Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

No reference stereo video quality assessment based on motion feature in tensor decomposition domain

Gangyi Jiang[a,b,*], Shanshan Liu[a], Mei Yu[a,b], Feng Shao[a], Zongju Peng[a], Fen Chen[a]

[a] *Faculty of Information Science and Engineering, Ningbo University, Ningbo, China*
[b] *National Key Lab of Software New Technology, Nanjing University, Nanjing, China*

ARTICLE INFO

Keywords:
No reference stereo video quality assessment
Tensor decomposition
Motion feature
Entropy
Random forest

ABSTRACT

A no reference stereo video quality assessment method based on motion features extracted in tensor decomposition domain is proposed. Tensor decomposition is used to reduce dimension of color, view and time of stereo video, and motion information maps containing time-varying information of inter-views and intra-views are obtained. Statistical features such as generalized Gaussian distribution (GGD), asymmetric GGD, spatial entropy, spectral entropy associated with two views, and spectral entropy related to depth perception of stereo video, are extracted. Random forest is utilized to establish relationship between stereo video quality and the extracted features. Experimental results on NAMA3DS1-COSPAD1 database demonstrate that the proposed method achieves good performance on JP2K, resolution reduction, sharpening and their combination distortions, Pearson linear correlation coefficient (PLCC) values of these types of distortions are higher than 0.97, while for H.264 distortion the PLCC value is 0.8850, which means that the proposed metric is consistent with human visual perception.

1. Introduction

Stereo video has board application prospect due to its excellent immersive experience [1]. However, stereo video technologies have the same problem as that for 2-dimensional (2D) videos, that is, the distortions which cannot be avoided during the process of the video [2,3]. In order to assess process performance during the processes of acquisition, processing, compression, storage and transmission of videos, and provide satisfied display for the user of 3-dimensional (3D) video system, objective video quality assessment (VQA) is necessary for the user experience [4,5].

According to the availability of original video as the reference, objective VQA methods can be roughly classified into three categories: full reference (FR), reduced reference (RR), and no reference (NR). FR methods need full original video as the reference, however in most cases this is impossible in practice. By contrast, RR methods only need representative features of the original video, while NR methods do not require any information of reference video. Therefore, from a practical perspective, NR methods are most expected for VQA.

Generally, stereo video has two views, and obviously the quality of each view is related to the overall quality of the stereo video. However, compared with 2D video, stereo video further provides depth information which enhances the viewing experience. Thus, depth information is also significant to the overall quality of stereo video.

Yu et al. proposed a stereo VQA method, which considered the influence of temporal characteristics of video and binocular perception in human visual system (HVS) [6]. Galkandage et al. presented a FR method based on an extended HVS model including the phenomena of binocular suppression and recurrent excitation, etc [7]. But these two methods need to know all or partial information of the original video, which results in limitation when applied to practical applications. Zhao et al. thought human eyes being more sensitive to moving object and edge information [8], and put forward a NR 3D video metric on the basis of visual attention and edge difference. Nevertheless, the accuracy of disparity has large influence on this objective metric. Han et al. considered the correlation between network packet loss and perceptual video quality for different bit-rate video sequences [9]. Moreover, they modeled the impact of network packet loss at different bit-rates and frame rates on the perceived quality of stereo video to make the video quality metric more generic [10]. But the metric proposed by them is only suitable for network delivery effects. For a specific network, parameters need to be computed. In addition, in the case of bit stream

[*] Corresponding author at: Faculty of Information Science and Engineering, Ningbo University, Ningbo, China.
E-mail address: jianggangyi@126.com (G. Jiang).

https://doi.org/10.1016/j.jvcir.2017.12.001
Received 14 April 2017; Received in revised form 1 November 2017; Accepted 9 December 2017
Available online 11 December 2017

unavailable due to the fact that it is encrypted or processed by the third part decoders, this kind of bit stream information based metric is invalid [11].

