The consistency features will characterize the temporal consistency of motion within the segment. For each segment, we also select one representative frame by comparing the average of the LCFs against each frame triplets' LCF. The selected representative frame is the closest to the average of all the reference frames in the sequence, and it is used to compute the spatial High Complexity Features (HCFs). In real-life video sequences, the spatial features of frames tend to be highly correlated, and therefore we can assume that the HCFs computed for the representative frame give a realistic approximation of the overall spatial quality features of the whole segment.

On the sequence level, average temporal pooling is applied to the segment level LCFs, HCFs and consistency features to obtain the feature vector for the whole video sequence. Block diagram of the whole process is shown in Fig. 2. It should be noted that for long sequences, temporally weighted pooling should be used instead of average pooling. In addition, scene change detection should be employed and quality scores should be estimated separately for different scenes. However, in the sake of simplicity, we have restricted our scope in this paper for short single scene sequences, and in this case, average temporal pooling works nearly as well as any more complex pooling strategies.

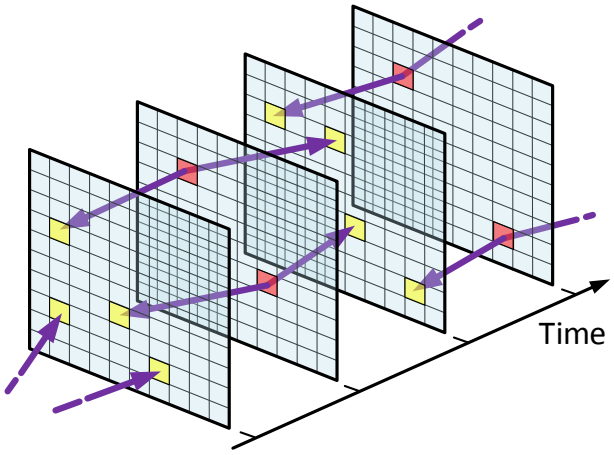


Fig. 3. Motion estimation from key pixels (marked as red) backwards and forwards (the best matching pixels are marked as yellow).

B. Low Complexity Features

The low complexity features serve for two purposes: first, they will be used in the overall feature vector to cover the temporal characteristics and dynamics of the video sequence, and second, they will be used for selecting the representative frames for computing the high complexity features. Since the main focus of the low complexity features is in temporal characterization, the main challenge is motion estimation. Block-based motion estimation is conventionally used in many video applications, in particular video compression. However, for NR-VQA, we are interested in statistical characteristics of the motion rather than accurate estimation of the motion at the pixel level. This is why we have chosen a simpler motion estimation method.

In the proposed motion estimation scheme, we first determine a set of key pixels in every second frame. They represent pixels of special interest, such as corners or tips of lines. Then, motion estimation is performed both backwards and forwards to find the respective pixels in the neighboring frames. The resulting motion vectors are depicted in Fig. 3, where key pixels in the reference frames are marked as red, the respective matching pixels in the target frames are marked as yellow, and the motion vectors are marked as purple. Even though this method may occasionally fail to detect the motion accurately for some pixels, we the method is accurate enough to collect statistically meaningful information about the motion flow, covering characteristics such as the general intensity and consistency of the motion. We assume that it is even possible to roughly classify motion in different regions as egomotion (panning, zooming, etc.) and motion related to moving objects. Before motion estimation, we resize the frames so that the width will be 480 pixels. Resizing removes high frequency noise that can disturb motion estimation, and it also helps to ignore motion that is so small that it is perceptually irrelevant.

First challenge is to find the key pixels. Several corner pixel detection methods are known in the prior art (e.g. Scale-Invariant Feature Transform, SIFT), but they tend to be unnecessarily complex to our purpose. Therefore, we have used a set of appropriately selected convolution filters and applied

them to the grayscale version of the input frame to find the candidates as key pixels. Then, we choose among the filtered pixels the pixel with the maximum value as the first key pixel and exclude the pixels in the surrounding window in order to avoid spatial clusters of key pixels. Then, the procedure is repeated with the pixel with the highest value in the remaining area of the screen, and the process is repeated until there are no remaining candidate pixels for key pixels.

When the candidate pixels are found, motion estimation is done using 3x3 pixel window around each key pixel to find the best match in the two neighboring frames. Block matching is performed using frames obtained by vertical and horizontal Sobel filter, instead of using the original pixels. Since the key pixels are supposed to be located at special points, such as corners or line tips, motion estimation can be done more accurately in the Sobel domain than in the original pixel domain, and since we only use 3x3 pixel blocks, even exhaustive search is possible without extensive computational load. To demonstrate that the proposed technique is indeed capable of capturing statistically meaningful information about the motion field, an example of motion vectors obtained with the method is shown in Fig. 4. In this frame, the camera follows the person in the center walking to the right. The motion vectors are formed by combining the motion vectors backwards and forwards, and even though some obviously erroneous motion vectors can be noticed, there is generally a good agreement between the motion vectors and our intuitive understanding of the motion present in the frame.

After the motion vectors have been computed, it is possible to compute statistical information about motion characteristics for each triplet of frames, such as density and intensity of the motion, average prediction accuracy of motion estimation, consistency of motion across the field, etc. This information is then used as temporal features for the segment. In addition to purely motion related features, we also compute some spatial features, such as indices for spatial activity, blurriness and blockiness. To comply with the computational constraints, simple techniques based on e.g. predefined convolution filters are used.

