Journal of Electronic Imaging

JElectronicImaging.org

Perceptual quality estimation of H.264/AVC videos using reduced-reference and no-reference models

Muhammad Shahid
Katerina Pandremmenou
Lisimachos P. Kondi
Andreas Rossholm
Benny Lövström

SPIE. IS&T imaging.org

Muhammad Shahid, Katerina Pandremmenou, Lisimachos P. Kondi, Andreas Rossholm, Benny Lövström, "Perceptual quality estimation of H.264/AVC videos using reduced-reference and no-reference models," *J. Electron. Imaging* **25**(5), 053012 (2016), doi: 10.1117/1.JEI.25.5.053012.

Perceptual quality estimation of H.264/AVC videos using reduced-reference and no-reference models

Muhammad Shahid,[a,b] **Katerina Pandremmenou,**[c,*] **Lisimachos P. Kondi,**[c] **Andreas Rossholm,**[a] **and Benny Lövström**[a]
[a]Blekinge Institute of Technology, Karlskrona 37179, Sweden
[b]Prince Sultan University, Rafha Street, Riyadh 12435, Saudi Arabia
[c]University of Ioannina, Department of Computer Science & Engineering, Ioannina GR-45110, Greece

Abstract. Reduced-reference (RR) and no-reference (NR) models for video quality estimation, using features that account for the impact of coding artifacts, spatio-temporal complexity, and packet losses, are proposed. The purpose of this study is to analyze a number of potentially quality-relevant features in order to select the most suitable set of features for building the desired models. The proposed sets of features have not been used in the literature and some of the features are used for the first time in this study. The features are employed by the least absolute shrinkage and selection operator (LASSO), which selects only the most influential of them toward perceptual quality. For comparison, we apply feature selection in the complete feature sets and ridge regression on the reduced sets. The models are validated using a database of H.264/AVC encoded videos that were subjectively assessed for quality in an ITU-T compliant laboratory. We infer that just two features selected by RR LASSO and two bitstream-based features selected by NR LASSO are able to estimate perceptual quality with high accuracy, higher than that of ridge, which uses more features. The comparisons with competing works and two full-reference metrics also verify the superiority of our models. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.5.053012]

Keywords: no-reference; packet loss; perceptual quality estimation; reduced-reference; video quality.

Paper 15785P received Oct. 26, 2015; accepted for publication Aug. 23, 2016; published online Sep. 20, 2016.

1 Introduction

The video portion of the global mobile data traffic has increased tremendously and it is estimated to be nearly 75% by 2019, from being 55% in 2014.[1] Therefore, with this growing usage of videos, it is believed that the end-users are becoming more aware of the perceptual quality characteristics of video services. A required amount of compression of the raw (original) videos has to be performed in order to meet the practical limits of data storage devices and transmission channels. Depending upon its intensity, the compression can introduce different visual artifacts in a video that may decrease its perceptual quality as compared to its original version.

Besides compression, video quality can also suffer from degradations due to transmission over lossy networks. Losses of video data in a network can occur for various reasons, such as network fluctuations, buffer overflows, and any operational management procedures. However, there is a growing trend of video communications through reliable transmission methods, where losses can be recovered through retransmissions, though it might be difficult to avoid all losses in the case of varying network characteristics. It is believed that retransmissions cannot avoid all losses in real-time video transmissions due to delay constraints. Moreover, real-time communications, such as video-conferencing and other low-delay demanding video services, may suffer from packet losses, as the underlying transport mechanisms generally do not apply retransmissions, e.g., user datagram protocol (UDP), real-time transport protocol (RTP), and so on. A parameter that is commonly used by service providers in order to evaluate the quality of service for an end-user is the packet loss rate (PLR), which is generally considered as a useful measure for quantifying the losses in a network. Hence, a study involving PLR or features that attempt to model this quantity so as to evaluate the performance of video communications in lossy networks can be quite useful.

In most scenarios of processing or transmission of visual information, the ultimate judges of quality are human observers. Despite the fact that many evaluation methods of objective performance have been developed, subjective assessment is the most authoritative solution, since it provides the ground truth of quality. The recommended procedures for subjective video quality assessment (VQA) involve the collection of quality scores from a viewers' panel, usually under a controlled laboratory environment[2] or relatively less controlled environments.[3] The product of such assessments is typically a mean opinion score (MOS)[4] for each test sample, which corresponds to the average value of the scores given by the panel. Crowdsourcing-based subjective VQA is an emerging technique, where test material is transferred to the viewer's premises through the Internet and the quality scores are collected through a loosely controlled environment.[5] However, subjective VQA is rather tedious, time-consuming, and impractical to be incorporated in many real-time applications.

In the last two decades, many modern models/metrics of perceptual VQA have been developed and they can be computed automatically based on quality-relevant features of a video. The goal of such objective metrics is the computation of a perceptual quality estimate that correlates well with the results of subjective assessment. A classification of

*Address all correspondence to: Katerina Pandremmenou, E-mail: apandrem@gmail.com

1017-9909/2016/$25.00 © 2016 SPIE and IS&T

the objective metrics can be made on the basis of the reference information used for quality estimation.[6] Given that "original" refers to the unprocessed pristine video and "impaired" refers to its processed version (including coding and/or transmission losses), full-reference (FR) metrics have full access to both the original and impaired videos, reduced-reference (RR) metrics have access to some suitable features transmitted from the server's side and full access to the impaired video, and no-reference (NR) metrics have access only to the impaired video.

It is generally believed that FR metrics have the capacity to provide the most accurate estimations of video quality, since they use input information from both the original and impaired videos. However, because of the dependence on the original video, FR metrics are mostly suitable for offline applications, such as encoder performance comparisons. In addition to the processed video, RR metrics can also access selected features of the original video. These features can be sent to the receiver through an ancillary channel[7] or alternatively, they can be embedded in the video content itself, by using techniques, such as watermarking.[8] For the purpose of quality estimation, NR metrics make use of either the bitstream of the impaired video or the decoded pixels of it, or a combination of both to build NR hybrid metrics.[9,10] A detailed review of NR metrics using this classification can be found in Ref. 11. Because of the limited or lack of dependence on the original video, RR and NR metrics are suitable for real-time applications and online quality monitoring of the streaming videos.[12]

In this article, we present a study on the design, implementation, and evaluation of RR and NR models, which employ two different sets of perceptually motivated features. These features are extracted from H.264/AVC encoded videos impaired with different amounts of packet losses. In order to make an accurate quality estimation, we decided to extract a large number of features from the bitstreams as well as from the pixels of a video and hence, our proposed models belong to the class of hybrid metrics. The use of least absolute shrinkage and selection operator (LASSO) regression[13–15] is proposed aiming at the dual goal of feature selection and MOS estimation. As a baseline, ridge regression[16–18] is applied on the preselected set of features, having performed sequential feature selection (SFS)[19] on the complete RR and NR set of features. Thus, ridge performs MOS estimation without performing any feature selection. An overall scheme of this work is presented in Fig. 1.

The rest of this article is organized as follows: Sec. 2 presents an overview of the related work and concludes with a summary of the key points and contributions of this paper. A discussion on the employed features that are potentially related to perceptual video quality takes place in Sec. 3 and the problem of video quality estimation based on a set of quality-relevant features and its solution using LASSO is presented in Sec. 4. In the same section, the procedure followed for the model's development is also described. The employed measures of performance, the provided experimental results and their analysis, and the performance comparison of our proposed models with that of ridge in combination with feature selection as well as with related works are given in Sec. 5. Finally, conclusive remarks on this work are given in Sec. 6.

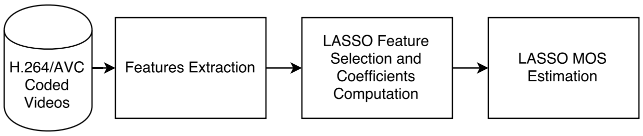


Fig. 1 Overall scheme of the proposed approach.

2 Related Work

During the last decade, a considerable part of the scientific community has focused its interest on efforts for the development of objective video quality metrics that target at reliable and accurate modeling of subjective VQA. Some RR metrics of VQA are presented in Refs. 20–22. In Ref. 20, the authors designed an RR metric that is targeted for applications related to wireless communications. It is built based on the principle that humans tend to have different impairment perceptibility based on the spatial and temporal affected regions of a video sequence. The work in Ref. 21 presented a family of RR VQA models that differ in the amount of reference information required for video quality measurement, while Ref. 22 proposed a wavelet-based video distortion metric that can operate in FR or RR mode, as required. Actually, RR metrics can be an alternative to FR metrics when the original video is not accessible. However, in some cases, the cost of maintaining an ancillary channel may be high for an RR approach, while such metrics may not meet the requirements of quality estimation in the event of a failure in RR data delivery to the receiver's end.

For these reasons, NR metrics are the most broadly applicable solution for VQA, though quality estimation with limited available input information can be challenging.[11] An NR metric tested on MPEG-4 compressed video that estimates the peak signal to noise ratio (PSNR) at the macroblock (MB) level was proposed in Ref. 23 and a similar method that estimates the structural similarity (SSIM) index was introduced in Ref. 24. The study presented in Ref. 25 described a PSNR estimator that considers only the compressed bitstream of an H.264/AVC coded video. However, the estimation of perceptual quality in terms of MOS could be an applicable improvement for the works presented in Refs. 23–25.

A set of bitstream-based features related to slice coding type, coding modes, various statistics of motion vectors, and quantization parameter (QP) values were employed in Ref. 26 with the goal of quality estimation of high definition television video, encoded by H.264/AVC. For the same purpose, statistics of boundary strength values of the deblocking filter, QP, and average bitrates were used in Ref. 27 for H.264/AVC-encoded videos. Also, a motion-based quality metric was explored in Ref. 28 for H.264/AVC-encoded videos as well. For this metric, some statistical features related to motion vectors along with the bitrate and frame rate were calculated, and the principal component analysis method was used to identify the parameters that can be the most influential in quality value. Similarly, a low complexity solution of VQA based on bitstream features was proposed in Ref. 29. An improvement of this approach was included in Ref. 30, in which the required number of features was reduced so as to promote computational efficiency. In that work, an improvement was noted in estimation accuracy by the virtue of the usage of an artificial neural network. A further improvement of Ref. 30 can be found in Ref. 31, in which a larger set of parameters was used and the estimation of subjective MOS was also considered. However, the models built in Refs. 26–31 are oriented toward capturing distortions due to lossy source coding only, and thus, they cannot be applied in the case of packet-loss impaired videos.

In Ref. 32, the authors extracted a set of features from the MPEG-2 bitstream and proposed two different modeling approaches: (1) a tree classifier to decide if a packet loss is visible or invisible and (2) a generalized linear model (GLM) to estimate the probability that a packet loss is visible. In Ref. 33, the GLM approach was extended for H.264/AVC bitstreams to model the visibility of individual and multiple packet losses. An application of the proposed GLM scheme to packet prioritization of a video stream, considering factors not only within a packet but also in its vicinity, was suggested in Ref. 34. The visual effect of whole B-frame losses was investigated in Ref. 35. For this purpose, a GLM was used to estimate the probability of the visibility of a B-frame loss and a router was able to decide about which frames to drop in a video transmission scenario, in which the incoming bitrate was higher than the outgoing rate. However, the methods presented in Refs. 32–36 classified packets in a binary mode as visible or invisible based on the viewers' responses to the glitches they spotted. For example, a packet loss was assumed to be visible when the percentage of the viewers that identified an impairment was over a threshold and invisible when this percentage was under a threshold. On the contrary, Argyropoulos et al.[37] introduced a NR bitstream-based model that predicts continuous estimates for the visibility of packet losses, and the impact of the lost packets on perceptual video quality was also studied. However, most of these metrics mainly target the visibility of packet losses and a direct estimation of perceptual quality is not made by also including the features related to video coding.

In addition to bitstream-based features, some approaches include pixel-based features or network-level impairments as well for better prediction performance[11] resulting in hybrid models. In Ref. 9, the proposed model combines the impact of network distortions with quality-related information of the video data. Specifically, the impact of jitter, delay, and packet loss on the video quality is assessed. Secondly, it estimates the impairments incurred due to compression while considering the content characteristics of a video. An analysis of the significance of different loss types (I-, P- or B-slices), video content characteristics, and quantizer scale on the video quality prediction is also presented. Similarly, the method proposed in Ref. 38 makes use of the impacts of spatial and temporal error concealment, missing prediction residuals, and the temporal propagation as a result of motion compensation. Additionally, the contribution of channel distortions that occur with relation to intra-MB prediction and deblocking filter is considered. The ITU-T standardized approach of hybrid video quality is published through ITU-T: J343, which uses bitstream data in addition to processed video sequences.

The NR method presented in Ref. 39 estimates the quality of videos transmitted over wireless networks, using information from MBs of interframe encoded pictures of a video. The proposed method analyzes the impact of both encoding and channel conditions to the video quality degradation by using motion vectors and residual error from the received P-frame and/or B-frame. In addition, in Ref. 40, a quality of experience (QoE) evaluation model was proposed to estimate the end-users' perception on a video streaming service considering different video content types. This QoE model extracts key parameter information directly from degraded video frames in order to estimate the video QoE. A similar NR quality metric for networked video was introduced in Ref. 41 using information extracted from the compressed bitstream only. This metric accounts for picture distortion caused by quantization, quality degradation due to packet losses and error propagation, and temporal effects of the human visual system.

2.1 *Related Work Selected for Comparison*

Keeping in view the usage of test stimuli, the performance of our proposed LASSO models is compared with that of ridge models, two FR metrics, as well as with the following related works. The study presented in Ref. 42 proposes an FR method that uses both singular values and singular vectors as visual features, and a machine learning technique for feature pooling is also introduced. The work presented in Ref. 43 proposes an RR metric that compares the phase and magnitude of the two-dimensional (2-D) discrete Fourier transform of the reference and distorted images in order to compute visual quality. An NR bitstream-based quality metric that considers both the effects of lossy H.264/AVC video encoding and packet losses over Internet Protocol networks is proposed in Ref. 44. In Ref. 45, an NR video quality metric for H.264/AVC video transmissions in packet-based networks is introduced, which uses features from the headers that encapsulate compressed video data. Similarly, in Ref. 46, an enhanced algorithm based on the G.1070 model[47] is developed that compensates for the impact of varying video content characteristics on encoding bitrate. Lastly, genetic programming-based symbolic regression is used in Ref. 48 in order to build a bitstream-based NR model. The used features characterize encoding settings, parameters related to network distortions, and video content.

2.2 *Goals of the Article*

Accordingly, in the context of the aforementioned related works, we propose an approach that directly estimates video quality by employing perceptually important video features, by extending our previous studies presented in Refs. 29–31, and 49. The key points of this article as well as the contributions that it brings are summarized as follows:

1. We propose RR and NR models in order to estimate the perceptual quality of H.264/AVC video sequences, which are affected by packet losses.
2. A variety of features that are expected to have an effect on perceptual video quality are collected in order to be used for building the proposed models. It is worth mentioning that the RR set of features as a whole and the NR set of features as a whole are employed for the first time, while this study also introduces the utilization of 11 new features.
3. LASSO regression[13–15] is utilized in order to indicate the most useful features for making MOS estimations. To the best of our knowledge, this is the first time that LASSO is employed in video quality estimation problems. From the obtained experimental results, we

confirm that this is a very efficient tool for feature selection, while producing accurate quality estimations through the use of sparse models, at the same time. As a baseline, ridge regression[16–18] is used, which requires a much larger number of features, as compared to LASSO for making MOS estimations, even after a feature selection preprocessing.[19]

4. Two different sets of experiments are conducted: (1) a case in which each of the test video sequences is excluded from the training set (for all different test sequences) and (2) a case in which a fixed training and test set are considered. For the second case, we recommend a specific set of regression coefficient values that can be used in practice for the prediction of perceived video quality for other databases.
5. The proposed models exhibit high performance as gauged by different statistical measures. Particularly, they offer impressively high accuracy, nearly perfect monotonicity, and very low estimation errors. In addition, a performance comparison of the proposed approaches is made with a number of related studies. Moreover, the performance statistics for two FR metrics that are oriented toward measuring video quality of digital video systems are explored, namely perceptual evaluation of video quality (PEVQ)[50] and video quality metric (VQM),[51] which are used for comparison. A close inspection of the results reveals that our proposed RR and NR models offer better performance than that of the FR metrics, while the comparative advantage over the related works is apparent in terms of all used performance measures and number of used features.

