




Error sensitivity model based on spatial and temporal features

Ran Ma^{1,2}  • Tong Li¹ • Dezhi Bo¹ • Qiang Wu³ • Ping An^{1,2}

Received: 29 September 2019 / Revised: 11 July 2020 / Accepted: 21 July 2020 /

Published online: 25 August 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Packet loss and error propagation induced by it are significant causes of visual impairments in video applications. Most of the existing video quality assessment models are developed at frame or sequence level, which can not accurately describe the impact of packet loss on the local regions in one frame. In this paper, we propose an error sensitivity model to evaluate the impact of a single packet loss. We also make full use of the spatio-temporal correlation of the video and analyze a set of features that directly impact the perceptual quality of videos, based on the specific situation of video packet loss. With the aid of the support vector regression (SVR), these features are used to predict the error sensitivity of the local region. The proposed model is tested on six video sequences. Experimental results show that the proposed model predicts sensitivity of videos to different packet loss cases with certain reasonable accuracy, and provides good generalization ability, which turns out outperform the state-of-art image and video quality assessment methods.

Keywords Packet loss · Spatial and temporal features · Error sensitivity · Regression

1 Introduction

With the development of video applications, video has become more and more important in daily life. The demand for high quality video is still increasing, which means more capacity and bandwidth are needed. Due to the unstable bandwidth and complex transmission

✉ Ran Ma
maran@shu.edu.cn

¹ School of Communication and Information Engineering, Shanghai University, 99 Shangda Road, Baoshan District, Shanghai 200444, China

² Shanghai Institute for Advanced Communication and Data Science, Shanghai University, 99 Shangda Road, Baoshan District, Shanghai 200444, China

³ Global Big Data Technologies Centre, University of Technology Sydney, NSW, Sydney 2007, Australia

environment, packet loss often occurs in streaming video, resulting in the degradation of perceived video quality. According to the characteristics of codec system, a frame subject to packet loss may cause some impairments in the successive frames. Even in the same frame, packet loss appearing in different regions (e.g., the regions with intense or slow movement), causes the video quality degradation at different degrees [4, 33]. Therefore, how to accurately measure the influence of packet loss faces challenges.

Many research works have been devoted to measure the impact of packet loss on video quality [16]. The most reliable way is to collect the judgements from many viewers, since humans are the final receivers of videos, such as the works [11, 17]. At the same time such subjective methods consume large time and human resource, which makes objective quality evaluations popular. Considering the diversity of lost packets, numerous studies have been explored quality change of video under some specific conditions, which mainly refer to different packet loss rates [6], different distributions of lost packets [5, 25], packet loss with different frame types (e.g., I, P, and B frame) [12, 27], packets lost in different Group of Pictures (GOP) patterns [31] and videos with different resolutions [3]. These works usually analyze the relationship between packet loss and traditional objective metrics, e.g., peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). It is noteworthy that traditional objective metrics are simple and easy to calculate, usually need the complete original video. Unfortunately, the original video is not always available. Especially when packet loss occurs, much of the information in the video is lost. There is the need for the development of no-reference (NR) methods, which estimate the influence with limited available input information.

During the past years, several NR methods have been proposed. In general, it is consisted of two main steps: analyzing the characteristics of corrupted video stream and predicting video quality. The prediction models are usually established by formulation and learning-based methods. The former incorporates the analyzed characteristics into explicit formulations, such as works [7, 21, 29, 30] and the latter uses machine learning techniques for deriving the quality from a set of features. Since the selected characteristics determine the accuracy of the prediction model, how to choose effective characteristics is a key point for NR methods. There are several features extracted from bit stream level [8, 28, 34, 35], such as the bit rate, the frame type, packet length, and the packet loss rate. These bit-stream based features can be easily obtained after a decoding of the packet and the video frame header. But sometimes using these parameters cannot exactly capture the impact of packet loss on a specific video, since for different video contents, the same parameters may have different impacts on video quality [1]. In that case, pixel level features are analyzed by utilizing the decoded video information. Because of the multidimensional nature of video, pixel-based techniques usually take into consideration both spatial and temporal features. For example, 3D shearlet transform and 3D discrete cosine transform is used in [19, 20], respectively. And the statistical analysis of transform coefficients can characterize the spatiotemporal statistics of videos in different views. Spatial features come from the statistical analysis of spatial activity and edge discontinuity indices, and temporal features are derived from the motion vector (MV) information [18]. A quality model based on time-domain characteristics was proposed to statistically analyze the residual diagram of adjacent frames of video to predict video quality [22]. Literature [2] extracted features from spatial domain and frequency domain: Canny operator was used to extract edge information in spatial domain; In the frequency domain, the Discrete Cosine Transform (DCT) was applied to the video frames to obtain the frequency domain features.

Up to now, most of the existing NR methods focus on the impact of packet loss on the whole frames or the video sequence, and rarely study the impact on the local regions in one

frame. Unfortunately, not all packet loss artifacts are the same, that is, the sensitivities of video to packet loss artifacts vary widely. Valenzise et al. [32] shows that after using error concealment (EC), some packet loss artifacts can be significantly alleviated, whereas the other may still be visible. More or less effective performance of EC depends on several factors, e.g., motion complexity and local texturing of the lost region. In addition, because of the temporal-prediction characteristic of codec system, errors in one frame may be spread to the following frames, which is called error propagation. Thus, packets lost in different regions with different coding patterns may cause various results. So, exploring the impact of packet loss on local regions can be of great importance. Korhonen [17] focuses on the visibility of packet loss artifacts appearing in spatially and temporally limited regions of a video sequence. The corrupted macroblocks (MBs) are combined into error clusters by using a methodology. A subjective test is then implemented to obtain the visibility of error clusters. This work narrows the region into partial spatiotemporal space, but the results significantly depend on the division of error clusters. A recursive distortion model [7] is proposed by analyzing the propagating behavior of transmission errors due to packet loss. Although the model has a good accuracy at the MB level, it entails recursive operation for every pixel in a MB, which increases the computational complexity.

In this paper, an error sensitivity model is proposed to measure the video quality affected by various packet losses. Different from the traditional methods that evaluate the quality on basis of the whole frames or the video sequence, the proposed model focuses on measuring the impact of a single lost packet on the local region. Once a block is lost, its internal information is also completely lost, which means the impact of the packet loss on the region is unknown. We firstly describe the error sensitivity index, based on the number of error pixels that still exist after using EC algorithm. Then, inspired by the spatiotemporal relativity of video sequences, we extract available features from the correctly received blocks in spatial and temporal domain. Finally, machine learning technology is applied to learn a mapping from feature space to error sensitivity of videos. Our model can give some guidance on feature selection in related fields, and at the same time, it can provide directions for error concealment algorithm improvement. Moreover, most of the existing methods are based on whole video frames, which cannot provide theoretical support for local region improvement, and our approach complements this nicely. Due to the complex factors that will affect the video quality (compression, blur, packet loss, etc.), unless otherwise noted, the error mentioned in this paper refers to packet loss. In H.265/HEVC, numerous largest coding units (LCUs) can be packetized into one packet for transmission. For simplifying the problem, we assume that each LCU is considered as a separate packet, and the packet loss in this paper refers to loss in the form of blocks. The major contributions of this paper are summarized as follows.