Natural scene statistics (NSS) models have been researched extensively and achieve good performance in image quality assessment (IQA). Based on statistical characteristics including generalized Gaussian distribution (GGD), asymmetric GGD (AGGD), entropy, etc., many NR methods have been proposed [12,13]. But the performance of this kind of metrics for VQA is not as well as for IQA, because some of these statistical features do not have the ability to distinguish the distortion degree of the video. In the past few years, many researchers have focused on employing NSS feature to measure video quality. Saad et al. studied the NSS features of frame difference in DCT domain [14]. Soundararajan et al. researched the frame difference of wavelet coefficients [15]. These researches confirmed that the frame difference of original videos has certain distribution regularity, and frame difference can represent the structure of the motion edge. Therefore, the frame difference can be used to obtain temporal information which is important for VQA.

Scalars and vectors are commonly used in traditional data processing. But the real-world data are usually multi-dimensional. For example, a gray image is 2-dimensional, a color image is 3-dimensional, color video is 4-dimensional, and color 3D video with two views composes of 5-dimensional data. Therefore, scalars and vectors can not reflect the complex structure of the real world data. By contrast, tensor decomposition is suitable for multi-dimensional data processing. Over the past few decades, tensor decomposition based methods have been widely adopted in medical imaging, surveillance, machine learning, etc [16]. The CANDECOMP/PARAFAC (PC) [17] and Tucker [18] families are the mainly used classes of tensor decomposition, and many other tensor decomposition methods are derived from them.

In this paper, we propose a NR stereo video quality metric called motion feature based no reference stereo video quality metric (MNSVQM). The color stereo video, which can be represented as 5-dimensional tensor, is processed by Tucker decomposition implemented through N-mode singular value decomposition (SVD) [18]. By analyzing the principal component of the N-dimensional data, the time-varying information of the video is obtained to construct the motion information map, and four kinds of features are then extracted from the motion information map in tensor domain based on statistical models such as spatial entropy and spectral entropy, GGD and AGGD model. Benefiting from their obvious statistical regularity, these features are used to distinguish distortion type and degree of the stereo video. Finally, random forest is adopted to model human visual perception based on these features so as to predict the quality of the stereo video.

The remaining parts are organized as follows. Section 2 describes the proposed NR stereo VQA metric in details. In Section 3, experimental results and discussions are presented. Conclusions are drawn in Section 4.

2. The proposed motion feature based no reference stereo video quality metric

In this paper, we use tensor decomposition to extract main motion information from the video, and propose a NR stereo video quality assessment method, the diagram of which is shown in Fig. 1. The proposed method consists of three parts: motion information map acquisition, feature extraction, random forest model training and video quality prediction.

It is clear that time-varying information is important to the quality of a video. In this paper, motion information map is acquired with tensor decomposition. The tensor decomposition can be implemented through N-mode SVD, which can realize the principal component analysis (PCA) of high-order data. PCA process can obtain main information of the signal and achieve dimensionality reduction, it can compress the data without loss of data as much as possible, and obtain the linearity of the original variable combination [19]. PCA is good at one-dimensional and two-dimensional data processing, but has some problems when used for high-order data, for example, the monocular gray-scale video which contains three dimensions such as space and time, that is, it is the 3rd-order tensor. PCA processing will ignore the temporal and spatial relations of the video. By contrast, tensor decomposition can take into account the temporal and spatial information of the video [20], which is quite important for VQA. In this paper, the purpose of feature extraction in tensor decomposition domain is to obtain the features related to stereo video quality on the basis of time-varying information as well as depth information. Then, these features are utilized for the subsequent random forest model training and video quality prediction.