3 Features Related to Perceptual Video Quality

In this section, we describe the video features that we used in order to model the impact of various impairments on video quality, and we also discuss the motivation of extracting the specific features. Table 1 summarizes these features along with their type. It is worth noting that all of the features described in Table 1 are used to build the RR models, while only features 1 to 47 are used for the development of our NR models.

3.1 *Examined Features*

In H.264/AVC based coding, several coding modes are typically dependent on the content of a video. Each frame of a video is divided into a fixed number of slices, where in this work, each slice consists of a full row of MBs.[52] Mainly, the coding starts with the prediction of one part (block) of a video frame from its adjacent frames so as to eliminate any temporal redundancies. The first frame is intra (I) coded, followed by a predetermined sequence of forward predictive (P) and bidirectional predictive (B) frames, with a periodic recurrence of I frames if required. These predictions can be applied on an MB, i.e., a 16×16 block of pixels or on its subsized blocks. The available information regarding these coding modes provides an estimation of the structural content of a video. The features that we compute from the lossy bitstream and that can be grouped in this category are listed from 1 to 20 in Table 1. Features 1 and 5 are useful for providing relative information on the percentage of blocks whose loss can be more significant as they might be used in the prediction of other blocks. Moreover, more flexibility on the usage of bipredictive coding leads to better compression performance. In this context, we employ feature 13. The percentage of "intra" coded blocks in an "inter" slice may represent rapid change of spatial content in a video and it is captured through feature 4. The percentage of blocks coded as "Skip" indicates the possibility of no need for any residual or motion vector information data that in turn represents the level of SSIM of the content between various frames of a video. Encoding blocks of size 16×16 is preferable as compared to 4×4 because, generally, the use of higher block sizes exhibits better compression performance. Accordingly, features 2, 3, 7 to 12, and 15 to 20 represent the percentages of different block sizes chosen for encoding.

In addition, interframe prediction, which takes advantage of the temporal redundancy between neighboring frames, involves the determination of motion vector information. This information can be used to estimate the relative motion found in the blocks of different frames of a video. Besides using the absolute values of the motion vectors, a number of related statistics were computed so as to better represent the motion content of a video (features 21 to 32 in Table 1). Except for the features 13 to 20 and 31 to 32, which are first proposed in this study, the others have been used in Ref. 31.

Driven by the fact that a packet loss is significantly less visible in still video scenes,[34] we propose the use of feature 33, in order to define if a video slice includes motion or not. Using the motion vector magnitude values, as they were computed by feature 28, we assume that a slice includes motion (NotStill = 1), if its magnitude value is greater than 1/10th of the highest magnitude value of all slices. Similarly, we assume that a slice includes high levels of motion (feature 34) if its magnitude value is greater than 8/10th of the highest magnitude value of all slices. Additionally, features 35 to 36 represent the maximum and mean residual energy over all the MBs of a slice, in which the residual energy for an MB is computed as the sum of squares of its transform coefficients. These additional parameters are used in order to validate whether the calculated motion vectors represent the underlying scene motion well or not. A higher residual energy value implies that the motion vectors probably do not represent the actual scene motion well. If a slice is lost, then even after applying a concealment strategy in order to accurately estimate the lost motion vectors, the resultant slice still differs from the original. Thus, residual energy is one way to assess the magnitude of this difference.[33]

Continuing with the features 37 to 45 in Table 1, they capture the effect of a packet loss in a video sequence, under various aspects. They are all computed from the lossy bitstream, except for feature 45, which is calculated from the reconstructed video sequence after error concealment. Specifically, features 37 to 42[34] model the impact of a packet loss based on its frequency, location, duration, and so on. The use of feature 37 is proposed for the first time as a means of quantifying the severity of distortion introduced within a frame due to the possible slice losses. The vertical location of the lost slice in a frame is represented by feature 38, where its use for quality estimation is motivated by the fact that a lost slice in the middle of a frame can have a different perceptual impact as compared to a lost slice in the top or bottom of a video frame.

Except for feature 38, another content-independent feature that is used to characterize the duration of time an

Table 1 Description of the examined features.

Feature	Description	Type
1. Intra (%)	The percentage of I coded MBs in a slice.	NR
2. I4 × 4 inIslice (%)	The percentage of MBs of size 4 × 4 in an I slice.	NR
3. I16 × 16 inIslice (%)	The percentage of MBs of size 16 × 16 in an I slice.	NR
4. IinPslice (%)	The percentage of I coded MBs in a P slice.	NR
5. P (%)	The percentage of P coded MBs in a slice.	NR
6. PSkip (%)	The percentage of P MBs coded as PSkip in a slice.	NR
7. P16 × 16 (%)	The percentage of P MBs coded with no subpartition of MBs in a slice.	NR
8. P8 × 16 (%)	The percentage of P MBs coded with 8 × 16 and 16 × 8 partitions of MBs in a slice.	NR
9. P8 × 8 (%)	The percentage of P MBs coded with 8 × 8 partition of MBs in a slice.	NR
10. P8 × 8 sub (%)	The percentage of P MBs coded with 8 × 8 in a subpartition of MBs in a slice.	NR
11. P4 × 8 (%)	The percentage of P MBs coded with 4 × 8 and 8 × 4 subpartitions of MBs in a slice.	NR
12. P4 × 4 (%)	The percentage of P MBs coded with 4 × 4 subpartition of MBs in a slice.	NR
13-20. B modes	B modes that correspond to the same features as given in features 5 to 12, but for B coded MBs.	NR
21-22. ΔMV_x, ΔMV_y	The average measures of motion vector difference values for x and y directions in a slice.	NR
23-24. $\text{avg}(\text{MV}_x)$, $\text{avg}(\text{MV}_y)$	The average measures of motion vector values for x and y directions in a slice.	NR
25. MV_0 (%)	The percentage of motion vector values equal to zero for x and y directions in a slice.	NR
26. ΔMV_0 (%)	The percentage of motion vector difference values equal to zero in a slice.	NR
27. Motion Intensity 1	Defined as: $\sum_{i=1}^{M}\sqrt{\text{MV}_{x_i}^2+\text{MV}_{y_i}^2}$ where MV_a, $a\in[x,y]$ represents the average value of motion vector in an MB in a-direction and M is the total number of MBs in a slice.	NR
28. Motion Intensity 2	Defined as $\sqrt{\text{avg}(\text{MV}_x)^2+\text{avg}(\text{MV}_y)2}$.	NR
29-30. $\lvert\text{avg}(\text{MV}_x)\rvert$, $\lvert\text{avg}(\text{MV}_y)\rvert$	The average measures of absolute value of motion vector for x and y directions in a slice.	NR
31. Motion intensity 3	Defined as: $\sum_{i=1}^{M}\sqrt{\lvert(\text{MV}_x)\rvert_i^2+\lvert(\text{MV}_y)\rvert_i^2}$ where MV_a, $a\in[x,y]$ represents the average value of motion vector in an MB in a-direction and M is the total number of MBs in a slice.	NR
32. Motion intensity 4	Defined as: $\sqrt{\lvert\text{avg}(\text{MV}_x)\rvert^2+\lvert\text{avg}(\text{MV}_y)\rvert^2}$.	NR
33. NotStill	Boolean. True, if a slice includes motion.	NR
34. HighMot	Boolean. True, if a slice includes a high motion level.	NR
35-36. MaxResEngy, MeanResEngy	The maximum and mean residual energy over all the MBs of a slice.	NR
37. LostSinFrm	Number of lost slices in a frame.	NR
38. Height	Vertical location of the lost slice within a frame.	NR

Table 1 (*Continued*).

Feature	Description	Type
39. TMDR	Number of frames affected by a lost slice.	NR
40. SpatialExtend	Number of consecutive lost slices in a frame.	NR
41. SpatialExtend2	Boolean. True, if SpatialExtend = 2.	NR
42. Error1Frm	Boolean. True, if TMDR = 1.	NR
43. DistToRef	Distance in frames between the current frame and the reference frame used for concealment. Based on the considered GOP pattern, P frames are concealed using images three frames ago, while both I frames and B frames are concealed using images one frame ago.	NR
44. FarConceal	Boolean. True if $\lvert\text{DistToRef}\rvert \geq 3$.	NR
45. Slice boundary mismatch	Impact of the impairment on slice boundaries.	NR
46-47. SigMean, SigVar	The mean and variance of the slice luminance.	NR
48-49. MeanMSE, MaxMSE	The mean and maximum MSE, over all MBs of a slice.	RR
50-51. MeanSSIM, MinSSIM	The mean and minimum SSIM, over all MBs of a slice.	RR

error persists is feature 39, and features 40 to 42 are also video content-independent and are used since intuitively, they may help in better describing the effect of losing consecutive slices. Moreover, features 43 to 44 are related to the concealment strategy applied to the decoder and particularly, they deal with the distance from the frame that is used as reference for the concealment of a frame impaired with a slice loss. Thus, these features take into account the considered group of pictures (GOP) structure and size. The motivation behind extracting the specific features lies in the fact that when the concealment image is temporally closer to the current image, fewer temporal artifacts occur and thus, reduced impairment visibility is observed compared to the case in which the concealment image is far away from the current image.[53]

As it was discussed earlier, even after applying an error concealment technique, imperfections in the concealed parts of a video cannot be avoided. Thus, when a slice loss occurs, we may have temporal and horizontal discontinuities between the correctly received and concealed slices, which increase the visibility of the impairment.[34] In this work, having detected the location of the lost slice, we applied the slice boundary mismatch metric (feature 45 in Table 1), as it is described in Ref. 53, with the goal of capturing the mismatch on the boundaries between correctly received and concealed slices in the decoded frames, on a pixel-by-pixel basis.

Last, features 46 to 51 in Table 1 are calculated on a pixel-by-pixel basis. Particularly, features 46 to 47 are computed from the compression-and-network-impaired videos, while features 48 to 51 are from the compression-impaired and compression-and-network-impaired versions of a video. The magnitude of distortion induced by a slice loss is also influenced by the presence of luminance masking, that is, the sensitivity of the human visual system to the distortion introduced in darker and brighter image areas. For this purpose, we utilized features 46 to 47 in order to model the mean and variance of the luminance of the signal.

Features 48 to 51 model the mean squared error (MSE) and SSIM metrics, which are commonly used in order to characterize the error amplitude and perceptual quality. In the current study, we precompute the MSE and SSIM values for each MB at the server side. Since it is considered that human attention is mostly drawn to the worst-case errors, in addition to the mean MSE and SSIM values, we also keep the maximum MSE and minimum SSIM values over all MBs in a slice. Afterward, all MSE and SSIM block values are averaged to obtain a representative value for each of them over each slice, and the resulting values, along with the maximum MSE and minimum SSIM for each slice, are next sent to the client's side. Thus, once it is known which slices are actually lost, we are able to know the corresponding MSE and SSIM values. This process of precomputing and transmitting the values from the server to the client renders these features of RR type.[34]

However, both MSE and SSIM metrics, which in the current work play the role of the RR features used for the model's development, present a number of weaknesses.[53] MSE cannot quantify the spatio-temporal characteristics of the error and it implicitly calculates the error size and duration, being unable to capture any information about error location or pattern. Moreover, MSE captures the error between the compression-impaired and compression-and-network-impaired versions of a video, but it does not give any information about the encoded and decoded signals individually. Similarly to MSE, SSIM does not offer any information about the error size or duration. Although it gives an intuition about the signal at the location of the impairment, it does not directly measure the decoded impairment attributes. Therefore, we confirm that the extraction of each of the network-error-related features presented in Table 1 is prudent, as each of them focuses on a different aspect of the effect incurred by a lost slice.

With regard to comparison of the proposed approach of RR-based VQA, our method has some advantages over the standardized RR model called ITU-T J.342[54] in the following ways. J.342 is based on the edge PSNR measurement, which is performed on the edge pixels of the video being transmitted over the ancillary channel. In our case, it is required to transmit a single MSE and a single SSIM value for the whole sequence, and hence, it may require less bandwidth. Our RR features are not dependent on the video content; on the other hand, edge pixels may vary for different contents (spatial details, frame resolution, and so on), requiring less or more bandwidth.

In the used test-stimuli, a slice of a video frame corresponds to a packet. Therefore, considering the impact of a packet loss in terms of data loss on the test-stimuli, it is noted that an integral number of slices are lost as a result of a packet-loss event. In light of this, in this study, the features that are related to the occurrence of a packet loss were computed at the slice level. On the other hand, some features, such as those related to motion vectors, are more suitably computed at the MB level. Hence, we found it reasonable to follow a bottom-up approach for computing most of these features at the MB level and subsequently, an average value was obtained at the slice level. Henceforth, we computed the average values of the slice level features to obtain their values at the frame level. Moreover, the frame-level feature values were averaged further to obtain their values at the video sequence level. For frame-level data to video-level data conversion, we tried Minkowski summation[55] by investigating a large number of Minkowski exponents. However, we confirmed that the overall performance of the estimation models was not significantly improved, and thus, we eventually employed the simple arithmetic mean.

4 Video Quality Estimation Using Linear Regression

The problem of perceptual video quality estimation based on a set of quality-relevant features is solved by building computational models that take the given set of feature values as input and produce appropriate quality estimates. The choice of a particular solution to be used for regression, linear or nonlinear, depends upon the requirements of the problem under consideration as well as the tradeoff preferences between the complexity and performance of a method. However, the theory associated with linear regression is well understood and allows for the construction of different types of easily interpretable, stable, and sparse regression models. In the current study, we propose the use of the LASSO[13–15] regression method.

4.1 *Least Absolute Shrinkage and Selection Operator*

LASSO is a regression method that can be used for both feature selection and computation of regression coefficients. Feature selection is useful when a collection of input features is available, from which we expect to select a small subset for the efficient estimation of a response variable, e.g., the perceptual quality of a video. The particular regression technique is able to effectively address possible issues that arise when the matrix of observations is not of full rank and thus, it is infeasible to be inverted using the ordinary least squares (OLS) method.[16] In addition, LASSO has the benefit of not only shrinking some coefficients close to zero but also setting some others equal to zero, offering feature selection and producing interpretable models. Thus, it combines the stability of ridge[16–18] and interpretability of subset selection, at the same time.

Practically, it minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Therefore, it solves the following minimization problem:

$$\min_{w}\left\{\frac{1}{2}\sum_{i=1}^{n}[y_i - w^{\top}\phi(x_i)]^2 + \frac{\lambda}{2}\sum_{j=1}^{m}|w_i|\right\}. \tag{1}$$

In this equation, the vector y includes the measured quality values for all n observations, which is the total number of videos of the test-stimuli, and w is an $m \times 1$ vector of regression coefficients, including the intercept. The basis function $\phi(x_i)$ is an $m \times 1$ vector at observation x_i, which includes the values for all examined features for a particular video sequence. The shrinkage parameter λ is a positive parameter that controls the amount of the regularization. As λ is increased, an increasing number of regression coefficients becomes equal to zero, while for $\lambda = 0$, no shrinkage is obtained. For the LASSO methodology, the regression coefficients w have no closed form and the solution involves quadratic programming techniques using convex optimization.