- 1) Considering the specific situation of video packet loss, we extract a collection of calculated features from the spatial and temporal domain. Most of these features are simple but closely quality related.
- 2) The proposed error sensitivity model pay attention to the levels of severity of damage in the local regions of the video, and can accurately predict the sensitivity of videos to different packet loss cases. The proposed model is appropriate to applications such as video processing and transmission, alleviating the impact of packet loss on video quality.

The rest of this paper is organized as follows. The theory of error sensitivity is introduced in Section 2. In Section 3, we describe details of the extraction process of the spatial and temporal

features. The error sensitivity prediction process from the extracted features is also presented in this section. Experiment results are shown in Section 4. Finally, conclusions are drawn in Section 5.

2 The theory of error sensitivity

As we all know, the quality impact of packet loss on local regions can be considered as a combination of quality degradation directly due to packet loss and that induced by error propagation. When a packet loss is detected, the decoder usually uses some EC algorithms to mitigate the degradation. However, there exists local differences in video sequences, leading to different performances of the EC algorithms. It is difficult to accurately measure the effect of packet loss on local region quality. Not all the lost information can be reconstructed intact, and errors in some areas are still very obvious, affecting the video viewing quality. Due to the diversification of the content characteristics of the video area (such as diversification of motion or diversity of texture), packet loss at different positions within the same video frame has a great difference in the impact on video quality [15]. What is the impact of different packet loss on video quality is an important issue to be solved urgently in the research field of video compression and processing. In this paper, error sensitivity, reflecting the sensitivity of the damaged region to errors, is used to evaluate the impact of packet loss. The regions with high sensitivity are more susceptible to errors, and the distortion in these regions usually remains quite noticeable even after simple EC operations. Thus, the error sensitivity can be obtained by counting the number of error pixels that still remain after concealing [10]. Since packet loss is loss in the form of blocks, the error sensitivity of a corrupted block y_B can be described as:

$$y_B = n/N_B \quad (1)$$

where n is the number of error pixels within the block, and N_B is the total number of pixels in the block. The more error pixels, the higher error sensitivity of the block.

Figure 1 illustrates the highly sensitive regions of the Traffic and Cactus sequences. Figure 1(a) and (b) are the original frames of Traffic and Cactus respectively. After subject to random packet loss at the rate of 20%, the damaged frames are shown in Fig. 1(c) and (d), where the black blocks refer to the lost regions. It is assumed that when the value of error sensitivity is higher than 10%, the damaged region is considered to be a highly sensitive region. In Fig. 1(e) and (f), the blocks left in lost regions are highly sensitive regions, and other white regions are low sensitive regions. It is worth noting that the regions containing objects or their boundaries are usually more sensitive to errors, whereas background or the regions where video content is consistent are not.

3 The proposed error sensitivity model

Since the original undistorted videos are not always accessible in many practical applications, it is not straightforward to obtain the sensitivities of regions to errors. Our proposed error sensitivity model predicts the sensitivity based on the quality-related features. The flowchart is given in Fig. 2. Numerous lost blocks in video sequences constitute the sample set. For every lost block, the related spatial and temporal features are extracted. After concealing the lost

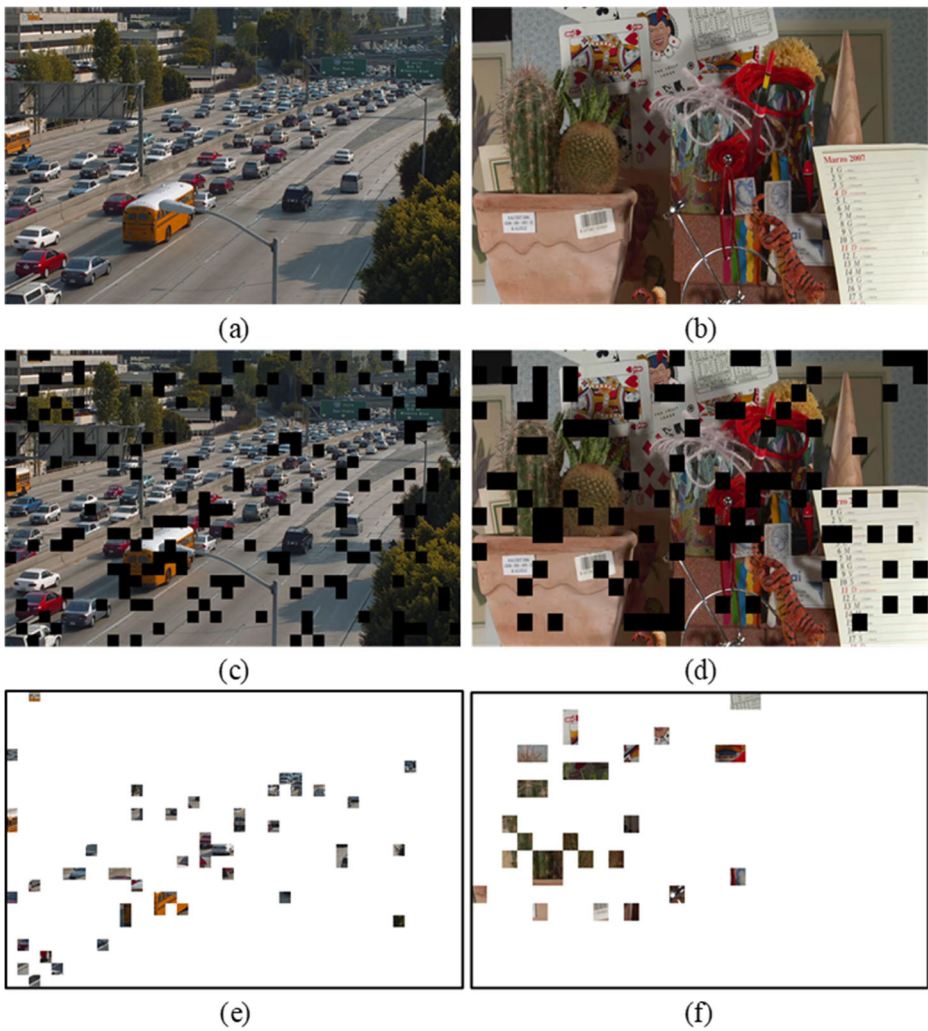


Fig. 1 The highly sensitive regions of the Traffic and Cactus sequences. (a) the 12th frame of Traffic; (b) the 11th frame of Cactus; (c), (d) the random packet loss results of (a), (b) with packet loss ratio: 20%; (e), (f) the highly sensitive regions in the lost regions of (c), (d)

block in training set, error sensitivity of the block is calculated, as described in Section 2. Then, the features and error sensitivity are used to train a regression module. Lastly, the trained model is used to map the features in testing dataset to error sensitivities.