For each segment, average pooling is applied to combine the frame level features into segment level features. In addition, we compute the consistency features by computing the standard deviation in the temporal dimension for each frame level feature. It is well known that human visual system is sensitive to temporal artifacts, such as flickering and jerkiness of motion, and consistency features can reveal such impairments in a simple and straightforward manner. We have also included correlation coefficient of motion intensity index and blurriness index as a consistency feature. In total, there are 22 temporal features and 23 consistency features generated for each segment.

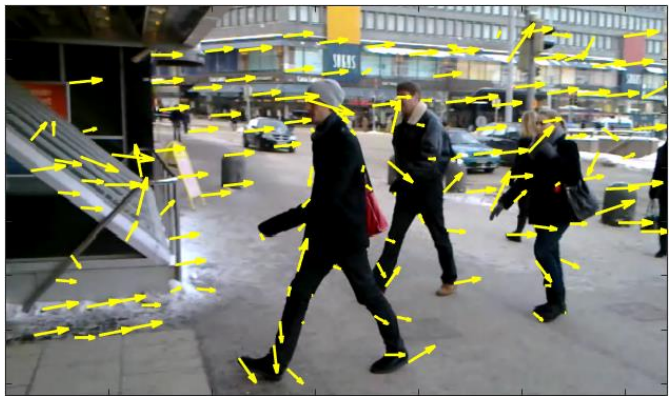


Fig. 4. Example of motion vectors obtained by the proposed technique, using a frame from CVD2014 dataset.

C. *High Complexity Features*

High complexity features are computed only for one frame per second, and therefore the constraints for computational complexity are more relaxed than for low complexity features. However, we still aim at reasonable complexity, in order to comply with timing requirements of real-life applications. Traditionally, spatial features aim at capturing the typical visual impairments, such as noise, blurriness, blockiness, or color distortion. Some of the used features are based on well-known techniques in image and video analysis, but we have also tailored some features for the specific purpose of video quality assessment. The space of a research paper does not allow detailed description of all the features; nevertheless, the used features are discussed on a general level below. The proposed spatial features can be divided in seven different categories as follows:

1) *Spatial activity features:* We use the Sobel filtered pixel intensities to compute the strength and distribution of spatial activity.

2) *Exposure features:* We use a simple segmentation algorithm to find areas where the pixel intensity is either close to the maximum (overexposed, or "burned", areas) or the minimum (underexposed areas), and variance of the pixel intensities is very small. The total sizes and the average sizes for both over- and underexposed areas are used as features.

3) *Blockiness features:* To detect blockiness, we use Sobel filter to detect horizontal and vertical edges, then we compute the mean edge value for each vertical and horizontal line of pixels, and finally we compute the autocorrelation of the horizontal and vertical mean edge values. Peaks in the output of the autocorrelation function indicate blockiness.

4) *Contrast and colorfulness features:* To characterize global contrast, we compute the cumulative difference between the evenly spread pixel intensities and the actual histogram for the pixel intensities. We also characterize the colorfulness and chrominance diversity by using the chroma components of the frame representation in the CIELAB color space.

5) *Noise features:* The algorithm searches for the pixels that represent a local minimum or local maximum, and considers them as noise pixels. Density of the noise pixels, as well as their intensity (difference between noise pixels and the surrounding pixels) and standard deviation of intensity, are used as noise features.

6) *Sharpness features:* Our basic approach is to use Sobel filter for vertical and horizontal direction, and then compute autocorrelations for 16x16 pixel blocks of the filtered frames by shifting the blocks in vertical, horizontal and diagonal

directions. For sharp images, the shape of the autocorrelation is much steeper than for blurry images. From the results, we can compute nine features that are related to the perceived blurriness or overshoot at the edges in the frame.

7) *Discrete Cosine Transform (DCT) features:* First, we apply DCT to the input frame. Then, we compute the correlations between the upper left quadrant and the mirrored upper right and lower left quadrants to detect high frequency anomalies, such as interlacing. We also use the coefficients in the lower right quadrant to determine the level of high frequency spatial activity, and the relative weights of the upper right and lower left quadrants to determine the balance between spatial details in the vertical and horizontal directions.

In total, we use 30 high complexity spatial features, with a property of capturing a wide variety of different characteristics directly related to the human perception of video quality.

D. Regression

Different learning-based regression models can be employed to predict the actual quality scores from the features. In the field of NR-IQA and NR-VQA, Support Vector Regression (SVR) is the most commonly used regression model [5,9,36-37], but also neural networks and Random Forest Regression (RFR) have been employed successfully [7-8,38]. We have tried different regression models during the preliminary evaluation of the proposed set of features and the benchmark methods, and in general, SVR and RFR seem to perform the best. Therefore, we have focused on those two regression models in the experimental part of our study.

The main weakness of SVR is that it is sensitive to the model hyperparameters, in particular γ, C and ε, and optimization of hyperparameters is considered as a research problem of its own [39,40]. It can be argued that those hyperparameters are indeed part of the model, and therefore optimizing them separately for different datasets would lead to overfitting and resulting models would not generalize very well to other datasets. In this respect, RFR is more robust method that allows more fair comparison of the prediction power of different features. On the other hand, the results with SVR may give a better idea of the theoretical limits of the compared methods and scenarios. Assuming that the hyperparameter optimization is done in a similar fashion for all the benchmark methods, the comparison of the methods can still be considered fair. For this reason, we have included results using both SVR and RFR in the experimental part of our study.