4.2 *Model Learning*

The set of features described in Sec. 3 was extracted from the test-stimuli of the Ecole Polytechnique Fédérale de Lausanne (EPFL) and Politecnico di Milano (PoliMi) database.[52] The original SouRCe (SRC) videos were selected for the representation of a variety of spatiotemporal perceptual information, as suggested by ITU-T Rec. P.910.[56] The selected SRCs were in raw progressive format sampled at 4:2:0 ratio of luma and chroma components and were encoded using the H.264/AVC reference software, version JM 14.2,[57] with high profile setting. A GOP structure of IBBP with size 16 was used, while each video had a duration of 10 seconds in length.

The video sequences comprising the EPFL-PoliMi's database of common intermediate format (CIF) resolution are "Mother," "Foreman," "Paris," "News," "Mobile," and "Hall," each of 298 frames at 30 frames per second (fps) and of 4CIF resolution are "Harbour" and "Soccer" of 298 frames at 30 fps, "Parkjoy," "Crowdrun," and "Duckstakeoff" of 250 frames at 25 fps, and "Ice" of 238 frames at 30 fps. For each video sequence, a full row of MBs was coded as a separate slice, while the bitstreams of the coded videos were impaired by a PLR of 0.1%, 0.4%, 1%, 3%, 5%, and 10%. For each PLR and content, two decoded video sequences were obtained, by reading an error pattern from a different starting point. At the decoder, motion compensated error concealment was applied. It should be noted that this database also includes the MOS values as they were collected after subjective experiments separately conducted at EPFL and PoliMi. Further details on the generation of this dataset as well as the testing conditions can be found in Ref. 52.

Before building our model, first, we standardized the values of the input features by calculating their "zscore" values;

that is, we subtracted the "mean" from each feature vector and the obtained values were divided by the "standard deviation" of the particular feature vector. The available data used for the model training and testing consisted of 144 sequences. Specifically, for each of the 12 SRC videos, 12 different realizations of a packet-loss environment were simulated, as mentioned above.

In order to validate the robustness of our estimation models, we ensured a clear distinction between the training and test data such that no content is common between the sets and we followed two slightly different modeling approaches. In the first case, the test set comprised all distorted versions of one SRC sequence (12 sequences) and the training set comprised of the data of test-stimuli generated from the distorted versions of 11 SRCs (132 sequences). This process of splitting the dataset into training and test sets was iterated such that each impaired sequence set from each SRC takes its place on the test set. This procedure is the well-known k-fold cross validation (CV)[58] (in our case, 12-fold CV), in which data are partitioned into k equally sized subsets and an iterative procedure is repeated k times such that $k-1$ subsets are used for training and the remaining one subset is used for testing (validation). In the second case, we considered a 3∶1 ratio between the training and test sets, with the sequences "Mother," "Foreman," "News," "Mobile," "Hall," "Harbour," "Soccer," "Crowdrun," and "Duckstakeoff," comprising the training set and "Paris," "Ice," and "Parkjoy" the test set. Therefore, 108 sequences were used for training and the remaining 36 sequences were used for testing.

Once the training data were selected, the next important step was the initialization of the regularization parameter λ in Eq. (1). For both approaches, a number of 100 different λ values slightly above 0 were tested, in which each different λ results in a different number of selected features. Also, the MSE between the subjective and estimated MOS values was calculated. Therefore, by simultaneously examining the sparsity, i.e., the number of features that are assigned zero regression coefficients, as well as the estimation accuracy in terms of MSE, we selected the λ value among all 100 values of λ that gives the best tradeoff of these conditions. Thus, using the chosen λ value, we trained our models and we obtained the values for the regression coefficients.

The obtained regression coefficient values were applied to the data of the test set in order to get the MOS estimations. Using the estimated values, we were able to evaluate models' performance in comparison with the subjective MOS values. Algorithm 1 summarizes the methodology adopted in this work for the k-fold case in order to develop our proposed models. A similar procedure is also followed for the fixed dataset partition into training and test sets, with the difference that this process is followed just once as there is a single training set.

In addition, it is worth mentioning that the use of LASSO avoids the problem of overfitting because: (1) it builds a simple model and (2) it performs regularization. It is generally admitted that too complex models are prone to overfitting and thus, they give poor estimations. In this context, LASSO regression is able to autonomously perform feature selection within its learning process, producing estimations at the same time. Also, apart from feature selection, the specific method is able to perform L_1 regularization. Regularization works well when we have a lot of features, each of which contributes a bit to the response variable estimation and it deters overfitting since the magnitude of the regression coefficients is reduced and thus, a smoother curve for fitting the data is obtained.

Algorithm 1 Model development.

```
loop
  if (a SRC is not tested) then
    Split the dataset into training set and test set, such that the test
    set includes all impaired versions of the same SRC.
    Execute exclusively on the training set.
      a. Perform LASSO regression.
      b. Determine the optimal $\lambda$ value of Eq. (1).
      c. Using the optimal $\lambda$ value, train the whole training set of
         LASSO model.
      d. Get regression coefficient estimates.
    Apply regression coefficient estimates on the test set.
    Get video quality estimations.
    Evaluate performance.
  end if
end loop
```

5 Experimental Results

In this section, we present an analysis of the obtained results in order to evaluate the performance of the RR and NR LASSO models. Moreover, for the sake of comparison with the MOS estimations given by LASSO, we applied ridge regression,[16–18] which does not perform any kind of feature selection in its standard form. Ridge is an extension of the OLS regression method[16] and is able to improve the OLS estimates by allowing a little bias in order to reduce the variance of the estimated values, offering a good generalization capability to unseen data. Practically, it solves the following minimization problem:

$$\min_{w}\left\{\frac{1}{2}\sum_{i=1}^{n}[y_i - w^{\mathrm{T}}\phi(x_i)]^2 + \frac{\lambda}{2}\|w\|^2\right\}. \tag{2}$$

Therefore, ridge attempts to tradeoff the goodness-of-fit, as it is described by the first term of the above equation, and the penalty, as it is described by the second term of the same equation.

As no feature selection is performed using the aforementioned regression method, this implicates the risk of harming the estimations, when irrelevant or noisy features are employed. Due to this, a feature selection procedure[19] preceded ridge so as to keep only those features that are the most influential toward making MOS estimations. Specifically, we performed SFS and particularly forward feature selection

Table 2 Performance of LASSO, ridge and reference FR metrics using MOS collected by PoliMi for both training and test for the *k*-fold case.[52]

	CIF						4CIF					
Method	Test Sequence	λ	# Features	PCC	SROCC	RMSE	Test Sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Foreman	$1e-05$	9	0.977	0.958	0.297	Crowdrun	$1e-05$	11	0.981	0.972	0.223
NR LASSO		0.2674	2	0.979	0.972	0.284		0.2677	3	0.969	0.986	0.283
RR ridge		$1e-05$	9	0.986	0.986	0.230		$1e-05$	10	0.988	0.986	0.179
RR LASSO		0.7014	2	0.985	0.986	0.243		0.7178	2	0.987	0.986	0.187
PEVQ		—	—	0.983	0.963	0.792		—	—	0.966	0.986	0.317
VQM		—	—	0.971	0.979	1.548		—	—	0.991	0.986	0.343
NR ridge	Hall	$1e-05$	8	0.945	0.937	0.420	Duckstakeoff	$1e-05$	4	0.968	0.993	0.313
NR LASSO		0.2711	3	0.965	0.979	0.337		0.2714	2	0.973	1.000	0.289
RR ridge		$1e-05$	11	0.976	0.923	0.278		$1e-05$	6	0.996	1.000	0.113
RR LASSO		0.7099	2	0.980	0.944	0.258		0.7070	2	0.991	1.000	0.163
PEVQ		—	—	0.944	0.818	0.753		—	—	0.994	0.996	0.329
VQM		—	—	0.940	0.895	1.064		—	—	0.988	0.996	0.489
NR ridge	Mobile	$1e-05$	5	0.960	0.979	0.356	Harbour	$1e-05$	6	0.978	0.916	0.203
NR LASSO		0.2697	2	0.979	0.972	0.259		0.2754	2	0.978	0.930	0.204
RR ridge		$1e-05$	7	0.993	0.965	0.153		$1e-05$	5	0.973	0.923	0.224
RR LASSO		0.7070	2	0.993	0.979	0.153		0.7239	1	0.981	0.930	0.186
PEVQ		—	—	0.969	0.937	0.404		—	—	0.960	0.930	0.419
VQM		—	—	0.979	0.972	0.575		—	—	0.974	0.930	0.509
NR ridge	Mother	$1e-05$	7	0.945	0.937	0.318	Ice	$1e-05$	14	0.962	0.958	0.338
NR LASSO		0.2757	3	0.929	0.937	0.359		0.2706	3	0.970	0.965	0.303
RR ridge		$1e-05$	16	0.962	0.965	0.265		$1e-05$	10	0.977	0.979	0.267
RR LASSO		0.7381	2	0.966	0.944	0.252		0.7133	2	0.979	0.979	0.256
PEVQ		—	—	0.967	0.944	0.803		—	—	0.977	0.972	0.368
VQM		—	—	0.963	0.930	2.003		—	—	0.975	0.972	0.633
NR ridge	News	$1e-05$	14	0.965	0.965	0.345	Parkjoy	$1e-05$	6	0.954	0.951	0.317
NR LASSO		0.2704	1	0.969	0.972	0.328		0.2749	2	0.972	0.965	0.247
RR ridge		$1e-05$	9	0.987	0.979	0.212		$1e-05$	3	0.975	0.979	0.234
RR LASSO		0.7093	2	0.992	0.979	0.164		0.7180	2	0.984	0.979	0.188
PEVQ		—	—	0.976	0.972	0.343		—	—	0.979	0.972	0.446
VQM		—	—	0.980	0.979	1.078		—	—	0.981	0.979	0.550

Table 2 (*Continued*).

CIF							4CIF					
Method	Test Sequence	λ	# Features	PCC	SROCC	RMSE	Test Sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Paris	1*e* – 05	8	0.972	0.916	0.320	Soccer	1*e* – 05	4	0.991	0.979	0.162
NR LASSO		0.2677	2	0.978	0.944	0.286		0.2712	3	0.987	0.993	0.192
RR ridge		1*e* – 05	13	0.990	0.972	0.188		1*e* – 05	4	0.996	1.000	0.108
RR LASSO		0.7035	2	0.989	0.972	0.205		0.7120	2	0.996	0.993	0.103
PEVQ		—	—	0.976	0.951	0.587		—	—	0.989	0.996	0.382
VQM		—	—	0.970	0.951	0.833		—	—	0.991	0.986	0.672
NR ridge	Average	—	9	0.961	0.949	0.343	Average	—	8	0.973	0.962	0.261
NR LASSO		—	2	0.967	0.963	0.309		—	3	0.975	0.973	0.253
RR ridge		—	11	0.982	0.965	0.221		—	6	0.984	0.978	0.188
RR LASSO		—	2	0.984	0.967	0.213		—	2	0.986	0.978	0.181
PEVQ		—	—	0.969	0.930	0.614		—	—	0.977	0.976	0.377
VQM		—	—	0.967	0.951	1.183		—	—	0.983	0.975	0.533

(FFS) (“`sequentialfs`” function in MATLAB) in order to select the appropriate subsets of features from the initial RR and NR sets, respectively, that best estimate the actual MOS. Starting from an empty feature set, candidate subsets are created by sequentially adding each of the features not yet selected, in order of importance. For each candidate feature subset, ridge regression was applied in order to estimate the output values, and finally, the MSE between the actual and estimated MOS was returned. This process continued until adding more features did not further decrease MSE. Thus, the specific RR and NR feature subsets that resulted in minimum MSE were selected. Applications of FFS using linear regression models in similar problems of video quality estimation can be found in Refs. 59 and 60.

In addition, a comparison with related approaches is also reported in this section. In order to evaluate the models' performance, the following measures, as recommended by video quality experts group (VQEG)[61] were used.

- Linearity: The Pearson linear correlation coefficient (PCC) is used to describe the linearity of the estimation.
- Monotonicity: The Spearman rank order correlation coefficient (SROCC) is used to describe the monotonicity of the estimation.
- Accuracy: The root MSE (RMSE) is used to describe the accuracy of the estimation.

The values of PCC and SROCC lie in the range $[-1,1]$, where values closer to 1 represent high positive correlation.

5.1 *Results and Discussion*

As a result of the test setup described in Sec. 3, a number of simulations were performed for the RR and NR sets of video features, using the MOS values collected by both PoliMi and EPFL. It holds that subjective MOS values are usually compressed at the ends of the rating scale (0 and 5), while this is not the case for objective video quality models that are unable to mimic this weakness of subjective data. Therefore, following the VQEG report on validation of objective video quality models,[62] a third order monotonic mapping function was applied on the estimated values of our models before the computation of the performance measures.