3.1 Selection of spatial features

When a block is lost, it means that all the information about the block is also lost. Fortunately, natural videos possess substantial spatiotemporal regularities, in the sense that video frames at different times and spatial positions are highly correlated. According to this property, we extract the features from adjacent blocks in spatiotemporal domain. Selection of temporal features will be detailed in the next subsection. In this section, some spatial features are mainly discussed.

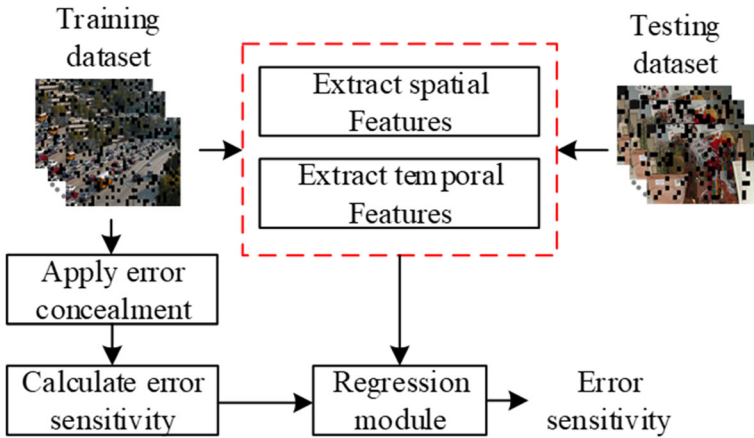


Fig. 2 The flowchart of error sensitivity model

3.1.1 Correctness of adjacent blocks in spatial domain

As a source of information for feature extraction, the correctness of adjacent blocks matters a lot. The number and locations of adjacent blocks may affect the video quality in different extent. Therefore, the eight surrounding blocks of current lost block, whose size is the same as that of current lost block, are considered. The relation of lost block and its adjacent blocks is demonstrated in Fig. 3, where B_0 is the lost block and B_k ($k = 1-8$) is the surrounding block of B_0 in spatial domain. It is assumed that the loss probability of each block is independent, and we come up with (2) to judge the blocks' correctness. Since there are eight adjacent blocks, each lost block corresponds to an eight-dimensional feature to characterize the correctness of adjacent blocks in spatial domain.

$$C_k = \begin{cases} 1, & B_k \text{ is correct} \\ 0, & B_k \text{ is lost} \end{cases} \quad (k = 1-8) \tag{2}$$

3.1.2 Textural features

Texture is one of the important characteristics for picture, describing the surface properties of the object in the region. When the region with detailed texture is corrupted, the performance of a normal EC method may not be satisfactory, since the surrounding texture information used is

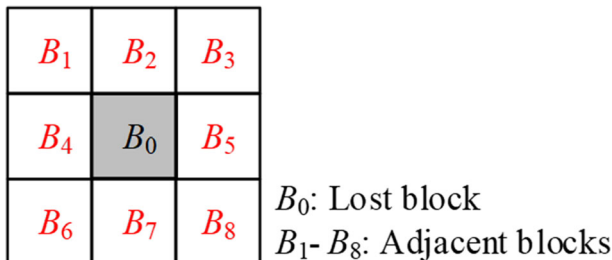


Fig. 3 The relation of lost block and its adjacent blocks

not reliable. On the other hand, for the region with simple texture, it is much easier to reconstruct region with good quality by using surrounding information. Hence textural features directly affect the sensitivity of regions to errors. In this work, textural features are calculated with the help of gray-level co-occurrence matrix (GLCM). GLCM is the representation of statistical joint probability of two pixels (I, j) held at distance d in direction θ , which reflects the characteristics of texture [9]. Various GLCMs can be calculated for different distances and directions, in order to reduce computational complexity, for each adjacent block received correctly, we only compute the GLCM matrices with a pixel distance of 1 in the directions 0° and 90° . These two matrices are then averaged and normalized to mitigate the effect of the direction on the results, expressed as $P_d(i, j)$. Considering the correlation of textural descriptors derived from GLCM, we focus on one of the descriptors, namely entropy, which measures the regularity versus disorder of pixel values in the block. The calculating formula of entropy H_k as follows:

$$H_k = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P_d(i, j) \log P_d(i, j) \tag{3}$$

where L is the gray level, and in this work L is set to 128. When all elements in co-occurrence matrix are equal or have the maximum randomness, the entropy becomes larger, which means the block has higher texture complexity. After computing the entropy features of all the available surrounding blocks, the texture complexity of the lost block E_c is defined as:

$$E_c = \begin{cases} \frac{\sum_{k=1}^8 H_k C_k}{\sum_{k=1}^8 C_k}, & \text{if } \sum_{k=1}^8 C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In particular, when all the adjacent blocks are lost, the texture complexity of the lost block is fixed to 0.

E_c reflects the average level of the texture complexity of the regions around the lost block. However, when the texture complexity of the surrounding regions varies widely, using the mean values alone cannot reflect the textural features effectively. Therefore, we also analyze the differences in the texture of available surrounding blocks, called texture consistency:

$$E'_c = \begin{cases} std(H_k, k = 1-8 \text{ and } C_k \neq 0), & \text{if } \sum_{k=1}^8 C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $std(\cdot)$ indicates the standard deviation operation.

3.1.3 Spatial activity

Spatial activity indicates the amount of detail of a video in spatial domain. It is derived from the gradient, describing the structural information of the video. As described in Fig. 1, regions with rich structural information, like edge regions, are more likely to be highly sensitive regions. Taking spatial activity into account can enhance the prediction accuracy of our model. In this paper, we use the definition of spatial activity [14] and modify it slightly. After Sobel operator, the spatial motion of the video frame calculates its standard deviation for all pixel values of the filtered image:

$$SA = std_{M,N}(Sobel(F)) \tag{6}$$

where F is the video frame, and $Sobel(\cdot)$ means Sobel filter operation on F . $std_{M,N}(\cdot)$ means standard deviation for all pixel values of a 5×5 size image, respectively.