IV. Experimental Validation

For the validation study, we have used all the three relevant public datasets (CVD2014, KoNViD-1k and LIVE-Qualcomm) with consumer video content that we are aware of, available for the public [41-43]. We have implemented the feature extraction for the proposed method in Matlab, and for benchmarking, we have chosen the most relevant NR-VQM and NR-IQM methods with publicly available implementations (all in Matlab). For fair comparison, we have only considered learning based models that have been successfully employed for NR-VQM in the related studies with a decent performance. Based on these criteria, we have chosen V-BLIINDS, FRIQUEE and Video CORNIA (V-CORNIA) as benchmark methods. According to the related studies using the relevant datasets, FRIQUEE and V-CORNIA have shown the best performance in terms of MOS prediction accuracy [15-16]. To ensure fair comparison between the methods, we have used Python with scikit-learn toolbox [44] for regression in a similar fashion for the proposed method as well as all the benchmark methods.

A. Sampling Rate for NR-IQMs

The most accurate NR-IQMs tend to be computationally complex, and it is therefore not convenient to extract the NR-IQM features for every frame of a sequence. We assume that it is not even necessary, since we use average temporal pooling to compute the sequence level features, and the spatial features of neighboring frames in practical video sequences tend to be highly correlated. However, the question of the optimal sampling frequency of frames for NR-IQMs remains.

To get an idea of the appropriate sampling rates, we conducted a small study by using FRIQUEE as NR-IQM and CVD2014 low resolution subset (48 video sequences in total) to find out at which sampling rate the MOS prediction accuracy converges. We used six different sampling rates: one frame per four seconds, one frame per two seconds, one frame per second, two frames per second and four frames per second. In each test case, we used 100 different standard random splits (80% of sequences for training and 20% for testing) and computed the average prediction accuracy using SVR with the same hyperparameters as a regression model.

We noticed that fixed sampling intervals gave inconsistent results. This is probably because video compression causes some periodic quality differences in individual frames (I-frames tend to have higher quality than predicted P- and B-frames). This is why we have selected the exact positions of the frames with small random variation within the applicable sampling intervals.

The results in terms of average Pearson linear Correlation Coefficient (PCC), Spearman Rank order Correlation Coefficient (SRCC) and Root Mean Squared Error (RMSE), as well as their standard deviations, are listed in Table II. The results show that the results improve when the sampling rate is increased from one frame per four seconds to one frame per second, but there is no systematic improvement at the higher sampling rates than that. Therefore, we can conclude that the sampling rate of one frame per second of typical consumer video content is sufficient for FRIQUEE to achieve optimal or nearly optimal MOS prediction accuracy, assuming that there is no drastic variation in the content, such as scene changes.

TABLE II
Performance of FRIQUEE with Different Frame Sampling Rates

Sampling Rate	PCC (±STD)	SRCC (±STD)	RMSE (±STD)
1 frame/4 sec	0.74 (±0.13)	0.68 (±0.15)	12.4 (±2.7)
1 frame/2 sec	0.78 (±0.10)	0.72 (±0.13)	11.7 (±2.5)
1 frame/sec	0.79 (±0.11)	0.74 (±0.13)	11.3 (±2.8)
2 frames/sec	0.79 (±0.11)	0.74 (±0.12)	11.4 (±2.8)
4 frames/sec	0.79 (±0.10)	0.74 (±0.11)	11.3 (±2.7)

We ran a similar experiment using also V-CORNIA as NR-VQM. In contrast to FRIQUEE, V-CORNIA uses a very large number of features (10,000 in total), and therefore the method requires larger training set to give stable results. This is why we used the full CVD2014 dataset, including both low and high resolution sequences (234 sequences). Fortunately, feature extraction for individual frames works much faster for CORNIA than FRIQUEE. The results are summarized in Table III. The results show that there is no systematic improvement in MOS prediction accuracy when the sampling rate gets higher than one frames per second. Therefore, we can conclude that sampling one frame per second is sufficient also for V-CORNIA to achieve optimal or nearly optimal results.

TABLE III
PERFORMANCE OF V-CORNIA WITH DIFFERENT FRAME SAMPLING RATES

SAMPLING RATE	PCC (±STD)	SRCC (±STD)	RMSE (±STD)
1 frame/4 sec	0.67 (±0.09)	0.63 (±0.10)	16.0 (±1.7)
1 frame/2 sec	0.69 (±0.09)	0.67 (±0.10)	15.5 (±1.8)
1 frame/sec	0.71 (±0.08)	0.68 (±0.09)	15.2 (±1.6)
2 frames/sec	0.70 (±0.08)	0.67 (±0.10)	15.4 (±1.7)
4 frames/sec	0.71 (±0.08)	0.68 (±0.09)	15.3 (±1.7)

B. Performance on Different Datasets

We conducted a performance validation study using the benchmark methods (V-BLIINDS, V-CORNIA and FRIQUEE) and the proposed method, named as Hierarchical Video Quality Model (HiViQuM), on the three available datasets (CVD2014, KoNViD-1k and LIVE-Qualcomm). In the sake of fair comparison, the random splits were initialized using the same set of seed numbers to ensure identical splits for different methods, and we also included indices for the frame rate for and resolution of CVD2014 sequences and frame rate of KoNViD-1k sequences in the feature vectors of all the methods. Based on the results in previous Subsection, we used average temporal pooling and frame sampling rates of one frame per second for V-CORNIA and FRIQUEE. We used 100 different random splits with 80% of data for training and 20% of sequences for validation in each case, and reported the average results and their standard deviations in Tables IV-VI, respectively. For SVR, we optimized the model parameters (γ, C and ε) separately for each dataset. The resulting values are listed in Table VII. For RFR, we used fixed model parameters for every test case.