Table 2 presents the obtained results for both RR and NR models using ridge and LASSO, when the PoliMi MOS values are used in both the training and test phases. Besides the performance measures, the chosen λ values and the number of used features for the prediction of the quality value of each test sequence using each proposed model are also mentioned. It is to be noted that for ridge, we set $\lambda = 10^{-5}$ as being a typical small positive value able to improve the conditioning of the problem and reduce the variance of the MOS estimates. Each cell of the table cites the results when 12 impaired versions of a specific SRC sequence are used as the test dataset (the training dataset comprises the 132 impaired sequences of the remaining 11 SRC sequences) and the bottom cell shows the arithmetic mean (average) of the performance over all the SRCs (of all cells) with resolution CIF and 4CIF separately. In the same table, the related performance of PEVQ and VQM metrics is also mentioned. PEVQ is an FR metric that is a part of the ITU-T Recommendation J.247.[50] VQM is also an FR metric that has been largely adopted in the research community for taking quality estimates.[51] In our experiments, we calculated VQM using the MSU video quality measurement tool, Version 4.3 Beta Professional, available in Ref. 63. In an effort to be fair when comparing the estimated and subjective MOS, we scaled

Table 3 Performance of LASSO and ridge models using MOS collected by PoliMi for training and MOS collected by EPFL for test for the *k*-fold case.[52]

	CIF						4CIF					
Method	Test Sequence	λ	# Features	PCC	SROCC	RMSE	Test Sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Foreman	$1e-05$	10	0.958	0.930	0.403	Crowdrun	$1e-05$	6	0.948	0.951	0.337
NR LASSO		0.2674	2	0.982	0.979	0.266		0.2677	3	0.955	0.993	0.312
RR ridge		$1e-05$	6	0.981	0.979	0.272		$1e-05$	4	0.986	0.993	0.178
RR LASSO		0.7014	2	0.983	0.979	0.261		0.7178	2	0.986	0.993	0.178
NR ridge	Hall	$1e-05$	2	0.928	0.916	0.479	Duckstakeoff	$1e-05$	6	0.942	0.986	0.420
NR LASSO		0.2711	3	0.939	0.958	0.442		0.2714	2	0.964	0.993	0.332
RR ridge		$1e-05$	7	0.972	0.930	0.300		$1e-05$	3	0.992	0.993	0.158
RR LASSO		0.7099	2	0.975	0.937	0.284		0.7070	2	0.991	0.993	0.166
NR ridge	Mobile	$1e-05$	10	0.940	0.986	0.487	Harbour	$1e-05$	9	0.977	0.972	0.224
NR LASSO		0.2697	2	0.973	0.979	0.328		0.2754	2	0.969	0.972	0.262
RR ridge		$1e-05$	5	0.993	0.986	0.174		$1e-05$	9	0.976	0.944	0.231
RR LASSO		0.7070	2	0.992	0.972	0.186		0.7239	1	0.977	0.944	0.227
NR ridge	Mother	$1e-05$	8	0.961	0.965	0.257	Ice	$1e-05$	3	0.952	0.965	0.376
NR LASSO		0.2757	3	0.936	0.930	0.329		0.2706	3	0.935	0.965	0.433
RR ridge		$1e-05$	12	0.963	0.979	0.251		$1e-05$	10	0.976	0.993	0.267
RR LASSO		0.7381	2	0.979	0.923	0.192		0.7133	2	0.979	0.993	0.248
NR ridge	News	$1e-05$	12	0.966	0.972	0.370	Parkjoy	$1e-05$	5	0.959	0.972	0.298
NR LASSO		0.2704	1	0.952	0.993	0.440		0.2749	2	0.971	0.986	0.251
RR ridge		$1e-05$	17	0.993	0.979	0.171		$1e-05$	7	0.993	0.993	0.120
RR LASSO		0.7093	2	0.994	0.972	0.158		0.7180	2	0.993	0.993	0.124
NR ridge	Paris	$1e-05$	10	0.964	0.916	0.383	Soccer	$1e-05$	10	0.983	0.965	0.217
NR LASSO		0.2677	2	0.972	0.944	0.337		0.2712	3	0.981	0.986	0.231
RR ridge		$1e-05$	13	0.971	0.958	0.344		$1e-05$	7	0.983	0.986	0.181
RR LASSO		0.7035	2	0.969	0.958	0.354		0.7120	2	0.991	0.986	0.155
NR ridge	Average	—	9	0.953	0.948	0.396	Average	—	7	0.960	0.969	0.312
NR LASSO		—	2	0.959	0.964	0.357		—	3	0.963	0.983	0.304
RR ridge		—	10	0.979	0.969	0.252		—	7	0.985	0.984	0.189
RR LASSO		—	2	0.982	0.957	0.239		—	2	0.986	0.984	0.183

Table 4 Performance of LASSO and ridge models using MOS collected by EPFL for training and MOS collected by PoliMi for test[52] for the *k*-fold case.

CIF							4CIF					
Method	Test sequence	λ	# Features	PCC	SROCC	RMSE	Test Sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Foreman	$1e-05$	10	0.961	0.916	0.390	Crowdrun	$1e-05$	9	0.973	0.972	0.264
NR LASSO		0.2222	3	0.981	0.972	0.272		0.2283	3	0.969	0.986	0.283
RR ridge		$1e-05$	15	0.988	0.993	0.216		$1e-05$	9	0.973	0.972	0.264
RR LASSO		0.5419	2	0.988	0.986	0.221		0.5576	2	0.985	0.986	0.120
NR ridge	Hall	$1e-05$	7	0.944	0.930	0.424	Duckstakeoff	$1e-05$	12	0.993	1.000	0.152
NR LASSO		0.2253	3	0.965	0.979	0.337		0.2259	2	0.977	1.000	0.265
RR ridge		$1e-05$	7	0.978	0.923	0.271		$1e-05$	3	0.996	1.000	0.106
RR LASSO		0.5481	2	0.978	0.923	0.266		0.5473	2	0.995	1.000	0.119
NR ridge	Mobile	$1e-05$	12	0.985	0.979	0.362	Harbour	$1e-05$	5	0.972	0.923	0.228
NR LASSO		0.2221	3	0.979	0.972	0.257		0.2280	3	0.978	0.930	0.202
RR ridge		$1e-05$	10	0.990	0.972	0.178		$1e-05$	12	0.983	0.930	0.177
RR LASSO		0.5408	2	0.993	0.979	0.154		0.5556	2	0.981	0.930	0.187
NR ridge	Mother	$1e-05$	5	0.920	0.923	0.381	Ice	$1e-05$	5	0.961	0.965	0.345
NR LASSO		0.2295	3	0.929	0.937	0.360		0.2261	3	0.970	0.965	0.304
RR ridge		$1e-05$	11	0.959	0.944	0.276		$1e-05$	8	0.961	0.965	0.345
RR LASSO		0.5716	1	0.966	0.944	0.252		0.5507	2	0.979	0.979	0.255
NR ridge	News	$1e-05$	4	0.972	0.979	0.314	Parkjoy	$1e-05$	15	0.975	0.951	0.234
NR LASSO		0.2225	3	0.970	0.972	0.324		0.2289	3	0.943	0.965	0.352
RR ridge		$1e-05$	8	0.987	0.986	0.210		$1e-05$	3	0.976	0.979	0.229
RR LASSO		0.5423	2	0.987	0.979	0.214		0.5539	2	0.986	0.979	0.176
NR ridge	Paris	$1e-05$	8	0.972	0.916	0.319	Soccer	$1e-05$	8	0.989	0.979	0.176
NR LASSO		0.2217	3	0.976	0.944	0.296		0.2254	3	0.986	0.986	0.198
RR ridge		$1e-05$	11	0.987	0.965	0.218		$1e-05$	9	0.990	0.979	0.163
RR LASSO		0.5414	2	0.990	0.972	0.197		0.5501	2	0.996	0.993	0.107
NR ridge	Average	—	8	0.954	0.941	0.365	Average	—	9	0.977	0.965	0.233
NR LASSO		—	3	0.967	0.963	0.308		—	3	0.970	0.972	0.267
RR ridge		—	10	0.982	0.964	0.228		—	7	0.980	0.971	0.214
RR LASSO		—	2	0.983	0.964	0.217		—	2	0.987	0.978	0.174

Table 5 Regression coefficient values for the *k*-fold case.

			NR LASSO				RR LASSO		
			Intercept	B (%)	TMDR	SpatialExtend	Intercept	TMDR	MinSSIM
PoliMi	CIF	Foreman	2.5776	0.0163	−0.7980	—	2.5768	−0.0481	0.3849
		Hall	2.5705	0.0205	−0.6572	−0.1567	2.5760	−0.0710	0.3639
		Mobile	2.5519	—	−0.6574	−0.1652	2.5534	−0.0102	0.4221
		Mother	2.5729	0.0132	−0.7344	−0.0971	2.5966	−0.0427	0.3908
		News	2.5523	—	−0.8205	—	2.5619	−0.0769	0.3603
		Paris	2.5640	0.0127	−0.8143	—	2.5759	−0.0054	0.4253
	4CIF	Crowdrun	2.5694	0.0068	−0.7737	−0.0510	2.5547	−0.0364	0.3837
		Duckstakeoff	2.5819	0.0473	−0.7902	–	2.5768	−0.0402	0.3824
		Harbour	2.5244	–	−0.7072	−0.1320	2.5235	–	0.4323
		Ice	2.5619	0.0052	−0.8038	−0.0156	2.5472	−0.0560	0.3686
		Parkjoy	2.5845	–	−0.8132	−0.0266	2.5792	−0.0534	0.3706
		Soccer	2.5677	0.0075	−0.7207	−0.0982	2.5635	−0.0497	0.3708
EPFL	CIF	Foreman	2.2503	0.0456	−0.7269	−0.0957	2.2539	−0.0430	0.5597
		Hall	2.2360	0.0403	−0.6865	−0.1589	2.2424	−0.0460	0.5609
		Mobile	2.2248	0.0235	−0.6404	−0.2027	2.2297	−0.0526	0.5501
		Mother	2.2547	0.0499	−0.6519	−0.2015	2.3005	—	0.6245
		News	2.2175	0.0199	−0.7574	−0.0870	2.2398	−0.0313	0.5684
		Paris	2.2311	0.0380	−0.7861	−0.0518	2.2439	−0.0303	0.5695
	4CIF	Crowdrun	2.2506	0.0386	−0.8327	−0.0227	2.2235	−0.0151	0.6098
		Duckstakeoff	2.2286	0.2054	−0.7126	—	2.2382	−0.0599	0.5670
		Harbour	2.2130	0.0145	−0.6804	−0.1922	2.2114	−0.0492	0.5699
		Ice	2.2491	0.0428	−0.7944	−0.0474	2.2257	−0.0279	0.5960
		Parkjoy	2.2673	0.0202	−0.7613	−0.1125	2.2574	−0.0800	0.5434
		Soccer	2.2476	0.0365	−0.6905	−0.1536	2.2412	−0.446	0.5768

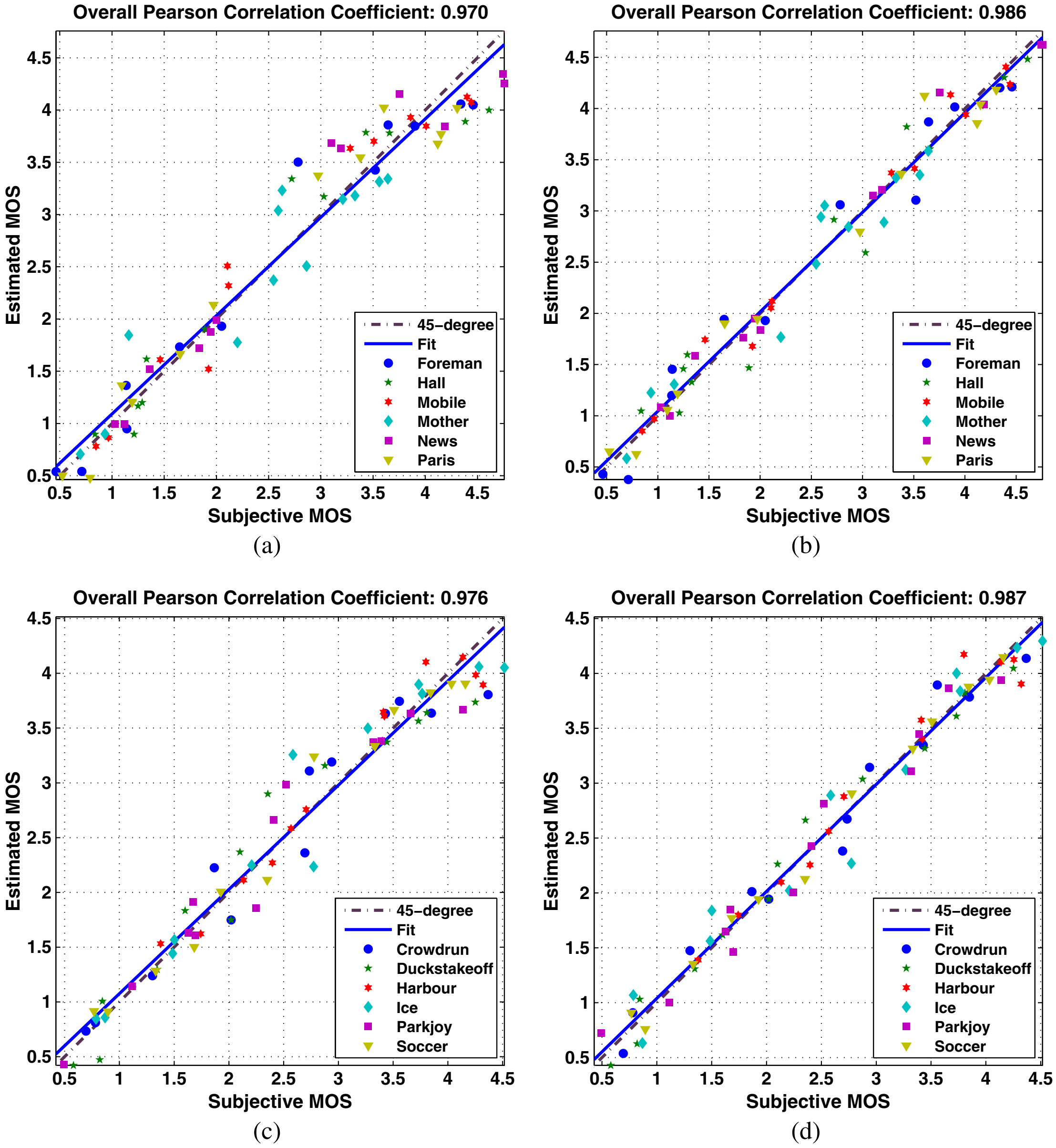
PEVQ and VQM values in the range [0, 5]. In addition, since for the VQM, the smaller the value the better the video quality, we “reversed” these values to follow the trend of MOS. It holds that comparing RR and NR models against FR metrics is a challenging task, as FR metrics have far more data to process for estimating quality. Nonetheless, as it turns out, the proposed models perform equally well, or somewhat better than the considered FR metrics. The advantage of our proposed models is more evident mainly in terms of fairly lower values of estimation error (RMSE).

The same information for the obtained results of both RR and NR models using ridge and LASSO is also provided in Tables 3 and 4. Specifically, in Table 3, we cite the results when the MOS obtained by PoliMi is used for models training and the MOS collected by EPFL is used in order to test the models. The opposite comes true for the results in Table 4, i.e., the MOS from EPFL is used to train the models and the MOS from PoliMi is used for models testing. In addition, in these tables, the values for the PEVQ and VQM values are omitted. Apparently, they are the same as presented in Table 2, due to the fact that they do not depend on the considered training set. Also from these tables (Tables 3 and 4), we confirm that the proposed models present a similar performance with that of Table 2, while their strong point against the FR metrics is also the much better estimation accuracy they provide in terms of RMSE.

One salient aspect of comparing the performance of ridge and LASSO is the level of accuracy and sparsity offered by each solution. Observing the performance results in Tables 2–4, we do not confirm an advantage of a particular regression method over the other, since their statistics are similar in both CIF and 4CIF resolutions. However, for all examined cases (with only one isolated exception), it can be construed that LASSO models use far less features than those employed by ridge for making quality estimations.

In fact, the values of the regression coefficients are considered as an indication of feature selection or not. If the coefficient associated with a certain feature acquires a zero value, this means that the specific feature is excluded from the estimation process. On the contrary, a feature is selected, if it is assigned a nonzero regression coefficient value. As the values of the input features are normalized to the same scale, a higher value of a coefficient implies higher significance of the related feature, and vice versa. Moreover, the features that are associated with positive-signed coefficients are considered to cause an increase in quality if their values are increased. By contrast, the features associated with negative-signed coefficients are considered to decrease quality if their values are increased.