After converting the video frame to grayscale image, the spatial activity of the available adjacent blocks SA_k can be calculated as:

$$SA_k = \frac{std_{H,W}(Sobel_x(B_k)) + std_{H,W}(Sobel_y(B_k))}{2} \tag{7}$$

where H is the height of the block, and W is the width of the block. $Sobel_x(B_k)$ and $Sobel_y(B_k)$ means applying the Sobel operator to B_k with the horizontal and vertical mask, respectively. Considering the computational complexity, 3×3 masks of Sobel operator are adopted. After obtaining the spatial activity values of all the available adjacent blocks, the spatial activity of the lost block E_{SA} can be calculated by:

$$E_{SA} = \begin{cases} \frac{\sum_{k=1}^8 SA_k C_k}{\sum_{k=1}^8 C_k}, & \text{if } \sum_{k=1}^8 C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

3.2 Selection of temporal features

It is not enough to only consider spatial features in our model. The degradation of video quality is not only because of impairments caused by packet loss, but also owing to error propagation along the direction of inter prediction. Packet loss artifacts may spread from frame to frame. Since temporal information plays an important role in error sensitivity estimation, in this paper some temporal features are analyzed.

3.2.1 Correctness of the corresponding block in temporal domain

According to the characteristics of inter prediction, if the reference blocks in previous frames are damaged, the probability of errors occurring in the current block increases, which means current block has higher sensitivity. Analyzing correctness of reference blocks is necessary. To simplify the problem, suppose the current frame only refers to its previous frame. The corresponding block in previous frame, which is in the same position as the lost block in the current frame, is taken into consideration, as shown in Fig. 4. Correctness of the corresponding block B_0 can be obtained by (2).

3.2.2 Motion features

Motion features reflect the motion activity of video content. Generally, temporal ECs work well if the video content is of low motion activity, but often cause noticeable errors in regions

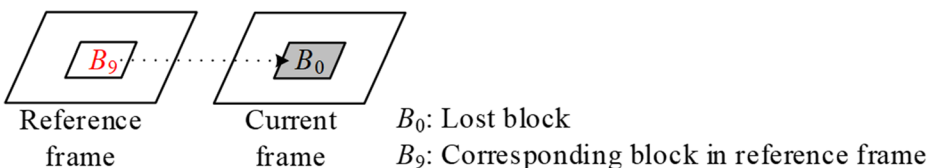


Fig. 4 The relation of lost block and its corresponding block in temporal domain

with high motion activity. In order to further explore the relationship between the motion characteristics of the region and the impact of packet loss in the region on the video, Fig. 5 shows the scatter plots of motion vector (MV) amplitude and error sensitivity of image blocks in BQMall sequence and BasketballPass sequence. It can be seen from the scatter plots of the two sequences that as the MV amplitude increases, the error sensitivity corresponding to the image block also shows an upward trend, that is, the two have a positive correlation. Therefore, it is necessary to analyze the motion characteristics of the video.

To capture the motion features of the lost block, we make full use of the spatio-temporal correlation of the video, and analyze the motion characteristics of related regions in spatial and temporal domain. For a lost block, the motion information of 4 surrounding blocks in spatial domain ($B_2, B_4, B_5,$ and B_7 in Fig. 3.) and the corresponding block in temporal domain (B_9 in Fig. 4.) is considered. Firstly, the motion vectors of the related blocks V_k can be calculated as:

$$V_k = \begin{cases} \sqrt{|MV_x(k)|^2 + |MV_y(k)|^2}, & C_k = 1 \\ 0, & C_k = 0 \end{cases} \quad (k = 2, 4, 5, 7, 9) \tag{9}$$

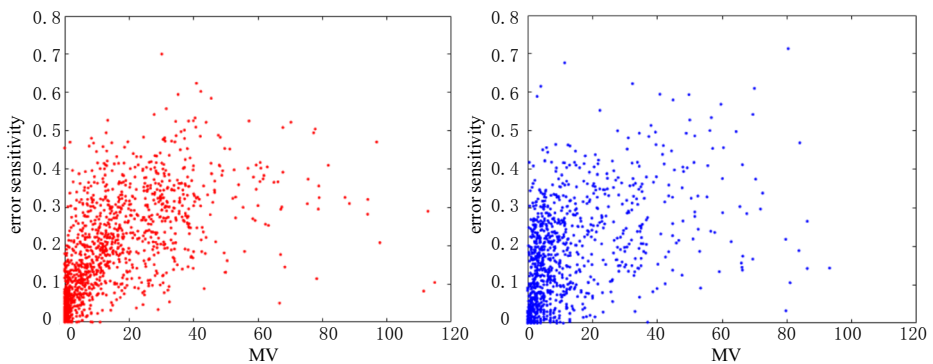
where $MV_x(k)$ and $MV_y(k)$ are the x -axis and y -axis components of the average motion vector of all pixels in the k th block, respectively. Then, the motion intensity of the lost block E_V can be estimated by those motion vectors:

$$E_V = \begin{cases} \frac{\sum V_k C_k}{\sum C_k}, & \text{if } \sum C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (k = 2, 4, 5, 7, 9) \tag{10}$$

In addition, by analyzing the motion differences of the related regions, we can obtain the motion consistency of the lost block:

$$E_V' = \begin{cases} std(V_k, k = 2, 4, 5, 7, 9 \text{ and } C_k \neq 0), & \text{if } \sum C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Finally, the motion features of the lost block are characterized by the motion intensity and the motion consistency.



(a) Scatter plot of BQMall sequence (b) Scatter plot of BasketballPass sequence

Fig. 5 Scatter plot of MV amplitude and error sensitivity of image blocks

3.2.3 Temporal randomness

Temporal randomness measures the temporal regularity of video content. Packet loss occurring in regions with regular and unregular motion can result in different visual impairments. According to [13], we can obtain temporal randomness by using previous frames to predict the current frame, which can be written as:

$$r = |f(t_2 + 1) - C\hat{A}X_{t_1}^{t_2}| \tag{12}$$

where $f(t_2 + 1)$ represents the current frame. C , \hat{A} , and $X_{t_1}^{t_2}$ are the related parameters of $F_{t_1}^{t_2}$ (a sequence from the t_1 th frame to the t_2 th frame), carrying the information of previous frames. They come from the theory that there are some connections between frames, and the video signal can be modeled as a dynamic system:

$$F_{t_1}^{t_2} = CX_{t_1}^{t_2} + W_{t_1}^{t_2} \tag{13}$$

$$X_{t_1}^{t_2} = AX_{t_1-1}^{t_2-1} + V_{t_1}^{t_2} \tag{14}$$

where $X_{t_1}^{t_2}$ and $X_{t_1-1}^{t_2-1}$ are the state sequence of $F_{t_1}^{t_2}$ and $F_{t_1-1}^{t_2-1}$, respectively. A is the state transition matrix reflecting the regularity of motion, and C is a metric to encode the regularity of spatial information. $W_{t_1}^{t_2}$ and $V_{t_1}^{t_2}$ are the noise that cannot be represented by C and A , respectively. By conducting the singular value decomposition on $F_{t_1}^{t_2}$, we can obtain the parameter C and $X_{t_1}^{t_2}$ in (13). Since A reflects the motion information and can be used to predict next frames, the optimal A is expected to represent information as much as possible, which can be calculated as:

$$\hat{A} = X_{t_1+1}^{t_2} \left(X_{t_1}^{t_2-1} \right)^{-1} \tag{15}$$

where $\left(X_{t_1}^{t_2-1} \right)^{-1}$ is the pseudo inverse of $X_{t_1}^{t_2-1}$. Once the related parameters of $F_{t_1}^{t_2}$ are obtained, we can calculate r by using (12). It is not hard to see that r reflects the unregular information that cannot be predicted from previous frames, which indicates temporal randomness.