Let the frame size of stereo video be $W \times H$, K be the color dimension, S denote the time dimension. For each view of the stereo video, makes every S frames to form a group so that a 4th-order tensor $\chi \in \mathbb{R}^{W \times H \times K \times S}$ can be obtained, and the tensor is called group of frame (GOF) tensor hereinafter. The first two dimensionalities (i.e. mode-1 and mode-2) of χ represent the spatial information, the third dimensionality (i.e. mode-3) represents the RGB color information, and the mode-4 represents the time information, respectively. Then the left and right views of the stereo video can be represented by sets $\{\chi_1^L, \chi_2^L, \cdots, \chi_T^L\}$ and $\{\chi_1^R, \chi_2^R, \cdots, \chi_T^R\}$, respectively, where T is the number of GOF tensor in one view of the stereo video. The mode-3 matrix of χ is processed with SVD decomposition, so as to reduce the color dimension from three-dimensional to one-dimensional. And then, same processing is performed in time dimension to obtain the motion information map. At the same time, to obtain depth information of the stereo video, for each pair of GOFs, for example, the t-th GOFs of the left and right views ($1 \le t \le T$), the frames can also form a new 4th-order tensor set $B_t\{y_1, \ldots, y_S\}$, $y_i \in \mathbb{R}^{W \times H \times K \times V}$, where the first three dimensionalities (or modes) of y_i have the same meanings as that of χ_t^L and χ_t^R, while the fourth dimensionality V represents the view of the stereo video. After Tucker decomposition on the view and color dimensions of y_i, the $W \times H \times 3 \times 2$ tensor y_i is transformed to a $W \times H$ matrix that contains the main view and color information of y_i, then the S matrices corresponding to $B_t\{y_1, \ldots, y_S\}$ forms a new 3rd-order tensor $Z \in \mathbb{R}^{W \times H \times S}$. The dimension of the mode-3 matrix of Z can be reduced to obtain the motion information map which contains the time-varying information in the views as well as depth information between views, because the mode-3 matrix of Z is derived from the last two dimensionalities of y which represent the color and view dimensions respectively. After that, some statistical features are extracted in tensor decomposition domain and further pooled to obtain the overall video features. Finally, the relationship between the features and the subjective evaluation scores is modeled with random forest, and an objective metric used for predicting the quality of stereo video is obtained.

2.1. Tucker tensor decomposition

Tucker tensor decomposition decomposes a tensor into a set of matrices and a core tensor, and the Tucker family includes the Tucker1, Tucker2 and Tucker3 models. Tucker tensor decomposition can be described by

$$\min_{C, \mathrm{A}^{(1)}, \cdots, \mathrm{A}^{(N)}} \|\chi - C \times_1 \mathrm{A}^{(1)} \times_2 \mathrm{A}^{(2)} \cdots \times_N \mathrm{A}^{(N)}\|_F$$

$$\text{s.t. } C \in \mathbb{R}^{R_1 \times R_2 \cdots \times R_N},\ \mathrm{A}^{(n)} \in \mathbb{R}^{I_N \times R_n}, \quad n = 1, \cdots, N \tag{1}$$

where the symbol $\chi \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}$ is a tensor and C is the core of the tensor. A is 2D matrices. $||\bullet||_F$ is the Frobenius norm of the matrix. The n-mode (matrix) product of a tensor is denoted by $\times_n$. The matricization (also known as unfolding or flattening) of a tensor is the process of reordering the elements of an N-order array into a matrix $\mathrm{X}_{(n)}$, whose dimension is $I_n \times \prod_{j=1, j \ne n}^{N} I_j$. For example, let the elements of $\chi \in \mathbb{R}^{2 \times 2 \times 2}$ be $x_{111} = 1$, $x_{121} = 2$, $x_{211} = 3$, $x_{221} = 4$, $x_{112} = 5$,