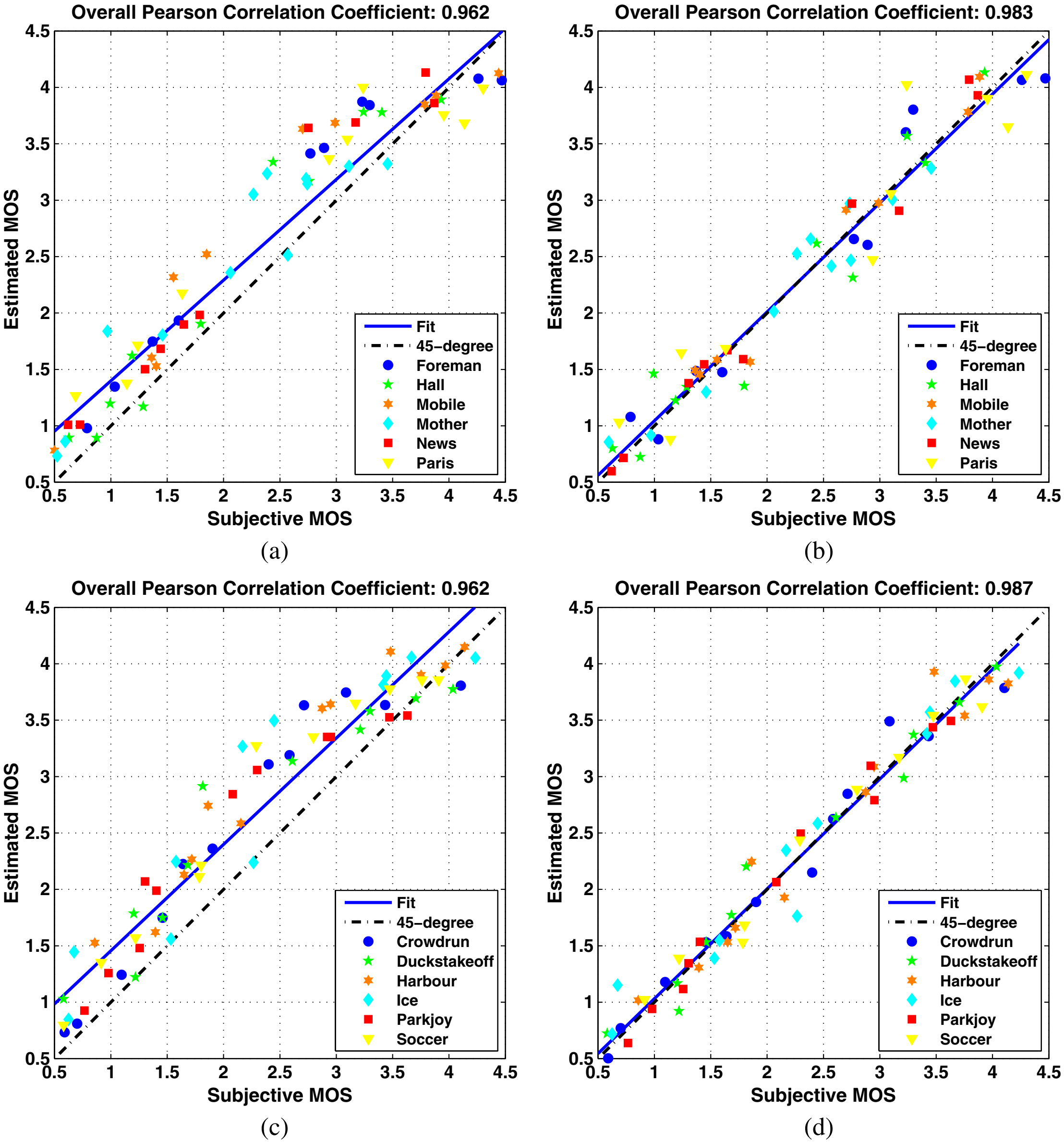
From the investigation of the results in Tables 2 and 3, we observe that the selected λ values as well as the number of selected features, for each corresponding test sequence, for both the NR and RR LASSO models, are the same. This is a reasonable conclusion as our models, in both cases



(a) (b)

(c) (d)

Fig. 2 Overall performance of NR LASSO and RR LASSO (a, b) for CIF and (c, d) for 4CIF, using MOS collected by PoliMi for both training and test, for the *k*-fold case.



(a)

(b)

(c)

(d)

Fig. 3 Overall performance of NR LASSO and RR LASSO (a, b) for CIF and (c, d) for 4CIF, using MOS collected by PoliMi for training and MOS collected by EPFL for test, for the *k*-fold case.

in Tables 2 and 3, have been trained on the same dataset. On the other hand, this does not hold for the RR and NR ridge models. Although we have set the same λ values, we notice that a different number of features is selected for each separate test sequence. This is attributed to the fact that the feature selection method we have used takes into account the considered test set at each time. In this context, Table 5 below presents the regression values that are assigned to each feature for each of the RR and NR LASSO models, when the MOS results from both the PoliMi and EPFL are employed for models' training. The features that do not appear on this table are assigned a zero regression coefficient value. Furthermore, on the same table, we present the corresponding intercept values.

A close inspection of the results provided by Table 5 reveals that for the NR LASSO models, three features (at the worst case) out of the 45 initially extracted ones are enough to make extremely precise MOS estimations, while for the RR LASSO models, this number shrinks to two features (at the worst case) out of the 51 in total, offering equally high prediction efficiency, for both CIF and 4CIF resolutions, and regardless of the considered MOS values used for models' training. Furthermore, it is intriguing to examine the specific features that are actually selected. As it can be seen from the same table, for the NR case, the selected features are B[%], TMDR, and SpatialExtend, while for the RR case, this list includes TMDR and MinSSIM. Features TMDR and SpatialExtend are closely related to the error propagation; the first captures the error propagation to the frames that depend on it, while the second one checks for impairments inside the same frame. In addition B[%], which describes the number of MBs coded as bipredictive in a slice also has a great significance in MOS estimation as it gathers information about the internal coding structure of the sequences. Among the features selected by the NR models, TMDR is definitely the most important one as it is selected by all different training

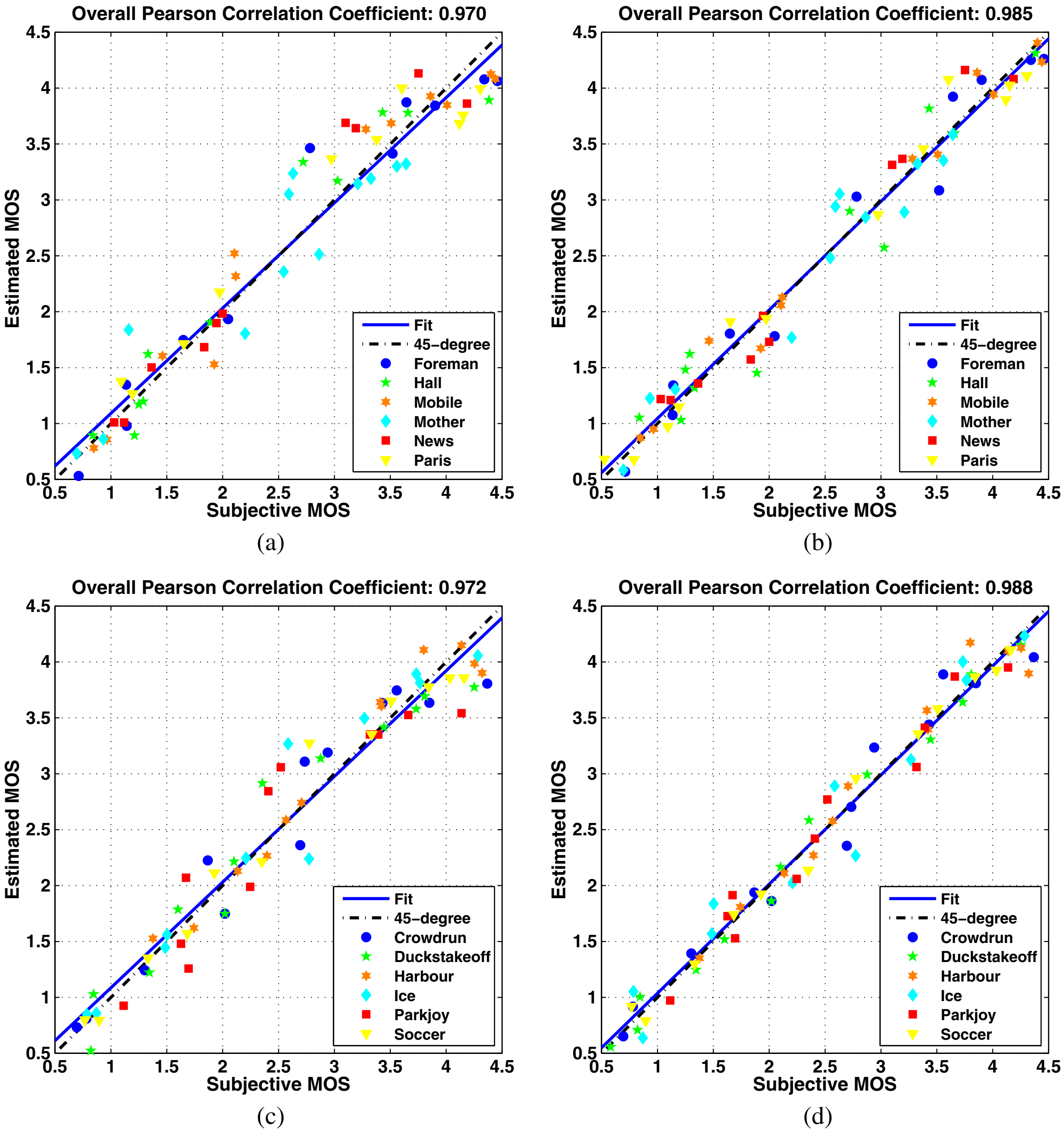


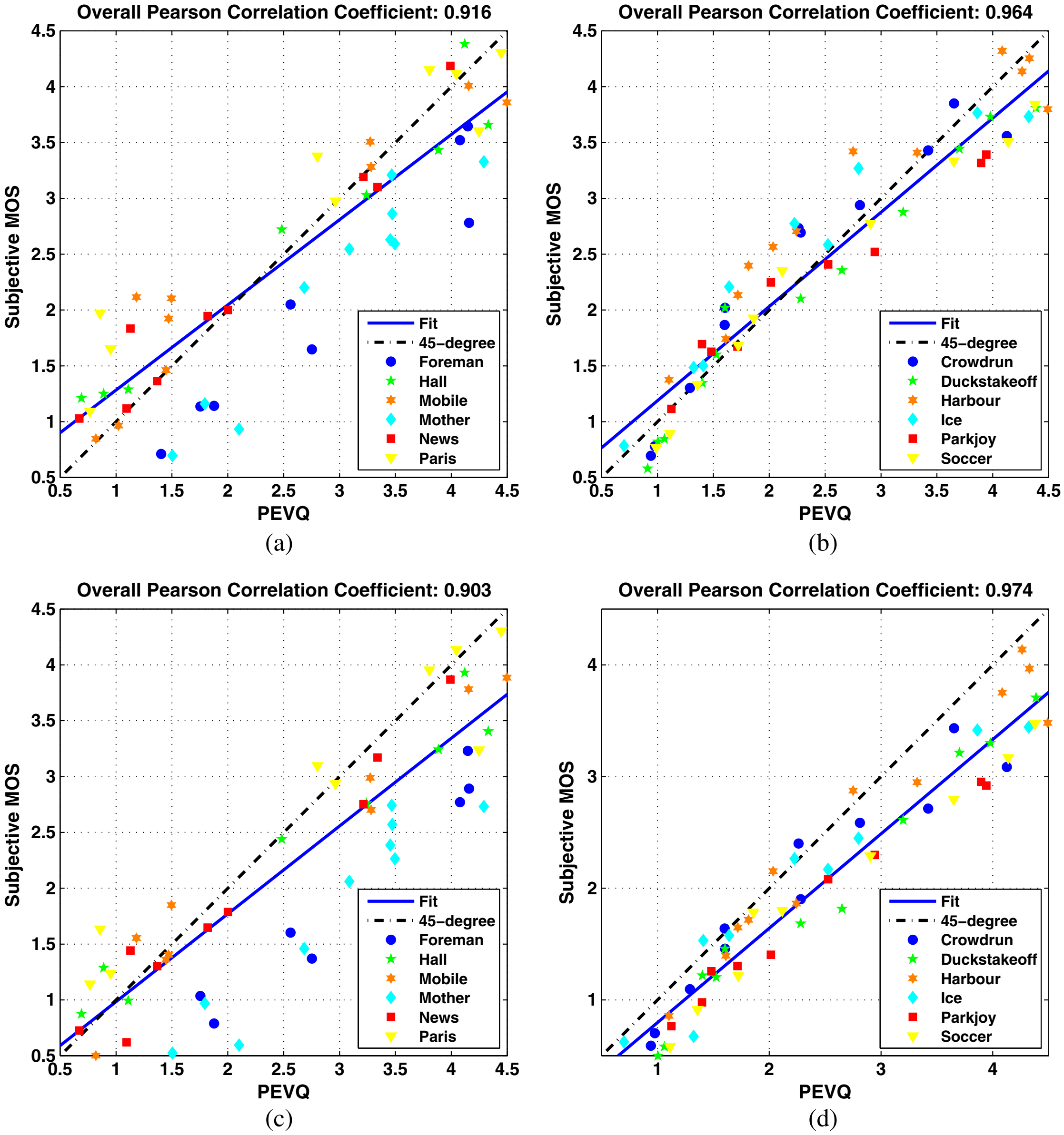
Fig. 4 Overall performance of NR LASSO and RR LASSO (a, b) for CIF and (c, d) for 4CIF, using MOS collected by EPFL for training and MOS collected by PoliMi for test, for the *k*-fold case.

sets; either the MOS from PoliMi or EPFL have been used for models' training.

For the RR LASSO models, TMDR is the one of the two selected features (interestingly, TMDR is selected by both the NR and RR models) and also MinSSIM is always selected in all cases presented in Table 5. In fact, MinSSIM keeps information about the error at the specific location of the loss and is able to describe the spatial extent of the error.

Commenting on the values of the regression coefficients, we observe that all employed features are assigned regression coefficient values that are very close to zero. In addition, TMDR has a negative impact toward MOS quality estimation as opposed to the other selected features. From Table 5, we can see that B[%] and SpatialExtend take almost zero values in most of the cases, while higher are the values assigned to TMDR, meaning that this feature affects more on the MOS estimation process. On the other hand, TMDR is assigned smaller values when it comes to the RR LASSO models. In this case, the MinSSIM feature is dominant, as it is assigned much higher values (with regards to its absolute value) than TMDR. Lastly, looking at the intercept values, we confirm that for both the NR and RR models, for each test sequence, they are almost identical, especially in the case of using the MOS values from the same institution for models' training.