In this paper, the temporal randomness is obtained by using the previous frame to predict the current frame. If the motion structures between adjacent frames are similar to each other, the temporal randomness should be small. To further visualize temporal randomness, by following [13], we transform the values of temporal randomness between 0 and 255, generating the temporal randomness map. The brighter the point in the map, the stronger the temporal randomness. Figure 6 shows the temporal randomness map for the Traffic sequence. Figure 6(a) and (b) are two consecutive frames in the sequence, and Fig. 6(c) shows the corresponding temporal randomness map. As seen in Fig. 6, the motion in background is regular, and its temporal randomness is rather small. But for the cars, the motion is unpredictable, corresponding to large temporal randomness.

Since the temporal randomness describes the regularity of video content between frames, the information from the temporal domain is as important as the information from the spatial domain when estimating the temporal randomness of the lost block. In fact, the temporal randomness represents the intensity of changes corresponding to the local region, and temporal

randomness is large when the movement is intense, otherwise, it is small. Therefore, the temporal randomness can well reflect the error sensitivity, which is also the reason for using this feature in this paper. For every lost block, we analyze the temporal randomness of its eight surrounding blocks and the corresponding block in the previous frame. After calculating the temporal randomness of available related blocks, we get the sum of temporal randomness of each block r_k , and then define the average value as the temporal randomness of the lost block E_r . The detail process is expressed as follows.

$$r_k = \sum_{i=1}^H \sum_{j=1}^W |r(i, j)| \quad (16)$$

$$E_r = \begin{cases} \frac{\sum_{k=1}^9 r_k}{\sum_{k=1}^9 C_k}, & \text{if } \sum_{k=1}^9 C_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where $r(i, j)$ is the temporal randomness of the pixel at position (i, j) .

3.3 Module regression

As the original video signals are not always available, we cannot directly calculate error sensitivity according to (1), which means how much it is sensitive to errors is unknown. Fortunately, machine learning methods have been widely used to derive the index from numerous features. They usually divide the sample dataset into a training dataset and a testing

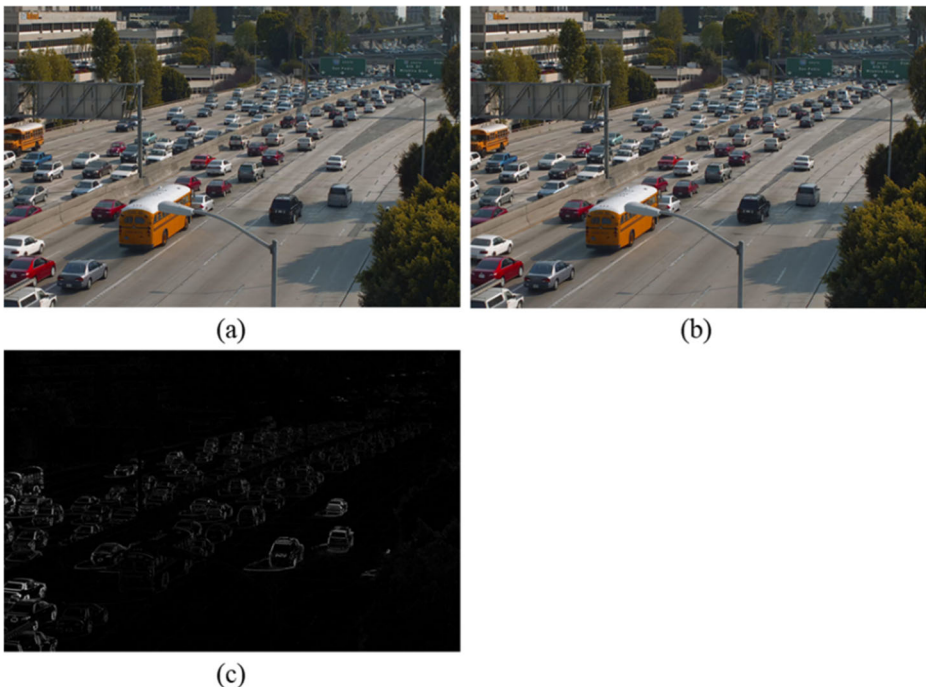


Fig. 6 The temporal randomness map for the Traffic sequence. (a)–(b) Consecutive frames in Traffic; (c) The corresponding temporal randomness map

dataset, as depicted in Fig. 2. For the training dataset, the above extracted features and their corresponding ground truth are used to train the model. So far, a total of 15 features are extracted for each lost block, including 11 spatial features and 4 temporal features, which are C_k ($k = 1 - 9$), E_c , E_c' , E_{SA} , E_V , E_V' , and E_r , respectively. We listed these features in Table 1. To get the ground truth of error sensitivity of the lost block in training dataset, we assume that the lost block is concealed by the simplest temporal EC method, where the lost block is directly replaced with the corresponding block in previous frame. Error sensitivity of the region is then obtained by using (1). In this paper, SVR is adopted to learn the relationship between the features and error sensitivity index. Specially, the LibSVM package is utilized to implement the SVR with the Radial Basis Function (RBF) kernel. The task of SVR is to train a regression model such as (18) so that $f(\mathbf{x})$ and y are as close as possible.

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (18)$$

where \mathbf{w}^T and b are the parameters of the model. Here \mathbf{x} is the feature set of the missing block, $f(\mathbf{x})$ is the predicted error sensitivity.

For the testing dataset, the extracted features are fed into the trained SVR model, and its corresponding error sensitivity is then predicted.

4 Experiment results

In this section, some experiments have been carried out to evaluate the performance of the proposed error sensitivity model. The experiment settings are introduced firstly, then the details of experimental results are reported.

4.1 Experiment settings

In our work, all the experiments are conducted in HM-16.9 and MATLAB R2016a. We use six video sequences (Traffic, Cactus, Kimono, BQMall, BasketballPass and FourPeople) to evaluate our model. These sequences have different spatial and temporal complexity. Detailed information of these sequences is summarized in Table 2. All the video sequences are firstly compressed using the H.265/HEVC encoding standard (HM-16.9), encoded in IPPP... sequence format with a fixed QP (32). 20% packet loss rate is considered to study the effect at worst channel conditions. The packet loss is simulated block wise, and each loss block size is 64×64 pixels.