TABLE VII
OPTIMAL SVR PARAMETERS FOR DIFFERENT DATASETS

	CVD2014			KONVID-1K			LIVE-QUALCOMM		
	γ	C	ε	γ	C	ε	γ	C	ε
V-CORNIA	$5\cdot10^{-4}$	2^{10}	0.01	$5\cdot10^{-4}$	2^{6}	0.1	$5\cdot10^{-4}$	2^{13}	1
V-BLIINDS	0.3	2^{7}	0.1	0.4	2^{6}	0.2	0.5	2^{6}	3
FRIQUEE	$5\cdot10^{-3}$	2^{7}	0.1	0.05	2	0.3	$5\cdot10^{-3}$	2^{9}	1
HiViQuM	0.2	2^{7}	0.3	0.2	5	0.2	0.1	2^{6}	0.3

As the results show, the proposed HiViQuM model outperforms all the benchmark methods on all the three datasets. Among the regression schemes, SVR performs better than RFR for all the methods, except with V-BLIINDS for CVD2014 and KoNViD-1k datasets, but the relative results for different NR-VQMs are similar with both regression models. The results with different datasets are in line with each other, suggesting that the conclusions concerning the relative performance of the methods are valid for a wide range of content and distortion types. Representative examples of scatter

TABLE IV
RESULTS ON CVD2014 DATASET USING DIFFERENT NR-VQMS

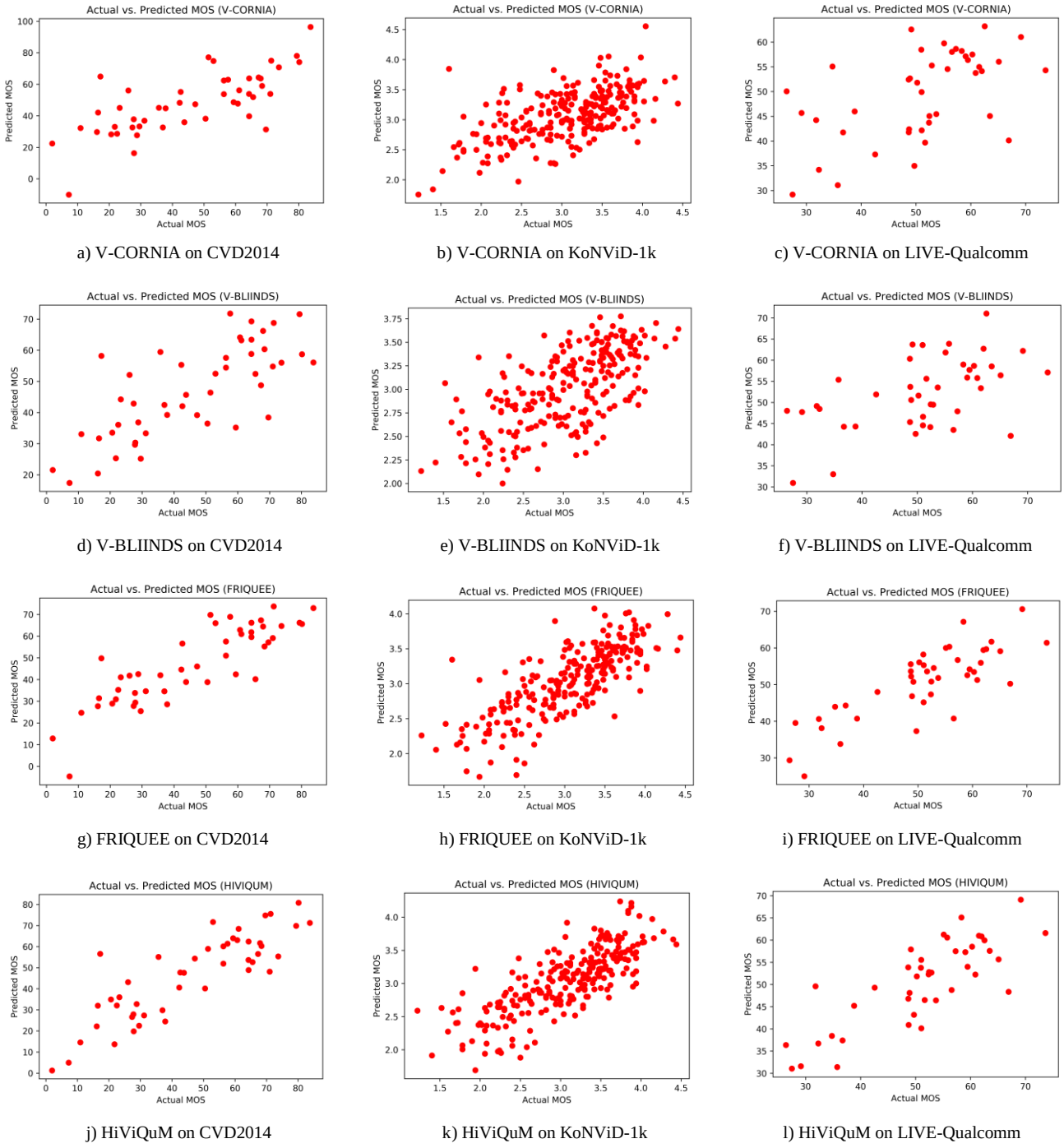
	SUPPORT VECTOR REGRESSION			RANDOM FOREST REGRESSION		
METHOD	PCC (±STD)	SRCC (±STD)	RMSE (±STD)	PCC (±STD)	SRCC (±STD)	RMSE (±STD)
V-CORNIA (1 fr/sec)	0.71 (±0.08)	0.68 (±0.09)	15.2 (±1.6)	0.63 (±0.10)	0.61 (±0.10)	16.9 (±1.5)
V-BLIINDS	0.71 (±0.09)	0.70 (±0.09)	15.2 (±2.0)	0.74 (±0.07)	0.73 (±0.08)	14.6 (±1.6)
FRIQUEE (1 fr./sec)	0.83 (±0.04)	0.81 (±0.05)	12.0 (±1.2)	0.77 (±0.07)	0.74 (±0.07)	13.9 (±1.6)
HiViQuM	**0.86 (±0.04)**	**0.84 (±0.04)**	**11.0 (±1.3)**	0.82 (±0.05)	0.79 (±0.05)	12.7 (±1.5)