Regarding the case in which ridge along with FFS is applied, the number of selected features, when each different SRC is tested, ranges between 2 and 15 for the NR case and between 3 and 17 for the RR case, examining all Tables 2–4. The corresponding list with all different selected features for the NR case includes the features 1 to 4, 6 to 10, and 12 to 45, which means that 43 out of 45 features in total are selected, excluding from the list features 5 and 11. Similarly for the RR case, the list with the selected features includes the features 1 to 3, 5 to 8, 10 to 28, 30, 34 to 43, and 45 to 51, i.e., 44 out of 51 features are selected, excluding the features 4, 9,



(a)

(b)

(c)

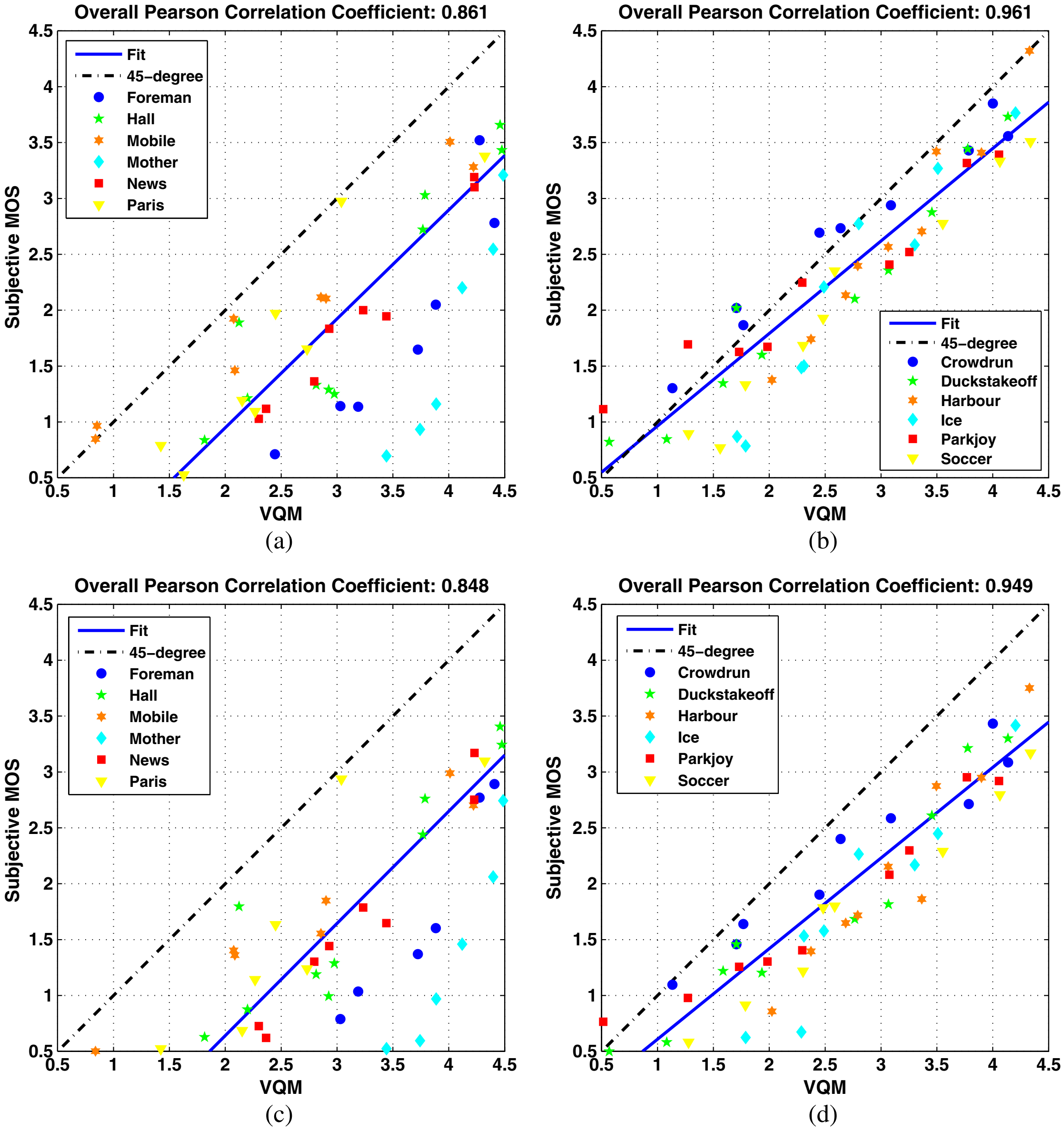
(d)

Fig. 5 Overall performance of PEVQ (a, c) for CIF and (b, d) for 4CIF, as compared to MOS collected by PoliMi and EPFL, respectively, for the *k*-fold case.

29, 31, 32, 33 and 44. The presentation of the individual regression coefficients values as they result after applying ridge are beyond the scope of this paper and for this purpose, we omit them from here. Also, it is interesting to observe that using the NR or the RR ridge models, all of the features presented in Table 1 contribute to the estimation of MOS, as the features excluded from NR ridge are included in RR ridge and vice versa.

Therefore, we confirm that using LASSO, we are able to reject a large number of features in order to achieve a desired level of sparsity, maintaining a high level of accuracy at the same time. LASSO keeps the advantage of providing much more sparse solutions as compared to ridge combined with FFS, and the same three and two specific features (at the worst case) that are selected in the NR and RR cases, respectively, are able to estimate quality with high precision, irrespective of the considered training set. On the other hand, when ridge with FFS is applied, a much larger number of the initially extracted features are kept, and overall, all of the features of Table 1 are employed for the MOS estimation of all of the examined video sequences, justifying our belief that the features we propose have an obvious impact on perceptual video quality. Additionally, it is interesting to highlight that all three features selected by NR LASSO are extracted from the lossy bitstream and do not require its decoding. Therefore, the lower computational complexity of LASSO is another benefit over ridge using FFS and hence, it can be used as an efficient solution for video quality estimation as well as feature selection, maintaining impressively good performance statistics.

Regarding the performance comparison of the RR and NR models, from the individual results of each SRC in Tables 2–4, it is observed that the RR models have slightly better performance than the corresponding NR models, as indicated through the correlation coefficients as well as the estimation errors, regardless of the regression method used.



(a)

(b)

(c)

(d)

Fig. 6 Overall performance of VQM (a, c) for CIF and (b, d) for 4CIF, as compared to MOS collected by PoliMi and EPFL, respectively, for the *k*-fold case.

This inference about the effectiveness of the RR models is also verified by the average results, depicted on the bottom cells of the same tables. However, we have to mention that the estimation accuracy in NR cases is very promising as well and hence, the proposed list of NR features (see Table 1) presents an acceptable solution for reference-free quality estimation of video transmissions over lossy networks.

In addition to the earlier mentioned statistics in Tables 2–4, the overall performance of RR LASSO and NR LASSO models for each corresponding case of the aforementioned tables is shown as scatter plots in Figs. 2–4, for both CIF and 4CIF resolutions. The values of the "overall" performance, as indicated in these plots, are obtained by comparing the values of estimated and subjective MOS, when all examined sequences of a specific spatial resolution are considered as a whole. The scatter plots not only indicate a very high overall performance in each case but also highlight the superiority of RR models over the corresponding NR ones, which are able to manage the MOS estimation in a better way. For instance, from all these figures, we notice that the quality estimation of "Mother" CIF sequence or "Ice" 4CIF sequence seems to be difficult for NR LASSO in all examined cases, while RR LASSO manages their estimation more effectively. One plausible reason behind this behavior could be that both of these video sequences have low spatiotemporal information indices[52] and the additional features selected by the RR approach could better represent perceptual quality of such videos. Thus, the slight advantage of using an RR model over an NR one in terms of performance is more obvious from these plots, at the cost of maintaining an ancillary channel and the risk of a possible failure in RR data delivery to the receiver's end.

Figures 5 and 6 below show the "overall" performance of PEVQ and VQM, respectively, and how they correlate with the subjective MOS. We present separate plots for each

Table 6 Performance of LASSO and ridge models using MOS collected by PoliMi for both training and test,[52] for fixed training and test sets.

Method	Test sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Paris	$1e-05$	6	0.982	0.951	0.259
NR LASSO		0.2242	2	0.975	0.944	0.301
RR ridge		$1e-05$	3	0.985	0.965	0.233
RR LASSO		0.7065	2	0.990	0.972	0.194
NR ridge	Ice	$1e-05$	6	0.965	0.972	0.326
NR LASSO		0.2242	2	0.970	0.965	0.303
RR ridge		$1e-05$	3	0.977	0.979	0.266
RR LASSO		0.7065	2	0.980	0.979	0.249
NR ridge	Parkjoy	$1e-05$	6	0.962	0.951	0.288
NR LASSO		0.2242	2	0.973	0.965	0.246
RR ridge		$1e-05$	3	0.975	0.979	0.234
RR LASSO		0.7065	2	0.986	0.979	0.177
NR ridge	Average	—	6	0.970	0.958	0.291
NR LASSO		—	2	0.973	0.958	0.283
RR ridge		—	3	0.979	0.974	0.244
RR LASSO		—	2	0.985	0.977	0.207
PEVQ		—	—	0.977	0.965	0.467
VQM		—	—	0.975	0.967	0.672

Table 7 Performance of LASSO and ridge models using MOS collected by PoliMi for training and MOS collected by EPFL for test,[52] for fixed training and test sets.

Method	Test sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Paris	$1e-05$	5	0.963	0.951	0.387
NR LASSO		0.2242	2	0.970	0.944	0.347
RR ridge		$1e-05$	3	0.964	0.958	0.380
RR LASSO		0.7065	2	0.965	0.958	0.377
NR ridge	Ice	$1e-05$	5	0.932	0.965	0.444
NR LASSO		0.2242	2	0.935	0.965	0.433
RR ridge		$1e-05$	3	0.979	0.993	0.250
RR LASSO		0.7065	2	0.978	0.993	0.254
NR ridge	Parkjoy	$1e-05$	5	0.953	0.972	0.317
NR LASSO		0.2242	2	0.968	0.986	0.262
RR ridge		$1e-05$	3	0.993	0.993	0.122
RR LASSO		0.7065	2	0.993	0.993	0.124
NR ridge	Average	—	5	0.949	0.963	0.383
NR LASSO		—	2	0.958	0.965	0.347
RR ridge		—	3	0.979	0.981	0.251
RR LASSO		—	2	0.979	0.981	0.252

specific resolution and we examine both cases of using the MOS from both PoliMi and EPFL. Comparing the performance of PEVQ and VQM with the proposed LASSO models, we infer that both NR and RR LASSO models far exceed these FR metrics under the "overall" context, especially in the cases of CIF resolution, where PEVQ and VQM are relatively weak in producing accurate MOS predictions.

In the next part of this section, we present the corresponding results for the case in which a specific set of sequences is always used for models' development and thus, the rest of the sequences of the database are used for testing the models. Specifically, Tables 6–8 present the results for both RR and NR models using both ridge and LASSO. In Table 6, the models have been trained and tested on MOS collected by PoliMi; in Table 7, the MOS collected by PoliMi have been used in the training set and the MOS collected by EPFL have been used in the test set; and last, in Table 8, the MOS collected by EPFL have been used in the training set and the MOS collected by PoliMi have been used in the test set. Moreover, the same tables provide information about the chosen λ values and the number of used features for each test sequence and proposed model. Last but not least, Table 6 at the bottom cell presents the average performance over all tested sequences for PEVQ and VQM.

Next, Table 9 shows the regression coefficient values that are assigned to the features that contribute to the MOS estimation. The features that are missing from this table are assigned zero regression coefficient values. The same table also presents the intercepts that are employed by each model, when both the MOS values from the PoliMi and EPFL are employed for the training of the models. From this table, we confirm that the same features that were employed by each separate test sequence in Table 5 are also used in this case, highlighting their significance in MOS estimation. Also, it is interesting to note that the specific values do not depend on a specific training set and are universal; thus, they can be used for the prediction of MOS of other databases. In addition, these values present a very slight variation depending on if the considered MOS values used to train the model are those of the PoliMi or the EPFL institutions.

The conclusions that arise by examining the results of this table are aligned with the conclusions derived by the results in Table 5. The only difference is that for the NR LASSO model of fixed training and test sets, the SpatialExtend feature is not selected, as in the present case we emphasized more obtaining an even sparser prediction model.

Table 8 Performance of LASSO and ridge models using MOS collected by EPFL for training and MOS collected by PoliMi for test,[52] for fixed training and test sets.

Method	Test Sequence	λ	# Features	PCC	SROCC	RMSE
NR ridge	Paris	$1e-05$	17	0.960	0.944	0.384
NR LASSO		0.2259	2	0.975	0.944	0.301
RR ridge		$1e-05$	9	0.988	0.965	0.212
RR LASSO		0.6561	2	0.990	0.972	0.196
NR ridge	Ice	$1e-05$	17	0.969	0.986	0.306
NR LASSO		0.2259	2	0.970	0.965	0.303
RR ridge		$1e-05$	9	0.974	0.979	0.279
RR LASSO		0.6561	2	0.979	0.979	0.253
NR ridge	Parkjoy	$1e-05$	17	0.941	0.951	0.358
NR LASSO		0.2259	2	0.973	0.965	0.246
RR ridge		$1e-05$	9	0.925	0.979	0.400
RR LASSO		0.6561	2	0.986	0.979	0.175
NR ridge	Average	—	17	0.957	0.960	0.349
NR LASSO		—	2	0.973	0.958	0.283
RR ridge		—	9	0.962	0.974	0.297
RR LASSO		—	2	0.985	0.977	0.208

From the provided results in Tables 6–8, one can notice that the λ values for the NR LASSO models are always the same, regardless of the sequence that is tested each time. The same holds also for the corresponding RR LASSO models. This is attributed to the fact that the λ values have been determined just once for a fixed training set and can be used for the prediction of MOS of every sequence. Specifically, for Tables 6–8, the MOS values from PoliMi have been used for models' training. By contrast, for Table 8, the MOS values from EPFL have been employed in the training phase of the models, and therefore, a different selection for the λ values of both the NR and RR LASSO models has been performed.

For this set of experiments, we have employed even sparser NR LASSO solutions. In more detail, from all Tables 6–9, we notice that just two features for both the NR and RR LASSO models are capable of producing extremely high MOS estimations. Although fewer features are employed as compared to the corresponding NR and RR ridge models preceded by FFS, the performance of the proposed LASSO models is usually marginally better, as it appears from both the separate test sequence results as well as the average results at the bottom cells of Tables 6–8. Similarly, with Tables 2–4, the slight advantage of the RR models over the corresponding NR ones is also evident from Tables 6–8 in terms of all examined measures of performance. Moreover, Table 6 at the bottom cell mentions the average performance of PEVQ and VQM metrics, over all the considered test sequences. We observe that our RR LASSO models keep a higher performance in terms of all examined measures of performance, while our NR LASSO models guarantee far smaller RMSE as compared to the corresponding values of PEVQ and VQM. Similarly, with the k-fold case, we omit the presentation of the average PEVQ and VQM results from Tables 7 and 8 as they are the same, due to the fact that they do not depend on the considered training set.

Table 9 Regression coefficient values for fixed training and test sets.

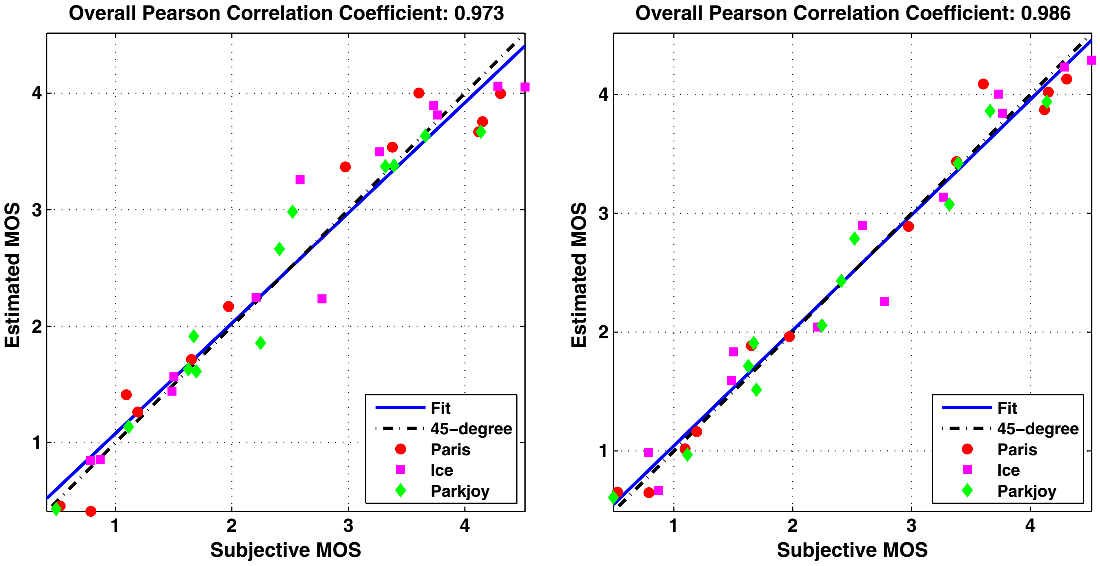
NR LASSO			RR LASSO		
	Coefficients			Coefficients	
Factors	PoliMi	EPFL	Factors	PoliMi	EPFL
Intercept	2.5828	2.2677	Intercept	2.5740	2.2553
B (%)	0.0266	0.0310	TMDR	−0.0585	−0.0317
TMDR	−0.8592	−0.8622	MinSSIM	0.3737	0.4707

In the following, Figs. 7–9 illustrate better the advantage of the RR LASSO models over the NR ones, as it was mentioned in the previous paragraph. In these figures, the values of the "overall" performance are obtained by comparing the values of estimated and subjective MOS, when all sequences of the test set are considered as a whole. Each of these figures corresponds to a different consideration for the employed MOS at each time, i.e., either they have been collected by PoliMi or EPFL. As it is shown, the RR LASSO estimations are closer to the actual MOS measurements as compared to the corresponding NR LASSO ones, although the solutions provided by the NR LASSO models are strongly efficient and thus acceptable. Last, Figs. 10 and 11 present the correlation of the PEVQ and VQM metrics with the actual MOS from both PoliMi and EPFL. A comparison between the efficiency of our proposed NR models and that of PEVQ and VQM reveals that when the MOS from the same institution is used for testing, both our RR and NR models are superior to both PEVQ and VQM. We could also say that PEVQ is able to produce measurements that are closer to the actual MOS ratings as compared to the corresponding measurements from VQM.