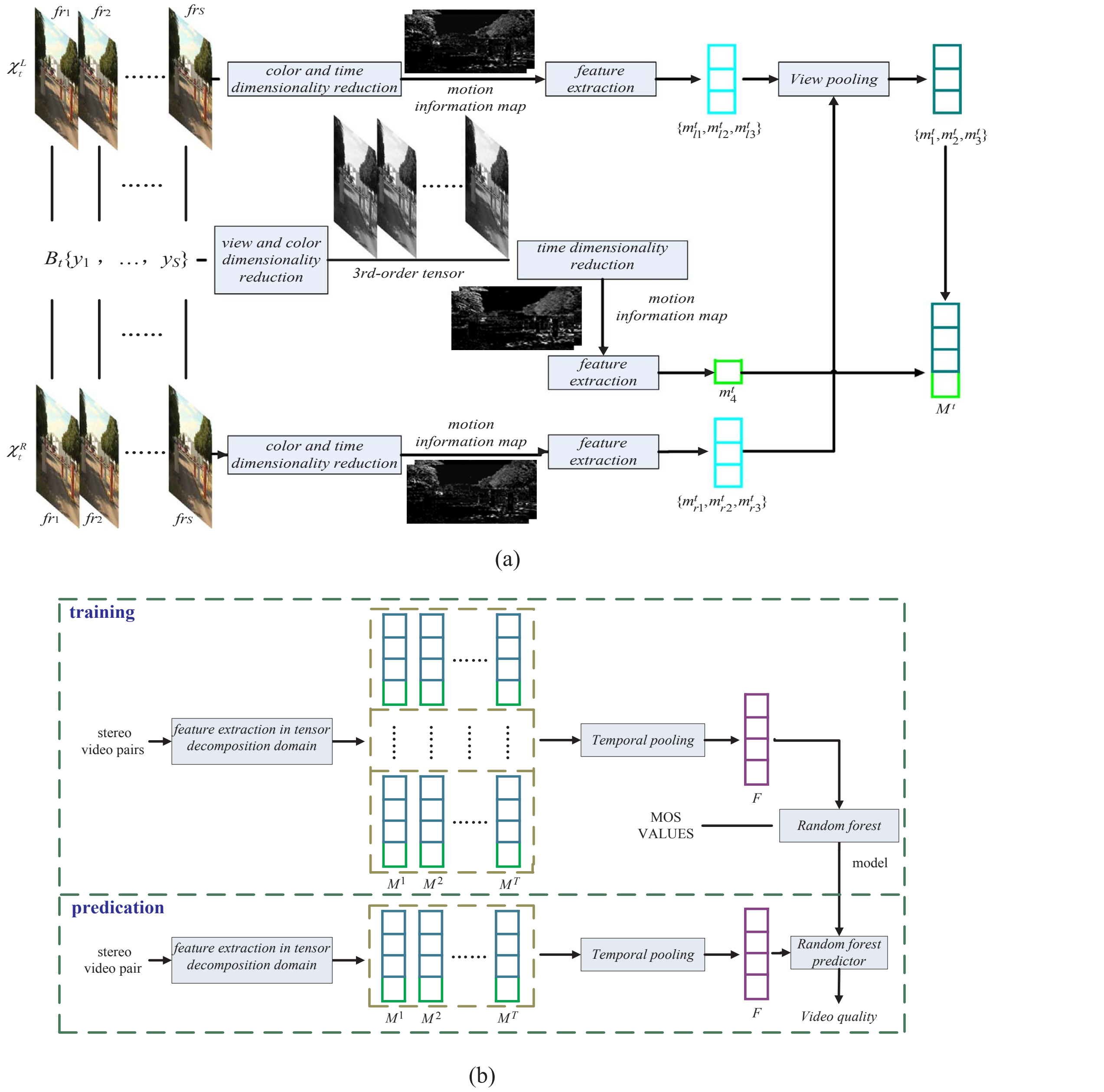


Fig. 1. Diagram of the proposed no reference stereo video quality assessment method. (a) Feature extraction in tensor decomposition domain. (b) Video quality prediction.