TABLE V
RESULTS ON KONVID-1K DATASET USING DIFFERENT NR-VQMS

	SUPPORT VECTOR REGRESSION			RANDOM FOREST REGRESSION		
METHOD	PCC (±STD)	SRCC (±STD)	RMSE (±STD)	PCC (±STD)	SRCC (±STD)	RMSE (±STD)
V-CORNIA (1 fr./sec)	0.55 (±0.07)	0.57 (±0.04)	0.542 (±0.042)	0.46 (±0.09)	0.46 (±0.09)	0.546 (±0.038)
V-BLIINDS	0.60 (±0.04)	0.63 (±0.04)	0.513 (±0.027)	0.64 (±0.04)	0.65 (±0.04)	0.490 (±0.022)
FRIQUEE (1 fr./sec)	0.75 (±0.03)	0.76 (±0.03)	0.423 (±0.022)	0.73 (±0.03)	0.73 (±0.03)	0.441 (±0.021)
HiViQuM	**0.78 (±0.02)**	**0.78 (±0.02)**	**0.402 (±0.018)**	0.74 (±0.03)	0.74 (±0.03)	0.434 (±0.021)

TABLE VI
RESULTS ON LIVE-QUALCOMM DATASET USING DIFFERENT NR-VQMS

	SUPPORT VECTOR REGRESSION			RANDOM FOREST REGRESSION		
METHOD	PCC (±STD)	SRCC (±STD)	RMSE (±STD)	PCC (±STD)	SRCC (±STD)	RMSE (±STD)
V-CORNIA (1 fr./sec)	0.60 (±0.09)	0.55 (±0.09)	9.8 (±0.9)	0.43 (±0.13)	0.40 (±0.13)	10.6 (±1.1)
V-BLIINDS	0.67 (±0.09)	0.60 (±0.10)	8.9 (±0.9)	0.63 (±0.10)	0.59 (±0.10)	9.4 (±0.9)
FRIQUEE (1 fr./sec)	0.77 (±0.05)	0.74 (±0.07)	7.7 (±0.8)	0.64 (±0.09)	0.62 (±0.10)	9.3 (±1.0)
HiViQuM	**0.81 (±0.06)**	**0.78 (±0.06)**	**7.0 (±1.0)**	0.71 (±0.10)	0.68 (±0.09)	8.8 (±1.1)



a) V-CORNIA on CVD2014 b) V-CORNIA on KoNViD-1k c) V-CORNIA on LIVE-Qualcomm

d) V-BLIINDS on CVD2014 e) V-BLIINDS on KoNViD-1k f) V-BLIINDS on LIVE-Qualcomm

g) FRIQUEE on CVD2014 h) FRIQUEE on KoNViD-1k i) FRIQUEE on LIVE-Qualcomm

j) HiViQuM on CVD2014 k) HiViQuM on KoNViD-1k l) HiViQuM on LIVE-Qualcomm

Fig. 5. Representative examples of scatter plots for different NR-VQMs on CVD2014, KoNViD-1k and LIVE-Qualcomm datasets.

plots for each NR-VQM on each dataset are shown in Fig. 5, using the best performing regression method for each combination of features and dataset.

In the related work, other authors have compared some of the same methods using the same datasets. In [13], V-CORNIA and V-BLIINDS were included as benchmark methods on CVD2014 dataset. The reported results are much worse than our results, but this is not surprising, since both methods were tested using default models trained with a database impaired by compression and transmission errors only. Therefore, it is expected that the performance for assessing capture artifacts is not optimal. In [15], V-BLIINDS and FRIQUEE were tested on LIVE-Qualcomm dataset. Apart from small differences that can be explained by minor differences in the training and validation procedures, our results are in line with the results in [15].

In [33], V-CORNIA and V-BLIINDS were tested on

KoNViD-1k dataset. Their results for V-BLIINDS were worse than our results, whereas the results for V-CORNIA were better than our results. There are several possible explanations for the differences, and without detailed information of the model parameter selection and the validation procedure, definite conclusions cannot be made. For V-BLIINDS, the difference between [33] and our study is not as large, and it can be explained by differences in hyperparameter optimization and validation procedure. For V-CORNIA, the authors in [33] have used all the 20,000 features originally defined for image based CORNIA [34], whereas we have used different spatial pooling to reduce the number of features to 10,000, as suggested in [35]. It could be speculated that the difference comes from the subsampling of frames, since we use only one frame per second, whereas in [33] the features are computed for every frame. To exclude this possibility, we computed the results using the full sequence for a subset of KoNViD dataset (300 video sequences) and compared the results with frame subsampling on the same subset, but we did not find essential difference in the results.

C. Performance across Datasets

Even though all the studied NR-VQMs show a reasonably good performance on different datasets separately, cross-database scenarios tend to be more challenging, due to the presence of different types of contents and quality impairments in different datasets. In order to study the robustness of NR-VQMs, we also measured the MOS prediction performance by using different datasets for training and validation. For the sake of fairness, we have used the SVR hyperparameters optimized with the training dataset also for testing. Since only CVD2014 dataset has multiple resolutions and LIVE-Qualcomm dataset has fixed frame rate for all the sequences, we did not include resolution and frame rate and indicators in the feature vectors. To make MOS values comparable, we applied simple linear mapping to convert all the results to the same scale (from one to five).