5.2 *Comparison with Related Work*

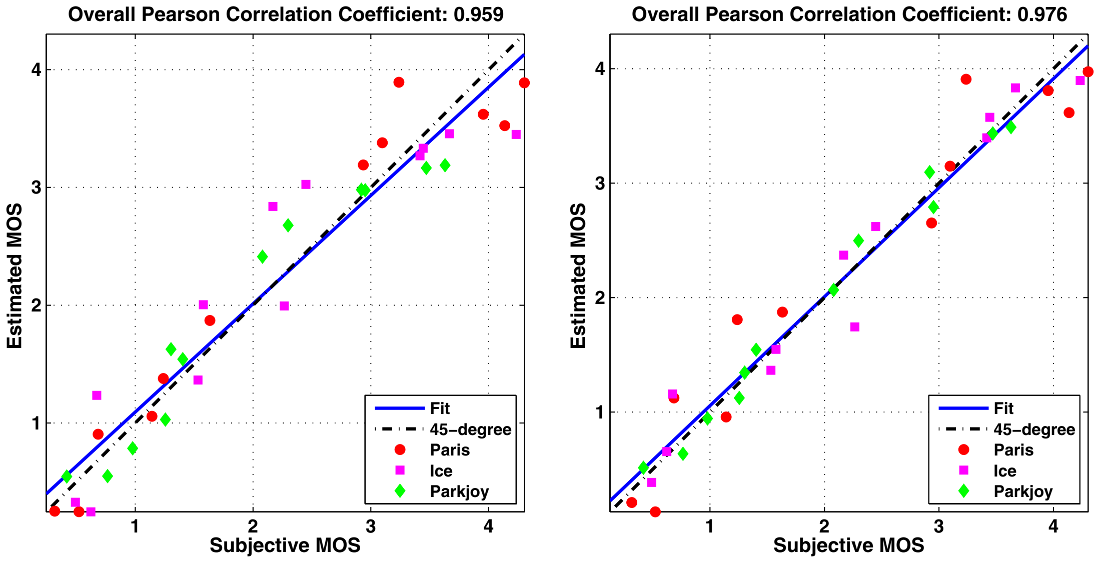
In this part of the article, we compare the results produced in this study for the k-fold case, with the results of existing publications that address video quality estimation problems in FR, RR, and NR modes. It is to be noted that Table 10 includes the results of the performance measures as they were calculated in "overall" fashion, that is considering estimated and subjective MOS values for all examined sequences as a whole, for each particular resolution. Also, the same table gives an intuition about the average number of features required by each employed model for the estimation of MOS.

From Fig. 8(b) of the work presented in Ref. 42, we observe that the PCC and SROCC values using the EPFL-PoliMi



(a) (b)

Fig. 7 Overall performance of (a) NR LASSO and (b) RR LASSO, using MOS collected by PoliMi for both training and test, for fixed training and test sets.

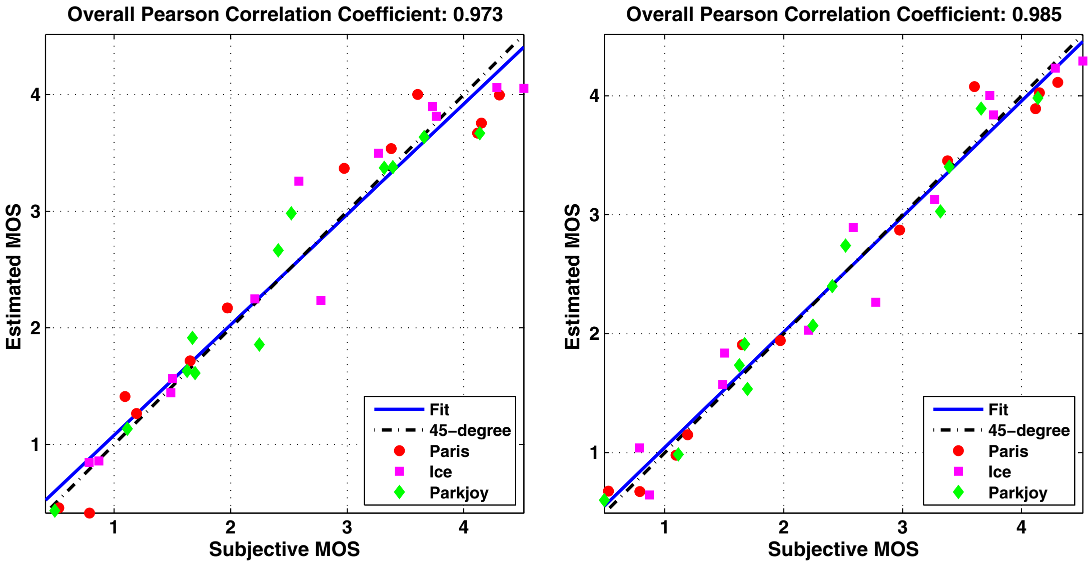


(a) (b)

Fig. 8 Overall performance of (a) NR LASSO and (b) RR LASSO, using MOS collected by PoliMi for training and MOS collected by EPFL for test, for fixed training and test sets.

video database[52] are between 0.85 and 0.95, for all proposed Q-mentioned FR metrics (Qvector, Qcsiq, Qtid, Qlive). These results were generated by using the singular values and vectors so as to quantify the visual distortions. Despite the fact that we propose RR and NR models, and thus, the task of making estimations is more challenging as compared to a FR model, from Table 10, we infer that our NR LASSO model offers PCC values equal to or higher than 0.960 and SROCC values equal to or higher than 0.970, and our RR LASSO model offers 0.981 and 0.974 as the least values of PCC and SROCC.

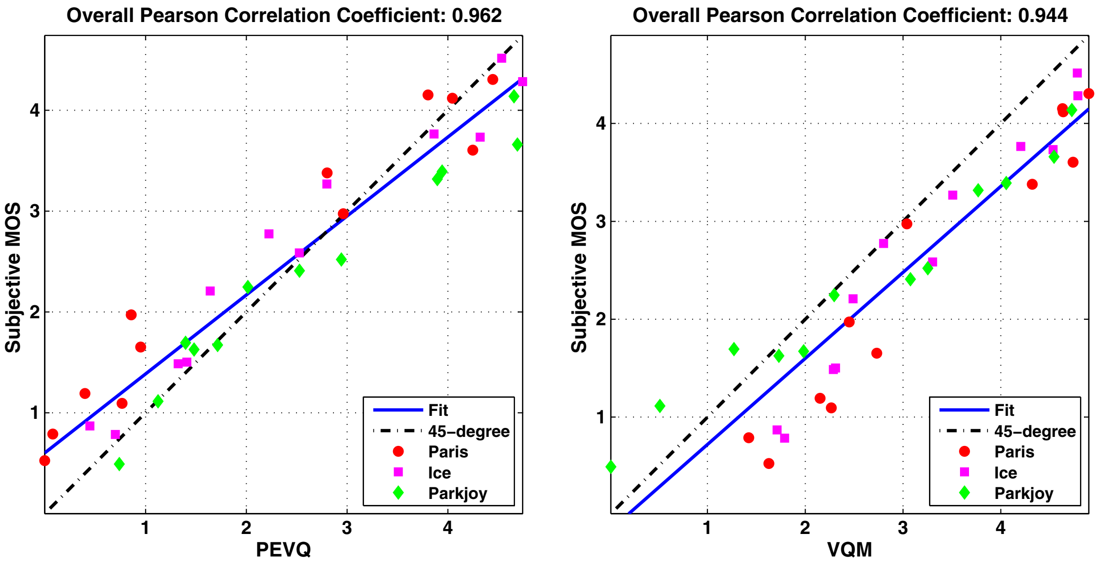
Moreover, the performance of the Fourier transform-based RR model proposed in Ref. 43 for its variant Q_{combined} that offers the best results is compared against our proposed models in the same table (Table 10). In Ref. 43, the basic idea is the comparison of the phase and magnitude of the reference and distorted images so as to compute the quality score. From the provided experimental results, we confirm the superiority of the RR LASSO model that we propose in all examined measures of performance. Interestingly, the estimation error of our model is nearly equal to half of the corresponding amount of Q_{combined} using on average over all tested sequences the same number of features (two) with the features used for the development of Q_{combined}. In addition, Table 10 also depicts the results of the RR ridge model when FFS is preceded. In this case, we can see that ridge achieves marginally (negligibly) better performance as compared to LASSO, but it requires six



(a)

(b)

Fig. 9 Overall performance of (a) NR LASSO and (b) RR LASSO, using MOS collected by EPFL for training and MOS collected by PoliMi for test, for fixed training and test sets.

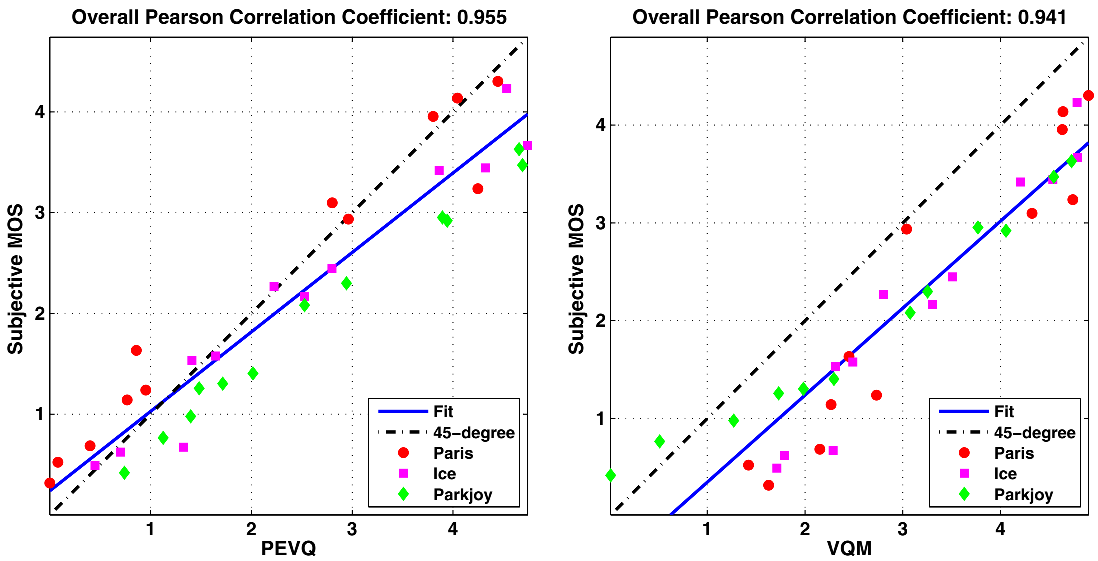


(a)

(b)

Fig. 10 Overall performance of (a) PEVQ and (b) VQM, as compared to MOS collected by PoliMi.

times more features as compared with the corresponding LASSO case. It is to be noted that the experiments in Refs. 42 and 43 were performed such that the proposed models were trained on a database different from what they used for validation of these models. However, the following references used the same database for training and validation. Furthermore, another work that utilizes the EPFL-PoliMi database[52] and specifically, the CIF resolution sequences to assess the performance of the proposed NR metric is the one presented in Ref. 44. In that work, the MOS values collected by PoliMi are used, while the best proposed model is called "frame-type and error pattern dependent packet-loss model," as denoted by "FE-PLM" in Table 10, which uses five features in total for making perceptual quality estimations. It can be seen that our proposed NR LASSO model is better in terms of all examined statistics compared to those of Ref. 44. One of the underlying reasons behind this difference in performance can be the fact that we use a variety of features to capture various characteristics of a video including the impact of packet losses. On the other hand, the models in Ref. 44 are based on the assumption that visual quality can always be exponentially related to the PLR which, in practice, may not hold in varying bitrates and different contents.[64] Also in this case, the proposed NR LASSO model not only gathers better performance statistics as compared to the FE-PLM model but also it requires on average less than half the number of the features required by FE-PLM. Moreover, it is worth mentioning that the NR ridge model combined with FFS also gathers better statistics as compared to the FE-PLM model, but it requires many



(a)

(b)

Fig. 11 Overall performance of (a) PEVQ and (b) VQM, as compared to MOS collected by EPFL.

Table 10 Comparison of the overall performance of the proposed models with ridge models and other related works for the *k*-fold case.

	Based on MOS values by PoliMi				Based on MOS values by EPFL		
CIF resolution							
Metric	NR ridge	NR LASSO	FE-PLM[44]		RR ridge	RR LASSO	Q_{Combined}[43]
PCC	0.963	0.970	0.95		0.983	0.981	0.944
SROCC	0.957	0.970	0.95		0.978	0.974	0.930
RMSE	0.345	0.311	0.43		0.244	0.259	0.446
Ave. Feat. Num	9	2	5		12	2	2
4CIF resolution							
Metric	NR ridge	NR LASSO	G.1070E[46]	$\text{SLR}_{\text{IP}} + \text{SLR}_{\text{B}}$[45]	NR ridge	NR LASSO	G.1070E[46]
PCC	0.973	0.976	0.93	0.963	0.962	0.960	0.926
SROCC	0.974	0.977	0.91	—	0.970	0.976	0.93
RMSE	0.268	0.256	0.373	0.337	0.314	0.325	0.533
Ave. Feat. Num	8	3	3	2	6	2	3

more features for making estimations. By contrast, NR LASSO outperforms NR ridge in terms of both performance and number of features used for estimating video quality.

Similarly, the NR model presented in Ref. 45 was evaluated using MOS values collected at PoliMi.[52] In order to design the model, the authors assumed that PLR and MOS can be characterized by a two-region piecewise linear relationship. Based on this assumption, a number of variants of the basic NR model was proposed, which differ mainly on the type of data used for estimating losses introduced by the network. The results that we considered from Ref. 45 are based on the quality estimation using the $\text{SLR}_{\text{IP}} + \text{SLR}_{\text{B}}$ model variant (based on slice loss rate of I/P slices and B slices) that offers the best results. The conclusion derived after looking at the results is that also in this case, the NR LASSO model achieves better performance in terms of all measures of performance. However, it is to be noted that LASSO utilizes one feature more for making estimations as compared to the model proposed in Ref. 45. On the other hand, we should notice that in this study, we propose a one-region linear model in contrast to Ref. 45, where a two-region piecewise linear model is employed. Similarly, the

NR ridge model is able to estimate video quality more accurately, as compared to Ref. 45, while it requires significantly more features for making estimations. Therefore, also in this case, the NR LASSO model's advantage is obvious.