$x_{122} = 6$, $x_{212} = 7$, and $x_{222} = 8$, then the three mode-n matricizations are

$$X_{(1)} = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix} X_{(2)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix} X_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \tag{2}$$

We use Tucker decomposition to realize the dimensionality reduction, specifically, N-mode SVD is employed in this paper. Tucker decomposition is equal to higher-order PCA and SVD [21], and the core of the decomposition result not only retains the main information but also is stable. The value of n in this paper is from k_1 to N. Algorithm 1 shows the process details and can be denoted by Eq. (3) whereχexpresses a tensor. C and A are the core tensor and decomposition matrix, respectively.

$$\chi \approx C \times_{k_1} \mathrm{A}^{(k_1)} \cdots \times_N \mathrm{A}^{(N)} \tag{3}$$

Algorithm 1. Tucker decomposition process for one frame

```
  Input: tensor $\chi \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}$
  ----------------------------------------
    The core tensor C's dimension sizes $J_1, J_2, \ldots, J_N$
  Output: C
for $n$: $= k_1$ to $N$ do
    $\mathrm{A}(n) \leftarrow J_n$ leading left singular vectors of $\mathbf{X}_{(n)}$
end for
$C: = \chi \times_{k_1} \mathrm{A}^{(k_1)} \times_{k_1+1} \mathrm{A}^{(k_1+1)} \cdots \times_N \mathrm{A}^{(N)}$
Return C
```

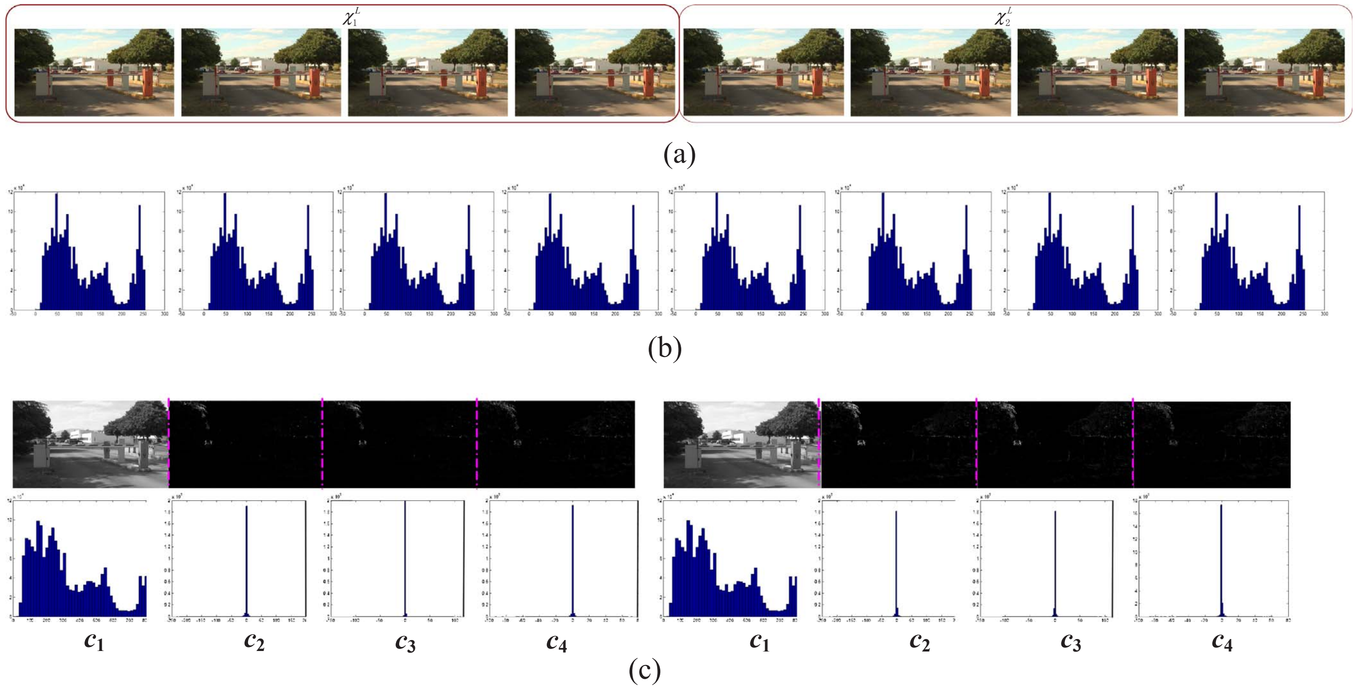


Fig. 2. Analysis of GOF tensor and Eigen-Actions. (a) Two GOF tensors in src1 sequence (from 1st frame to 8th frame). (b) Histogram of frames in (a) in pixel domain. (c) Mode-1 unfolding of the Eigen-Actions for the two GOF tensors.