Unfortunately, the results were not satisfactory with any of the methods. The best performing methods in terms of PCC/SRCC in different scenarios are shown in Table VIII. As seen from the Table, none of the methods shows consistently better performance than the others. The results with SVR and RFR were also inconsistent, and the numerical results were rather disappointing, PCC and SRCC ranging from 0.1 to 0.6, depending on the test case. It seems also that because the three datasets have a different range of impairments in terms of severity, the rating scales of the datasets are not directly comparable. In particular, KoNViD-1k has a substantially wider range of impairments than the two other datasets, but there are also qualitative differences between the dominating artifacts in CVD2014 and LIVE-Qualcomm datasets. This is one possible reason for the poor performance of all the models in this test scenario.

TABLE VIII
BEST PERFORMING METHODS IN CROSS-DATABASE EXPERIMENT

	Training Dataset		
Testing Dataset	CVD2014	KoNViD-1k	LIVE-Qualc.
CVD2014	-	FRIQUEE	V-CORNIA
KoNViD-1k	V-BLIINDS	-	HiViQuM
LIVE-Qualcomm	V-CORNIA	HiViQuM	-

Due to the qualitative differences between the datasets, we may argue that fair comparison between the methods in a cross-database scenario is not feasible: it is not realistic to expect that a learning based quality model could predict quality of the validation sequences accurately, if the respective combinations of frame rates, resolutions and impairment types in the validation dataset do not appear in the training dataset. Therefore, we can assume that the poor performance of the quality models in a cross-database experiment indicates different characteristics of the datasets, rather than poor generalization capability of the quality models.

To test the generalization ability of different models in a more realistic manner, we combined all the three datasets and run a similar test as we did for those datasets separately. MOS values of KoNViD-1k dataset were linearly mapped into rating scale 0-100 to obtain scales for all the datasets. The results are shown in Table IX. In general, the results are in line with the results obtained using each dataset separately. HiViQuM still outperforms FRIQUEE and is substantially better than V-CORNIA and V-BLIINDS.

D. Complexity

It is difficult to assess the computational complexity of NR-VQA algorithms analytically, because there is typically a large number of features and the actual execution time varies on different contents. However, our implementation of HiViQuM, as well as the implementations available for V-BLIINDS, V-CORNIA and FRIQUEE, are all running on Matlab, and therefore we can get a rough idea of the differences between their complexities by running the algorithms on the same content and the same computer, and then comparing the execution times.

We have selected ten representative video sequences from CVD2014 database, five with low resolution (640x480) and five with high resolution (1280x720). The lengths of the

TABLE IX
RESULTS ON LIVE-QUALCOMM DATASET USING DIFFERENT NR-VQMS

	Support Vector Regression			Random Forest Regression		
Method	PCC (±std)	SRCC (±std)	RMSE (±std)	PCC (±std)	SRCC (±std)	RMSE (±std)
V-CORNIA (1 fr./sec)	0.56 (±0.05)	0.58 (±0.04)	13.8 (±0.79)	0.54 (±0.04)	0.52 (±0.04)	14.0 (±0.50)
V-BLIINDS	0.65 (±0.03)	0.65 (±0.03)	12.6 (±0.49)	0.66 (±0.03)	0.65 (±0.03)	12.5 (±0.53)
FRIQUEE (1 fr./sec)	0.75 (±0.02)	0.75 (±0.02)	10.9 (±0.48)	0.71 (±0.02)	0.71 (±0.03)	11.7 (±0.42)
HiViQuM	**0.77 (±0.02)**	**0.77 (±0.02)**	**10.6 (±0.42)**	0.73 (±0.02)	0.72 (±0.02)	11.5 (±0.43)

sequences varies approximately from ten to twenty seconds. As the results in Table XI show, the proposed method is substantially faster than V-BLIINDS that is also using a combination of temporal and spatial features. The proposed method is also substantially faster than FRIQUEE, when one frame per second is sampled to compute the features. On the other hand, at the sampling rate of one frame per second, V-CORNIA is faster than HiViQuM and seems computationally less complex. V-CORNIA features are computed by using a predetermined codebook and standard built-in elementwise matrix operations in Matlab, allowing fast implementation.

We expect that both V-BLIINDS and the proposed method would run essentially faster (in the order of magnitude) if implemented in a low level compiled programming language, such as C++, but we also assume that the relative performance differences would remain roughly similar. We also expect that the performance of FRIQUEE and V-CORNIA cannot be improved in the same extent, since the core operations of FRIQUEE rely largely on compiled components in matlabPyrTools, and V-CORNIA uses optimized built-in Matlab matrix operations. The complexity of the proposed method can be adjusted easily by changing the maximum number of motion points and the size of the search window, but this will of course also influence the accuracy of the method.

TABLE XI
CROSS-DATASET PERFORMANCE USING KONVID-1K AND LIVE

METHOD	LOW RESOLUTION (640X480)	HIGH RESOLUTION (1280X720)
FRIQUEE (1 frame/sec)	466.7 sec	1355.9 sec
V-BLIINDS	455.6 sec	1050.2 sec
HiViQuM	69.4 sec	222.2 sec
V-CORNIA (1 frame/sec)	15.3 sec	24.9 sec

The computation time for regression with a pre-trained model is negligible in comparison to feature extraction. On the other hand, the time for training a model is proportional to the number of features. V-BLIINDS uses 47 features, the proposed method uses 75 features, FRIQUEE uses 561 features and V-CORNIA uses 10,000 features per video sequence. Therefore, training the regression models for V-BLIINDS and the proposed method is faster than training the regression model for FRIQUEE, and significantly faster than for V-CORNIA. Another disadvantage of V-CORNIA is that due to its large number of features, it is prone to the "curses of dimensionality," [45] such as need for large training datasets, redundant features and overfitting. As the performance comparison in Subsections IV.B and IV.C shows, V-CORNIA tends to predict subjective video quality less accurately than the three other methods.