Table 1 Features and feature description

Features	Feature description
$C_1 \sim C_8$	Correctness of adjacent blocks in spatial domain
C_9	Correctness of the corresponding block in temporal domain
E_c	Texture complexity
E_c'	Texture consistency
E_{SA}	Spatial activity
E_V	Motion intensity
E_V'	Motion consistency
E_r	Temporal randomness

For each database, 80% of samples are randomly selected as training set and the remaining 20% samples are used as test set. To avoid any performance bias, we repeat the training and test cycle using 10 different random splits, and the mean values are reported as the final performance score.

There are three criteria employed in this study to quantitatively measure the performance of the model: Pearson's linear correlation coefficient (PLCC), Spearman's rank-order correlation coefficient (SROCC), and root mean-squared error (RMSE). PLCC and RMSE are utilized for measuring prediction accuracy, whereas SROCC is used for measuring prediction monotonicity. Higher PLCC and SROCC values, and RMSE value closer to zero, indicate good correlation with the ground truth.

4.2 Performance comparison

In order to investigate the effectiveness of the proposed model, we conduct some experiments to compare the proposed model (donated as Proposed) with existing, state-of-art NR quality assessment models. These models are DIIVINE [23], NOREQI [24], VBLIINDS [26] and NR_VQA [18]. Among them, DIIVINE and NOREQI are good at evaluating the frame level quality, for convenience, called as type A models in the following. They use spatial or frequency features and do not consider any temporal features. In order to verify the effectiveness of the spatio-temporal feature selection in this paper, we also select two general video quality assessment model (VBLIINDS and NR_VQA), both of which analyze many characteristics of videos in spatial and temporal domain. These two models and the proposed model are collectively referred to as type B models. For the fair comparison, we extract the features from these models at block level, then train the SVR model with these features to predict error sensitivity. Moreover, the SVR parameters have been all optimized to achieve their best performance.

Table 3 lists the performances of all the methods on the six video sequences, where the best performance is highlighted in bold. From Table 3, the PLCCs and SROCCs of the proposed model are larger than other models in all the six video sequences, and the RMSEs are smaller than others, which means the proposed model can predict error sensitivity with higher accuracy. It is worth noting that, in most cases, the PLCCs and SROCCs of type B models considering video characteristics in spatio-temporal domain are larger than those of type A models only considering spatial or frequency features, and the RMSEs of the type A models are smaller than those of the type B models, which highlights the importance of the temporal information. Especially for the FourPeople, which has simple content and low motion activity, type B models is significantly superior to type A models, increasing the PLCCs and SROCCs at least by 0.06. This is due to the fact that these type B models extract temporal features more

Table 2 Information of video sequences

Sequence	Class	Resolution	Frame numbers	Frame rate
Traffic	A	2560 × 1600	60	30
Cactus	B	1920 × 1080	100	50
Kimono	B	1920 × 1080	100	24
BQMall	C	832 × 480	100	60
BasketballPass	D	416 × 240	500	50
FourPeople	E	1280 × 720	100	60

accurately when the motion of the video is slow. However, for the sequences with intensive motion, such as Kimono and BasketballPass, it is more difficult to predict error sensitivity for all the models. The PLCCs and SROCCs are all below 0.7, and the RMSEs are larger than 0.094. To visualize the statistical significance of the comparison, we take the Traffic as an example and show the box plots of PLCC, SROCC, and RMSE distributions of different models over 10 trails in Fig. 7(a), 6(b), and 6(c), respectively. It is clear that the proposed model works well among all the NR models under consideration.

As we know, generalization capability is a significant problem for all learning-based methods. To evaluate the generalization capability of our model, we implement cross-dataset experiments, where models are trained and tested on different datasets. Six video sequences are divided into two parts. Considering the balance of sample size, three video sequences (Traffic, BasketballPass and FourPeople) are used for training, and the remaining sequences (Cactus, Kimono and BQMall) for testing. The results are exhibited in Table 4. It can be observed that, compared with the results in Table 3, the cross-dataset experiments have reduced the prediction accuracy of both the proposed model and the other four models. Even so, our model maintains the stable performance across most sequences, showing better robustness. In addition, compared with the performances of type B models, the performances of type A models are comparatively low. Because the motions of different sequences vary widely, using spatial or frequency features alone cannot predict the error sensitivity well, which indicates the validity of the combination of spatial and temporal features.

4.3 Contributions of features

In the paper, 15 features are extracted to train the error sensitivity model. To further investigate their individual or combined contributions to the performance of the model, the following test is conduct. In our experiment, 15 features are classified into five categories: correctness of the related blocks (both in spatial and temporal domain), textural features, spatial activity, motion features, and temporal randomness, donated by I, II, III, IV, and V, respectively. The

Table 3 Performances of the proposed model and the other four models on the six video sequences

Sequence	Criterion	DIIVINE	NOREQI	VBLIINDS	NR_VQA	Proposed
Traffic	PLCC	0.6636	0.6456	0.6934	0.7240	0.7580
	SROCC	0.7211	0.7188	0.7436	0.7671	0.7812
	RMSE	0.0725	0.0740	0.0699	0.0669	0.0633
Cactus	PLCC	0.6561	0.6776	0.7370	0.7521	0.7679
	SROCC	0.7459	0.7593	0.8157	0.8250	0.8424
	RMSE	0.1044	0.1017	0.0935	0.0912	0.0886
Kimono	PLCC	0.6155	0.6006	0.6496	0.6258	0.6757
	SROCC	0.6288	0.6086	0.6650	0.6332	0.6857
	RMSE	0.1014	0.1029	0.0978	0.1003	0.0948
BQMall	PLCC	0.7186	0.6864	0.7430	0.7172	0.8044
	SROCC	0.7357	0.7130	0.7741	0.7535	0.8359
	RMSE	0.0990	0.1038	0.0955	0.0995	0.0848
BasketballPass	PLCC	0.5330	0.5669	0.6333	0.5916	0.6941
	SROCC	0.5491	0.5770	0.6408	0.6020	0.6952
	RMSE	0.1128	0.1098	0.1031	0.1075	0.0959
FourPeople	PLCC	0.5222	0.5554	0.6321	0.6186	0.7474
	SROCC	0.5032	0.4841	0.5843	0.5811	0.6320
	RMSE	0.0436	0.0425	0.0396	0.0401	0.0340