2.2. Extraction of motion information

Having motion information is the significant difference between video and image. Therefore, VQA must take into account the inherent motion information of the video. In traditional VQA methods, motion vector and frame difference are often used as the motion information [22]. However, these methods are close to each other in the process of obtaining the motion information, and the obtained motion information is too short in time dimension to sufficiently describe the motion characteristics of the long time video. In this paper, we use tensor decomposition to obtain time-varying information of the video. The obtained time-varying information contains motion information within a relatively long time so that the loss of motion characteristics is as less as possible.

Since N-mode SVD is not suitable for the case that the data set is too large, we implement Tucker decomposition for GOF tensors of a video instead of the whole video. Each GOF consists of S frames. Fig. 2(a) gives two formed GOF tensors χ_1^L and χ_2^L where $S = 4$, obtained by the first 8 frames of the src1 sequence. The video sequences used in Section 2 for statistical analysis are listed in Table 1. Fig. 2(b) shows histograms of frames in Fig. 2(a) in pixel domain. Fig. 2(c) is the results of mode-1 unfolding of the core (called Eigen-Actions in this paper) of the two GOF tensors χ_1^L and χ_2^L, which are divided into sub-bands, and their corresponding histograms are also given, respectively. Here we set $k_1 = 3$ and $N = 4$ in Eq. (3). In this step, only the color dimension is reduced, and the time dimension still remains as S, which means that the Eigen-Actions of χ_1^Lcan be represented by $\boldsymbol{C} \in \mathbb{R}^{W\times H\times S}$. The form of a matrix $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \cdots \boldsymbol{c}_S]$ is used to represent the Eigen-Actions as a result of mode-1 unfolding, where $\boldsymbol{c}_i \in \mathbb{R}^{W\times H}$. In Fig. 2, $\boldsymbol{C}$ contains four sub-bands, and the sub-bands starting from $\boldsymbol{c}_2$ are the so-called motion information in this paper. From Fig. 2(b), it is seen that the histograms of adjacent frames are similar because of the similarity of the adjacent frames in pixel domain, however, the histograms have no obvious statistical regularity. But the situation is different in Fig. 2(c). After the decomposition of the Eigen-Actions by mode-1, the first sub-band $\boldsymbol{c}_1$ is basically same as the pixel domain content, that is, the background information in these S images. This is because the background information of the S frames constituting the tensor is substantially the same, so that the first principal component is the background information. By contrast, $\boldsymbol{c}_2$, $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$ relate to the movement of cars and branches, that is, the motion information of the video. From the corresponding histograms, we can find that the distribution of background information is close to histogram of pixel domain frames, while the distribution of motion information has statistical regularity, which

Table 1
Stereo video sequences used for statistical analysis.

No.	Distortion type and degree	Video name
1	ORG	src01_hrc00_s1920x1080p25n400v0.avi
2	H.264 (MOS = 4.357)	src01_hrc01_ip64_qi32_qp32_qb32_hier0_s1920x1080p25n400v0.avi
3	JP2K (MOS = 1.107)	src01_hrc04_bitrate2_s1920x1080p25n400v0.avi
4	Reduction of resolution	src01_hrc08_RedRes4_s1920x1080p25n400v0.avi
5	Image sharpening	src01_hrc09_sharpen20_s1920x1080p25n400v0.avi
6	Downsampling & sharpening	src01_hrc10_RedRes4_sharpen20_s1920x1080p25n400v0.avi
7	H.264 (MOS = 3.214)	src01_hrc02_ip64_qi38_qp38_qb38_hier0_s1920x1080p25n400v0.avi
8	H.264 (MOS = 1.571)	src01_hrc03_ip64_qi44_qp44_qb44_hier0_s1920x1080p25n400v0.avi
9	JP2K (MOS = 2.857)	src01_hrc05_bitrate8_s1920x1080p25n400v0.avi
10	JP2K (MOS = 4.036)	src01_hrc06_bitrate16_s1920x1080p25n400v0.avi
11	JP2K (MOS = 4.464)	src01_hrc07_bitrate32_s1920x1080p25n400v0.avi
12	H.264 (MOS = 4.357)	src01_hrc01_ip64_qi32_qp32_qb32_hier0_s1920x1080p25n400 v1.avi