An enhanced version of the ITU-T Recommendation G.1070: Opinion model for video-telephony applications,[47] called G.1070E, can be found in Ref. 46. The estimation accuracy of the G.1070E model is validated using the 4CIF resolution sequences of the database presented in Ref. 52, with the MOS data collected from the subjective tests, conducted both at the EPFL and PoliMi institutions. Specifically, in Ref. 46, the estimation models take into account the video bit rate, frame rate, and PLR so as to measure video quality and they were trained on a large variety of CIF resolution sequences, other than those included in the EPFL-PoliMi's database, which were compressed at various bitrates and were impaired with different PLRs. Comparing the performance of the proposed NR LASSO model with the G.1070E[46] model, we easily perceive a clear advantage of LASSO and considerably better performance in terms of all presented measures of performance; either the PoliMi or the EPFL MOS values are used. Especially for the case when the EPFL MOS is employed, our proposed model is able to produce better performance statistics, requiring on average fewer features compared to G.1070E, at the same time. Regarding the NR ridge model, it also surmounts in perfomance the model proposed in Ref. 46 but it has the disadvantage of requiring a much larger number of features as compared to Ref. 46. Therefore, the advantage of LASSO is prominent also in this case, despite the fact that for the case of EPFL, output ridge offers slightly better MOS estimations.

Lastly, we studied the performance achieved by a genetic programming-based NR regression model presented in Ref. 48. In that work, the authors validated the performance of their proposed model, exploiting eight different features that are influential for modeling perceptual video quality, by considering the video quality estimates and subjective MOS values together, for both the CIF and 4CIF resolution sequences of the EPFL-PoliMi's database.[52] Accordingly in this case, our NR LASSO model offers PCC and SROCC values equal to 0.973 and 0.975, respectively, as compared to the corresponding values offered by Ref. 48, which are equal to 0.881 and 0.883, respectively. Besides the better statistics achieved by our proposed model, it is important to point out that NR LASSO uses less than half the number of the features employed by the model of Ref. 48.

Despite the fact that our proposed models gather a number of advantages and are able to surmount other related works, in our future plans, we aim to extend their use on a broader set of databases with different characteristics than the one used in this work. In this context, we plan to examine the generalization capabilities of our models in sequences with other resolutions than CIF and 4CIF, for various frame rates, as well as to investigate their performance when a different GOP structure and length is utilized. Interesting would also be the case of observing the models' behavior when different packet sizes are taken into account.

6 Conclusions

In this study, we investigated a fairly large variety of video features for estimating video quality. These features include different attributes related to perceptual quality and encompass the impacts of coding and network impairments on H.264/AVC encoded sequences. The vast majority of these features can be computed without any access to the original video and hence, they are applicable to design an NR model of quality estimation. The rest of the features can be precomputed and sent to the client's end for providing RR information of the original video. Based on these features, we propose RR and NR linear models of quality estimation, by employing the LASSO regression method. LASSO was investigated for its capability to make MOS estimations as well as feature selection at the same time. For comparison purposes, we applied sequential FFS using Ridge as the regression method, so as to get a baseline performance. The simulation results reveal that LASSO is able to achieve exceptionally high and marginally better performance statistics as compared to ridge using FFS, utilizing just two features, for both the RR and NR cases. Interestingly, both of the features required by the NR LASSO model for estimating MOS are extracted from the lossy bitstream, without the need for decoding. This means that significantly less computational complexity is involved in the feature extraction process, rendering the model practical for real-time applications. In addition, the proposed LASSO models outperform a number of existing FR, RR, and NR techniques used for video quality estimation, in terms of both performance statistics and required features. In the future, we aim to test the generalization capabilities of our models in more diverse databases with different encoding and transmission characteristics.

References

1. "Cisco visual networking index: global mobile data traffic forecast update, 2014–2019," Cisco white paper, http://www.cisco.com (2015).
2. ITU-R radio communication sector of ITU, "Recommendation ITU-R BT.500-13: methodology for the subjective assessment of the quality of television pictures," 2012, http://www.itu.int/rec/R-REC-BT.500-13-201201-I (05 May 2016).
3. ITU-T Telecommunication standardization sector of ITU, "ITU-T Recommendation P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," 2016, https://www.itu.int/rec/T-REC-P.913/en (05 May 2016).
4. S. Winkler, *Digital Video Quality: Vision Models and Metrics*, 1st ed., John Wiley & Sons Ltd., West Sussex, England (2005).
5. M. Shahid et al., "Crowdsourcing based subjective quality assessment of adaptive video streaming," in *Sixth Int. Workshop on Quality of Multimedia Experience* (2014).
6. S. Winkler, "Video quality and beyond," in *Proc. of European Signal Processing Conf.*, Vol. 446, pp. 150–153 (2007).
7. M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, **50**, 312–322 (2004).
8. K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features," in *17th IEEE Int. Conf. on Image Processing*, pp. 3229–3232 (2010).
9. S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Trans. Broadcast.* **54**, 660–668 (2008).
10. A. Raake et al., "IP-based mobile and fixed network audiovisual media services," *IEEE Signal Process. Mag.* **28**(6), 68–79 (2011).
11. M. Shahid et al., "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP J. Image Video Process.* **2014**, 40 (2014).
12. A. Reibman, V. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. Multimedia* **6**, 327–334 (2004).
13. R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc., Ser. B* **58**, 267–288 (1994).
14. R. Tibshirani, "The LASSO method for variable selection in the Cox model," *Stat. Med.* **16**(4), 385–395 (1997).
15. M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.* **20**(3), 389–403 (2000).

16. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, New Jersey (2006).
17. D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *Am. Stat.* **29**, 3–20 (1975).
18. A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics* **12**, 55–67 (1970).
19. F. J. Ferri et al., "Comparative study of techniques for large-scale feature selection," in *Pattern Recognition in Practice IV*, pp. 403–413, Elsevier Science B.V. (1994).
20. F. Yuan and E. Cheng, "Reduced-reference metric design for video quality measurement in wireless application," in *11th IEEE Int. Conf. on Communication Technology*, pp. 641–644 (2008).
21. R. Soundararajan and A. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.* **23**, 684–694 (2013).
22. M. Masry, S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. Circuits Syst. Video Technol.* **16**, 260–273 (2006).
23. T. Shanableh, "No-reference PSNR identification of MPEG video using spectral regression and reduced model polynomial networks," *IEEE Signal Process. Lett.* **17**, 735–738 (2010).
24. T. Shanableh, "Prediction of structural similarity index of compressed video at a macroblock level," *IEEE Signal Process. Lett.* **18**, 335–338 (2011).
25. M. Slanina, V. Ricny, and R. Forchheimer, "A novel metric for H.264/AVC no-reference quality assessment," in *14th Int. Workshop on Systems, Signals and Image Processing, 2007 and 6th EURASIP Conf. focused on Speech and Image Processing, Multimedia Communications and Services*, pp. 114–117 (2007).
26. C. Keimel et al., "No-reference video quality metric for HDTV based on H.264/AVC bitstream features," in *IEEE Int. Conf. on Image Processing*, pp. 3325–3328 (2011).
27. S.-O. Lee, K.-S. Jung, and D.-G. Sim, "Real-time objective quality assessment based on coding parameters extracted from H.264/AVC bitstream," *IEEE Trans. Consum. Electron.* **56**, 1071–1078 (2010).
28. M. Ries, O. Nemethova, and M. Rupp, "Motion based reference-free quality estimation for H.264/AVC video streaming," in *2nd Int. Symp. on Wireless Pervasive Computing* (2007).
29. A. Rossholm and B. Lövström, "A new low complex reference free video quality predictor," in *10th IEEE Workshop on Multimedia Signal Processing*, pp. 765–768 (2008).
30. M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor," in *4th Int. Congress on Image and Signal Processing*, Vol. 1, pp. 517–521 (2011).
31. M. Shahid, A. Rossholm, and B. Lövström, "A no-reference machine learning based video quality predictor," in *Fifth Int. Workshop on Quality of Multimedia Experience*, pp. 176–181 (2013).
32. S. Kanumuri et al., "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. Multimedia* **8**, 341–355 (2006).
33. S. Kanumuri et al., "Predicting H.264 packet loss visibility using a generalized linear model," in *IEEE Int. Conf. on Image Processing*, pp. 2245–2248 (2006).
34. T.-L. Lin et al., "A versatile model for packet loss visibility and its application to packet prioritization," *IEEE Trans. Image Process.* **19**, 722–735 (2010).
35. Y.-L. Chang, T.-L. Lin, and P. Cosman, "Network-based H.264/AVC whole-frame loss visibility model and frame dropping methods," *IEEE Trans. Image Process.* **21**, 3353–3363 (2012).
36. N. Staelens et al., "Viqid: a no-reference bit stream-based visual quality impairment detector," in *2nd Int. Workshop on Quality of Multimedia Experience*, pp. 206–211 (2010).
37. S. Argyropoulos et al., "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *3rd Int. Workshop on Quality of Multimedia Experience*, pp. 31–36 (2011).
38. M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Trans. Multimedia* **11**, 932–946 (2009).
39. Y. Yang et al., "A no-reference video quality metric by using inter-frame encoding characters," in *14th Int. Symp. on Wireless Personal Multimedia Communications*, pp. 1–5 (2011).
40. Y. Shen et al., "QoE-based evaluation model on video streaming service quality," in *IEEE Globecom Workshops*, pp. 1314–1318 (2012).
41. F. Yang et al., "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Trans. Circuits Syst. Video Technol.* **20**, 1544–1554 (2010).
42. M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," *IEEE Trans. Syst., Man Cybern. B* **42**, 347–364 (2012).
43. M. Narwaria et al., "Fourier transform-based scalable image quality measure," *IEEE Trans. Image Process.* **21**, 3364–3377 (2012).
44. M. Chin, T. Brandão, and M. Queluz, "Bitstream-based quality metric for packetized transmission of H.264 encoded video," in *19th Int. Conf. on Systems, Signals and Image Processing*, pp. 312–315 (2012).
45. J. Ascenso, H. Cruz, and P. Dias, "Packet-header based no-reference quality metrics for H.264/AVC video transmission," in *Int. Conf. on Telecommunications and Multimedia*, pp. 174–151 (2012).
46. T. Liu et al., "Real-time video quality monitoring," *EURASIP J. Adv. Signal Process.* **2011**, 1 (2011).
47. ITU-T Telecommunication standardization sector of ITU, "ITU-T Recommendation G.1070: Opinion model for videotelephony applications," 2012, http://www.itu.int/rec/T-REC-G.1070 (04 September 2015).
48. N. Staelens et al., "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Trans. Circuits Syst. Video Technol.* **23**, 1322–1333 (2013).
49. K. Pandremmenou et al., "A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses," *Proc. SPIE* **9394**, 93941F (2015).
50. ITU-T Telecommunication standardization sector of ITU, "ITU-T Rec. J.247: objective perceptual multimedia video quality measurement in the presence of a full reference," 2008, https://www.itu.int/rec/T-REC-J.247/en (05 May 2016).
51. Institute for Telecommunication Sciences (ITS), "ITS video quality research software," 2015, http://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx (05 September 2015).
52. F. De Simone et al., "Subjective quality assessment of H.264/AVC video streaming with packet losses," *EURASIP J. Image Video Process.* **2011**, 190431 (2011).
53. A. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in *Packet Video*, pp. 308–317, IEEE (2007).
54. ITU-T Telecommunication standardization sector of ITU, "ITU-T Rec. J.342: objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal," 2015, http://www.itu.int/rec/T-REC-J.342-201104-I/en (05 September 2015).
55. J. You, J. Korhonen, and A. Perkis, "Spatial and temporal pooling of image quality metrics for perceptual video quality assessment on packet loss streams," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 1002–1005 (2010).
56. ITU-T Telecommunication standardization sector of ITU, "ITU-T Recommendation ITU-R P.910: subjective video quality assessment methods for multimedia applications," 1999, https://www.itu.int/rec/T-REC-P.910-200804-I/en (05 September 2015).
57. Karsten Sühring, "H.264/AVC software coordination," 2015, http://iphome.hhi.de/suehring/tml (05 September 2015).
58. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, Burlington, Massachusetts (2011).
59. C. Wang, T.-L. Lin, and P. Cosman, "Network-based model for video packet importance considering both compression artifacts and packet losses," in *IEEE Global Telecommunications Conf.*, pp. 1–5 (2010).
60. C. Wang, T.-L. Lin, and P. Cosman, "Packet dropping for H.264 videos considering both coding and packet-loss artifacts," in *18th Int. Packet Video Workshop*, pp. 165–172 (2010).
61. ITU-T Telecommunication standardization sector of ITU, "ITU-T Recommendation P.1401: methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2012, https://www.itu.int/rec/T-REC-P.1401-201207-I (05 September 2015).
62. Video Quality Experts Group, "VQEG report on validation of video quality models for high definition video content," 2010, http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx (05 September 2015).
63. Moscow State University (MSU), "MSU video quality measurement tool," 2016, http://compression.ru/video/quality_measure/vqmt_pro_en.html (17 May 2016).
64. D. Mocanu et al., "When does lower bitrate give higher quality in modern video services?" in *IEEE Network Operations and Management Symp.*, pp. 1–5 (2014).

Muhammad Shahid received his PhD in applied signal processing and MSc in electrical engineering from Blekinge Institute of Technology, Sweden, in 2014 and 2010 respectively. He is currently serving as assistant professor in the Department of Communications and Networks Engineering, Prince Sultan University, Riyadh, Saudi Arabia. His research interests include video processing, objective and subjective methods of video quality assessment, and video coding. He has been active in VQEG for International Research Collaborations.

Katerina Pandremmenou received the BSc degree in computer science in 2008 from the Computer Science Department, University of Crete, Heraklion, Greece. In 2011, she received the MSc degree in technologies-applications and in 2015 the PhD degree in video processing and communications from the Computer Science and Engineering Department, University of Ioannina, Ioannina, Greece.

Her current research interests include biomedical signal processing, video quality assessment, video quality metrics and resource allocation over wireless networks.

Lisimachos P. Kondi received the PhD degree in electrical and computer engineering from Northwestern University, Evanston, USA, in 1999. He is currently an associate professor in the Department of Computer Science and Engineering, University of Ioannina, Greece. His research interests are in the general areas of signal and image processing and communications, including image and video compression and transmission over wireless channels and the internet, sparse representations and compressive sensing, and super-resolution of video sequences.

Andreas Rossholm received his PhD in applied signal processing in 2014 from Blekinge Institute of Technology. He has a position at Microsoft's Skype Division, Stockholm, Sweden. His research interests in audio and video signal processing include pre- and post-processing in mobile devices with efficient processing. He is now focused in the area of audio and video perceptual quality, where he endeavors significant questions for industry especially related to real-time communication of audio and video.

Benny Lövström received his MSc EE degree in 1983 and his PhD degree in signal processing in 1992, both from Lund University. In 1993 he was appointed senior lecturer at Blekinge Institute of Technology, where he was also head of the department 1994-1999 and dean of education 2002-2010. His current interest areas are signal processing, image processing, and digital communications. His research is focused on image and video processing and perceptual quality of image and video.