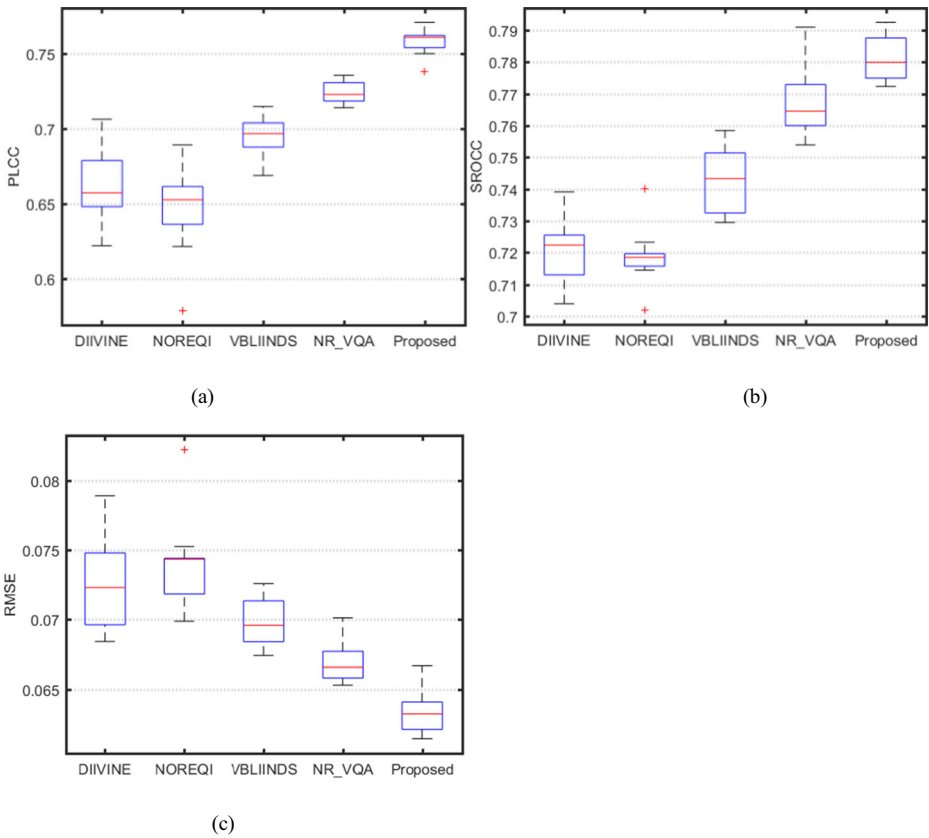


Fig. 7 Box plot of PLCC, SROCC, and RMSE distributions of the models over 10 trails on the Traffic. (a) Box plot of PLCC distribution; (b) Box plot of SROCC distribution; (c) Box plot of RMSE distribution

performances of different combinations of feature types on the BasketballPass is shown in Fig. 8. First, we test on each feature type in isolation and rank them in terms of PLCC (SROCC or RMSE can also be used): feature V (0.5393) > feature IV (0.5168) > feature I (0.3908) > feature II (0.3555) > feature III (0.2028). It can be observed from the rank that feature V is the most contributing feature. However, the prediction accuracy of the model using only one feature type is still unsatisfactory, reflected by the PLCCs below 0.6. Second, to find the most

Table 4 Results of cross-dataset experiments

Sequence	Criterion	DIIVINE	NOREQI	VBLIINDS	NR_VQA	Proposed
Cactus	PLCC	0.1784	0.0792	0.6107	0.7188	0.6941
	SROCC	0.2334	0.1216	0.7182	0.8194	0.8051
	RMSE	0.1372	0.1390	0.1105	0.0970	0.1004
Kimono	PLCC	0.3124	0.2261	0.4093	0.5057	0.5281
	SROCC	0.3425	0.1849	0.4043	0.5070	0.5361
	RMSE	0.1214	0.1244	0.1166	0.1103	0.1086
BQMall	PLCC	0.3354	0.1919	0.3841	0.6697	0.7847
	SROCC	0.3729	0.2195	0.4003	0.7168	0.8238
	RMSE	0.1325	0.1378	0.1299	0.1045	0.0872

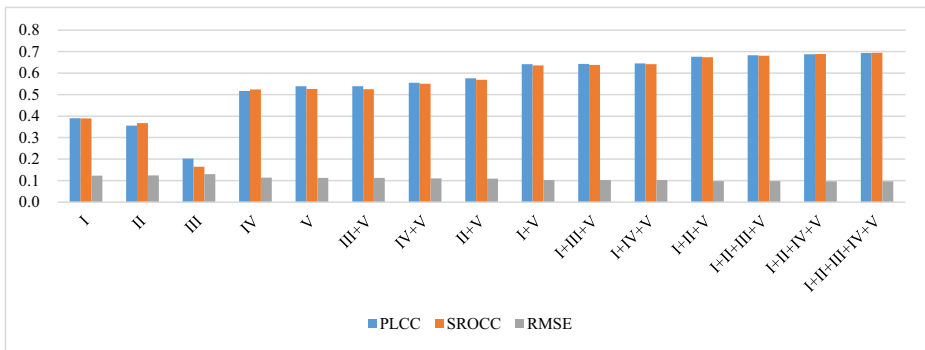


Fig. 8 Performances of different combinations of feature types on the BasketballPass

effective combination of feature types, we fix the best feature V and add each of the other four feature types individually. The new rankings are obtained: feature I + V (0.6414) > feature II + V (0.5762) > feature IV + V (0.5554) > feature III + V (0.5387). It is evident that the performance is much better if two feature types are used. Then, we fix the optimal feature I + V and add each of the other three feature types, and so on. Finally, the contributions of several combinations of feature types are evaluated.

As demonstrated in Fig. 8, all the five feature types play important roles in the performance of the error sensitivity model. With the addition of feature types, the values of PLCC or SROCC are on the rise, whereas the RMSE values reveal a trend of gradual decrease. When all the five feature types are utilized together, it achieves the best performance, indicating the feasibility of our model.

5 Conclusions

In this paper, a novel error sensitivity model, aiming to explore the impact of different packet losses on local regions, is presented. To solve the problem of missing information when packet loss appears, the available information from adjacent regions is further studied. Spatial and temporal features, which relate to error sensitivity, are considered comprehensively. We detail the features extracted, and then map it into error sensitivities using the SVR. The results of experiments conducted show that our model provides high accuracy for prediction of error sensitivity and is robust to different datasets as compared to state-of-the-art NR quality assessment models. Moreover, we demonstrate the effectiveness of feature selection of our model. More importantly, our error sensitivity model can give some guidance and improvement direction to the error concealment algorithm. Because improving the algorithm for high sensitivity region can greatly improve the error concealment effect.