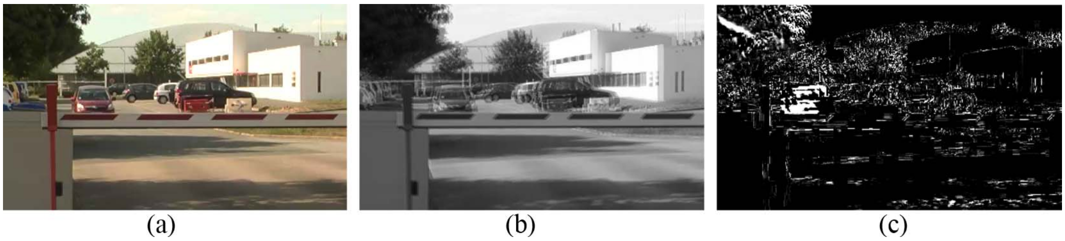


Fig. 3. (a) Partial region in one view of y_i. (b) partial region in color and view dimensionality reduction result of $y_i(\boldsymbol{c}_1)$. (c) partial region in time dimensionality reduction result of Z ($\boldsymbol{c}_2$).

means that motion information in tensor decomposition domain has potential to evaluate video quality. In this paper, the parameter S is set to 25 which is the frame rate of the video. Because the Tucker decomposition is realized by N-mode SVD, the more front the sub-band ranked, the more video information is included in the sub-bands. Therefore, we only utilize the second and the third sub-bands of the Eigen-Actions obtained by mode-1 unfolding, that is, $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$.

The most difference between 2D and 3D videos is that the 3D video can put viewers personally into the scene and enable human to be more deeply impressed. This kind of experience owes to depth perception caused by two views. Therefore, in this paper, the corresponding two views are grouped together to get a new tensor y_i, the 3rd and 4th dimensions of which represent color and view dimensions, respectively, and for each GOF there are S tensors, from y_1 toy_S. Then for each tensor y_i, which is in fact composed of the left and right views, Tucker decomposition is used to realize the dimensionality reduction of color and view, so that the dimension of y_i is changed from $W \times H \times 3 \times 2$ to $W \times H$. Fig. 3(a) and (b) show partial regions in one view of y_i and the color and view dimensionality reduction result of y_i. As we can see from Fig. 2(c), compared with other sub-bands, $\boldsymbol{c}_1$ contains the most common content of the tensor, so when we do tensor decomposition on view dimension, we will get the common content between views, which belongs to the content on the retina, and is important for depth perception, as shown in Fig. 3(b). The process is shown in Fig. 1(a). In this step, the parameters in Eq.(3) are $k_1 = 3$ and $N = 4$, too, but the meaning is different. After that, the S dimension reduced tensors are ordered in time dimension, so we get a new 3rd-order tensor $Z \in \mathbb{R}^{W \times H \times S}$. Then Eq. (3) is used again for time dimensionality reduction and here we set $k_1 = 3$ and $N = 3$ for Z to get the motion information maps of Z in a GOF, which is associated with the time-varying information in the views as well as depth information between views. Fig. 3(c) gives such a motion information map where the moving car and leaves shows relatively stronger response.