V. CONCLUSIONS

In this paper, we have proposed an objective NR-VQA method, aimed specifically at consumer-generated content, prone to capture artifacts such as camera shakiness, over- and underexposure, and sensor noise. Our method is based on the conventional approach, using a manually selected set of features as an input to a learning-based regression model. The main novelty of our proposal is in the mechanism for selecting the frames for which the temporal and spatial features are computed.

The traditional learning-based NR-VQA methods first compute frame level features, and then combine them in sequence level feature vectors. We have split the feature computation in four parts: first, we compute simple temporal features for every second frame, to capture the temporal artifacts. Second, we use the temporal features to select the most representative frames for computation of spatial features, to capture the spatial artifacts. Third, we use the temporal features to compute motion consistency features. Finally, we pool and merge all the features together. With this approach, we can reduce the computational complexity of feature extraction substantially. We have trained and validated the proposed method using three relevant annotated video quality databases with a wide variety of video contents and distortions, and the results show that the proposed method can outperform benchmark methods representing the state-of-the-art. In order to support reproducible research, we have published the source code for our method in [46].

REFERENCES

[1] S. Argyropoulos, A. Raake, M. N. Garcia, and P. List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Proc. QoMEX*, Mechelen, Belgium, 2011, DOI: 10.1109/QoMEX.2011.6065708, [Online].

[2] C. Keimel, J. Habigt, M. Klimpke, K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Proc. QoMEX*, Mechelen, Belgium, 2011, DOI: 10.1109/QoMEX.2011.6065711, [Online].

[3] Z. Chen and D. Wu, "Prediction of transmission distortion for wireless video communication: analysis," in *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1123-1137, Mar. 2012, DOI: 10.1109/TIP.2011.2168411, [Online].

[4] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," in *IEEE Trans. Circuits and Syst. for* Video Tech., vol. 22, no. 4, pp. 605-618, Apr. 2012, DOI: 10.1109/TCSVT.2011.2171211, [Online].

[5] M. A. Saad, A. C. Bovik, "Blind prediction of natural video quality*." IEEE Trans. Image Proc*., vol. 23, no. 3, pp. 1352-65, Mar. 2014, DOI: 10.1109/TIP.2014.2299154 [Online].

[6] K. Pandremmenou, M. Shahid, L. P. Kondi, B. Lovstrom, "A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses," in. P*roc. SPIE 9394 (HVEI)*, San Francisco, CA, USA, 2015, DOI: 10.1117/12.2077709, [Online].

[7] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung, "No-reference video quality assessment with 3D Shearlet transform and convolutional neural networks," in *IEEE Trans. Circuits and Syst. for Video Tech*. 26, 6 (Jun. 2015), 1044-57. DOI: 10.1109/TCSVT.2015.2430711 [Online].

[8] K. Zhu, C. Li, V. Asari and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," in *IEEE Trans. Circuits and Syst. for Video Tech*., vol. 25, no. 4, pp. 533-546, Apr. 2015, DOI: 10.1109/TCSVT.2014.2363737, [Online].

[9] J. Søgaard, S. Forchhammer and J. Korhonen, "No-reference video quality assessment using codec analysis," in *IEEE Trans. Circ. and Syst. for Video Tech*., vol. 25, no. 10, pp. 1637-50, Oct. 2015, DOI: 10.1109/TCSVT.2015.2397207, [Online].

[10] A. Mittal, M. A. Saad, and A. C. Bovik, "A Completely blind video integrity oracle," in *IEEE Trans. Image Proc.*, vol. 25, no. 1, pp. 289-300, Jan 2016, DOI: 10.1109/TIP.2015.2502725 [Online].

[11] M. Torres Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality," in *Signal Process.: Image*