However, the prediction accuracy for videos with high motion activity is still not encouraging. In the future, we will focus on the temporal characteristics and investigate other features to improve the prediction accuracy.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grant No. 61301112, 61828105 and 61601278, Chen Guang Project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant No.17CG41.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. A. Adeyemi-Ejeye, M. Alreshoodi, L. Al-Jobouri, M. Fleury, J. Woods and M. Medhi (2017) IEEE 802.11ac wireless delivery of 4kUHD video: The impact of packet loss. Proc ICCE-Berlin 258–259. Berlin, Germany
2. H. Bayrak, G. N. Yilmaz (2018) No-Reference Evaluation Of 3-Dimensional Video Quality Using Spatial and Frequency Domain Components. 26th Signal Processing and Communications Applications Conference (SIU) 1–4
3. Bondzolic BP, Pavlovic BZ, Petrovic VS, Andric MS (2016) Performance of peak signal-to-noise ratio quality assessment in video streaming with packet losses. Electron Lett 52(6):454–456
4. N Chen, X Jiang, C Wang and J Su (2011) Study on relationship between network video packet loss and video quality. Proc CISP 282–286. Shanghai, China
5. N. Chen, X. Jiang and C. Wang (2012) Impact of packet loss distribution on the perceived IPTV video quality. Proc CISP 38–42. Chongqing, China
6. Frnda J, Voznak M, Sevcik L (2016) Impact of packet loss and delay variation on the quality of real-time video streaming. Telecommun Syst 62(2):265–275
7. Gao P, Peng Q, Xiang W (2017) Analysis of packet-loss-induced distortion in view synthesis prediction-based 3D video coding. *IEEE Trans Image Process* 26(6):2781–2796
8. MN Garcia and A Raaqe (2010) Parametric packet-layer video quality model for IPTV. Proc Inform Sci Sign Proc Their Appl 349–352. Kuala Lumpur, Malaysia
9. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern -Syst Smc-3(6):610–621*
10. CTER Hewage, MG Martini, HD Appuhami (2012) A study on the impact of compression and packet losses on rendered 3D views. Proc SPIE 8290, Three-Dimensional Image Processing (3DIP) and Applications II, pp. 82901D–82901D–9, Burlingame, CA, USA
11. CTER Hewage, MG Martini, M Brandas and DVSX De Silva (2013) A study on the perceived quality of 3D video subject to packet losses. Proc IEEE ICC Workshops 662–666. Budapest, Hungary
12. Hu Z, Zhang Q (2018) A new approach for packet loss measurement of video streaming and its application. *Multimed Tools Appl* 77(10):11589–11608
13. Hu S, Jin L, Wang H, Zhang Y, Kwong S, Kuo C-J (2017) Objective video quality assessment based on perceptually weighted mean squared error. *IEEE Trans Circ Syst Video Technol* 27(9):1844–1855
14. ITU-T Rec. P.910 (1999) Subjective video quality assessment methods for multimedia applications, ITU, Geneva, Switzerland
15. Joskowicz J, Sotelo R (2014) A model for video quality assessment considering packet loss for broadcast digital television coded in H.264. *Int J Digit Multimed Broadcast* 2014:1–11
16. J Joskowicz, R Sotelo and JC Lopez Arado (2012) Comparison of parametric models for video quality estimation: Towards a general model. Proc IEEE Int Symp Broadband Multimed Syst Broadcast 1–7. Seoul, Korea
17. Korhonen J (2018) Study of the subjective visibility of packet loss artifacts in decoded video sequences. *IEEE Trans Broadcast* 64(2):354–366
18. J Korhonen (2018) Learning-based prediction of packet loss artifact visibility in networked video. Proc. 10th Int. Workshop QoMEX 1–6. Cagliari, Italy
19. Li Y, Po LM, Cheung CH, Xu X, Feng L, Yuan F, Cheung KW (2016) No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans Circ Syst Video Technol* 26(6):1044–1057
20. Li X, Guo Q, Lu X (2016) Spatiotemporal statistics for video quality assessment. *IEEE Trans Image Process* 25(7):3329–3342
21. K. Manasa, KVSNL Manasa Priya and SS Channappayya (2014) A perceptually motivated no-reference video quality assessment algorithm for packet loss artifacts. Proc 6th Int Workshop QoMEX 67–68. Singapore, Singapore
22. A. Mittal, M. Saad, A. C. Bovik (2014) Assessment of Video Naturalness Using Time-Frequency Statistics. *IEEE International Conference on Image Processing (ICIP)* 571–574

23. Moorthy AK, Bovik AC (2011) Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Trans Image Process* 20(12):3350–3364
24. Oszust M (2017) No-reference image quality assessment using image statistics and robust feature descriptors. *IEEE Sign Proc Lett* 24(11):1656–1660
25. R. Pauliks, I. Slaidins, K. Tretjaks and A. Krauze (2015) Assessment of IP packet loss influence on perceptual quality of streaming video. *Proc APMediaCast* 1–6. Kuta, Indonesia
26. Saad MA, Bovik AC, Charrier C (2014) Blind prediction of natural video quality. *IEEE Trans Image Process* 23(3):1352–1365
27. YM Saputra and Hendrawan (2016) The effect of packet loss and delay jitter on the video streaming performance using H.264/MPEG-4 Scalable Video Coding. *Proc TSSA* 1–6. Denpasar, Indonesia
28. Shahid M, Pandremmenou K, Kondi LP, Rossholm A, Lövfström B (2016) Perceptual quality estimation of H.264/AVC videos using reduced-reference and no-reference models. *J Electron Imaging* 25(5):053012–053012-26
29. Song J, Yang F (2014) No-reference video quality assessment model for distortion caused by packet loss in the real-time mobile video services. *Adv Multimed* 2014:1–15
30. Tang S, Alface PR (2014) Impact of random and burst packet losses on H.264 scalable video coding. *IEEE Trans Multimed* 16(8):2256–2269
31. M Uhrina, M Vaculík (2015) The impact of bitrate and packet loss on the video quality of H.264/AVC compression standard. *Proc. TSP* 1–6. Prague, Czech Republic
32. Valenzise G, Magni S, Tagliasacchi M, Tubaro S (2012) No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Trans Circ Syst* 22(4):605–618
33. S Wan, F Yang, Z Xie (2010) Evaluation of video quality degradation due to packet loss. *Proc ISPACS* 1–4. Chengdu, China
34. Wang Z, Wang W, Xia Y, Wan Z, Wang J, Li L, Cai C (2014) Visual quality assessment after network transmission incorporating NS2 and Evalvid. *Scientific World J* 2014:1–7
35. Yang F, Wan S, Xie Q, Wu HR (2010) No-reference quality assessment for networked video via primary analysis of bit stream. *IEEE Trans Circ Syst Video Technol* 20(11):1544–1554

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Ran Ma received her BS degree from Yangzhou University, Yangzhou, China, and her MS and PhD degrees from Shanghai University, Shanghai, in 1997, 2000, and 2018, respectively. She is currently an associate professor in the School of Communication and Information Engineering, Shanghai University. Her research interests include stereoscopic image and video processing, coding, and application.