Comm., vol. 52, pp. 20-32, Mar. 2017, DOI: 10.1016/j.image. 2016.12.001 [Online].
[12] J. Korhonen, "Learning-based prediction of packet loss artifact visibility in networked video," in *Proc. QoMEX*, Sardinia, Italy, 2018.
[13] M. Nuutinen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014 – A Database for evaluating no-reference video quality assessment algorithms," in *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073-86, May 2016, DOI: 10.1109/TIP.2016.2562513 [Online].
[14] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *Proc. QoMEX*, Erfurt, Germany, 2017, DOI: 10.1109/QoMEX.2017.7965673 [Online].
[15] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," in *IEEE Trans. Circuits and Syst. For Video Tech.* (in print), DOI: 10.1109/TCSVT.2017.2707479 [Online].
[16] H. Men, H. Lin, and D. Saupe, "Spatiotemporal feature combination model for no-reference video quality assessment," in *Proc. QoMEX*, Sardinia, Italy, 2018, pp. 72-74.
[17] T. Alpert, and J.-P. Evain, "Subjective quality evaluation – The SSCQE and DSCQE methodologies," EBU Technical Review, pp. 12-20, 1997. [Online]. Available: https://tech.ebu.ch/docs/techreview/trev_271-evain.pdf
[18] Subjective Video Quality Assessment Methods for Multimedia Applications, ITU-R Recommendation P.910, 1999.
[19] F. Kozamernik, V. Steinmann, P. Sunna, and E. Wyckens, "SAMVIQ-A new EBU methodology for video quality evaluations in multimedia," in *SMPTE Motion Imag. J.*, vol. 114, no. 4, pp. 152-160, Apr. 2005, DOI: 10.5594/J11535, [Online].
[20] Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R Recommendation BT.500-13, 2012.
[21] S. Winkler, "Analysis of public image and video databases for quality assessment," in *IEEE J. on Sel. Topics in Signal Process.*, vol. 6, no. 6, pp. 616-625, Oct. 2012, DOI: 10.1109/JSTSP.2012.2215007, [Online].
[22] Report on the Validation of Video Quality Models for High Definition Video Content. Video Quality Experts Group (VQEG), 2010. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/HDTV/VQEG_HDTV_Final_ Report_ version_ 2.0.pdf.
[23] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective quality assessment of H.264/AVC video streaming with packet losses," in *EURASIP J. Image and Video Process.*, Jan. 2011, DOI: 0.1155/2011/190431, [Online].
[24] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," in *IEEE Trans. Circuits and Syst. for Video Tech.*, vol. 20, no. 4, pp. 587-599, Apr. 2010, DOI: 10.1109/TCSVT.2010.2041829, [Online].
[25] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," in *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541-558, Feb. 2014, DOI: 10.1109/TMM.2013.2291663 [Online].
[26] S. Winkler, and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," in *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 660-668, Sep. 2008, DOI: 10.1109/TBC.2008.2000733, [Online].
[27] Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference, ITU-T Recommendation J.341, 2016. [Online] Available: https://www.itu.int/rec/T-REC-J.341-201603-I/en.
[28] Objective Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Reduced Reference Signal, ITU-T Recommendation J.342, 2011. [Online] Available: https://www.itu.int/rec/T-REC-J.342/en.
[29] R. Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *Proc. BMSB*, Cagliari, Italy, 2017, DOI: 10.1109/BMSB.2017.7986143, [Online].
[30] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," arXiv 1804.04813.pdf, 2018. [Online].
[31] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a 'completely blind' image quality analyzer," in *IEEE Signal processing Letters*, vol. 22, no. 3, pp. 209-212, Mar. 2013, DOI: 10.1109/LSP.2012.2227726 [Online].
[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," in *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012, DOI: 10.1109/TIP.2012.2214050 [Online].
[33] H. Men, H. Lin and D. Saupe, "Spatiotemporal feature combination model for no-reference video quality assessment," in *Proc. QoMEX*, Sardinia, Italy, 2018.
[34] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. CVPR*, Providence, RI, USA, 2012, DOI: 10.1109/CVPR.2012.6247789 [Online].
[35] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. ICIP*, Paris, France, 2014, DOI: 10.1109/ICIP.2014.7025098, [Online].
[36] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," in *J. of Vision*, vol. 17, no. 1, pp. Jan. 2017, DOI: 10.1167/17.1.32, [Online].
[37] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," in *IEEE Trans. Image Proc.*, vol. 21, no. 12 pp. 4695-08, Dec. 2012, DOI: 10.1109/TIP.2012.2214050, [Online].
[38] S.-C. Pei, and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," in *IEEE Trans. Image Proc.*, vol. 24, no. 11, pp. 3282-92, Nov. 2015, DOI: 10.1109/TIP.2015.2440172, [Online].
[39] K. Ito, and R. Nakano, "Optimizing Support Vector regression hyperparameters based on cross-validation," in *Proc. IJCNN*, Portland, OR, USA, Jul. 2003, DOI: 10.1109/IJCNN.2003.1223728, [Online].
[40] P. Tsirikoglou, S. Abraham, F. Contino, C. Lacor, and G. Ghorbaniasl, "A Hyperparameter selection technique for support vector regression models," in *Applied Soft Computing*, vol. 61, no. 12, pp. 139-148, Dec. 2017, DOI: 10.1016/j.asoc/2017.07.017, [Online].
[41] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, Jukka Häkkinen, "The CVD video database," [Online] Available: http://www.helsinki.fi/~tiovirta/Resources/CVD2014
[42] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang, "LIVE-Qualcomm mobile in-capture video quality database," [Online] Available: http://live.ece.utexas.edu/research/ incaptureDatabase/index.html, 2017..
[43] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "KoNViD-1k video database," [Online] Available: http://database.mmsp-kn.de/konvid-1k-database.html
[44] scikit-learn toolbox. [Online]. Available: http://scikit-learn.org
[45] N. Altman, and M. Krzywinski, "The curse(s) of dimensionality," in *Nature Methods*, vol. 15, no. 6, pp. 399-400, Jun. 2018, DOI: 10.1038/s41592-018-0019-x, [Online].
[46] J. Korhonen, "Hierarchical video quality model (HiViQuM)," [Online] Available: https://github.com/jarikorhonen/nr-vqa-consumervideo

Jari Korhonen (M'05) received the M.Sc. (Eng.) degree in information engineering from University of Oulu, Finland, in 2001 and the Ph.D. degree in telecommunications from Tampere University of Technology, Finland, in 2006. Currently, he is with the Institute of Future Media Technology, Shenzhen University, P. R. China, where he is currently working as Research Assistant Professor since 2017.

From 2001 to 2006, he was Research Engineer with Nokia Research Center, Tampere, Finland. In 2007, he was with École Polytechnique Fédérale de Lausanne, Switzerland, and from 2008 to 2010 with Norwegian University of Science and Technology, Trondheim, Norway. From 2010 to 2017 he was with Technical University of Denmark. His research interests include both telecommunications and signal processing aspects in multimedia communications, including visual quality assessment.