2.3. Feature extraction in tensor decomposition domain

Mittal et al. preprocessed image with local mean subtracted contrast normalized (MSCN) to obtain MSCN coefficients [22]. The MSCN coefficients of an image have some statistical regularities and can be simulated by the GGD model and the AGGD model. Then, image quality can be predicted by the damage degree of the MSCN coefficients of the distorted signal. For an image I, MSCN coefficients are defined by

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{4}$$

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i,j) \tag{5}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2} \tag{6}$$

where, $1 \le i \le H$, $1 \le j \le W$, H and W represent the height and width of the image respectively, $\omega = \{\omega_{k,l} | \text{-}K \le k \le K, \text{-}L \le l \le L\}$ is acircularly-symmetric Gaussian filter sampled out to 3 standard deviations. Fig. 4 shows histograms of MSCN coefficients of pixel domain and tensor decomposition domain with respect to some original and distorted videos in NAMA3DS1-COSPAD1 database [23]. Fig. 4(a) gives MSCN coefficient distributions with respect to original stereo video and videos distorted with different types of distortions, which illustrates the influence of different distortion types on video. The test video sequences include src1, src2, src3, src4, src5 and src6, as listed in Table 1. In Fig. 4(a), different colors represent different types of distortions. In Fig. 4(b), different colors indicate H.264 distorted videos with different Mean Opinion Score (MOS) values, and the test video sequences are src1, src2, src7 and src8. Fig. 4(b) shows how the distribution of MSCN coefficients changes along with the distortion degree of H.264 distortion. Similarly, Fig. 4(c) shows how the distribution of MSCN coefficients changes along with the distortion degree of JP2K distortion, and the test video sequences include src1, src3, src9, src10 and src11. From Fig. 4, it is seen that the histogram of MSCN coefficients in pixel domain in the first row of the figure looks like Gaussian distribution, and has capability to distinguish different distortion types to some extent. The second row of Fig. 4 is histograms of MSCN coefficients of $\boldsymbol{c}_1$ gotten from Eigen-Actions. It is seen that the distribution of the first sub-band $\boldsymbol{c}_1$ is similar to that of the pixel domain (the first row result). Therefore, both the pixel domain result and $\boldsymbol{c}_1$ have a certain capability to distinguish different types of distortions, but are poor in distinguishing distortion degree of certain distortion type, especially for H.264 distortion. The third row is the histograms of MSCN coefficients of the sub-band $\boldsymbol{c}_2$ (motion information map). By comparison of the results of the three rows, we can find that in regard to the capability of distinguishing distortion type and degree, $\boldsymbol{c}_2$ is relatively better than $\boldsymbol{c}_1$ and pixel domain, especially for the judgment of distortion degree of H.264.

The distribution of MSCN coefficients can be modeled by GGD model and AGGD model. GGD model is given by

$$f(x;\alpha,\sigma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} \exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right) \tag{7}$$

where, $\Gamma()$ is the gamma function, and

$$\beta = \sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{8}$$

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt \ a > 0 \tag{9}$$

The two parameters of GGD (α, σ^2) can be derived from L-moments estimation [12]. Furthermore, to measure the influence on MSCN coefficient distribute owing to distortions, AGGD model is used to extract feature, which is given by

$$f(x;\gamma,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\gamma}{(\beta_l+\beta_r)\Gamma(1/\gamma)} \exp\left(-\left(\frac{-x}{\beta_l}\right)^{\gamma}\right) & \forall\, x \leqslant 0 \\ \frac{\gamma}{(\beta_l+\beta_r)\Gamma(1/\gamma)} \exp\left(-\left(\frac{-x}{\beta_r}\right)^{\gamma}\right) & \forall\, x \leqslant 0 \end{cases} \tag{10}$$

$$\beta_l = \sigma_l\sqrt{\frac{\Gamma(1/\gamma)}{\Gamma(3/\gamma)}},\ \beta_r = \sigma_l\sqrt{\frac{\Gamma(1/\gamma)}{\Gamma(3/\gamma)}} \tag{11}$$

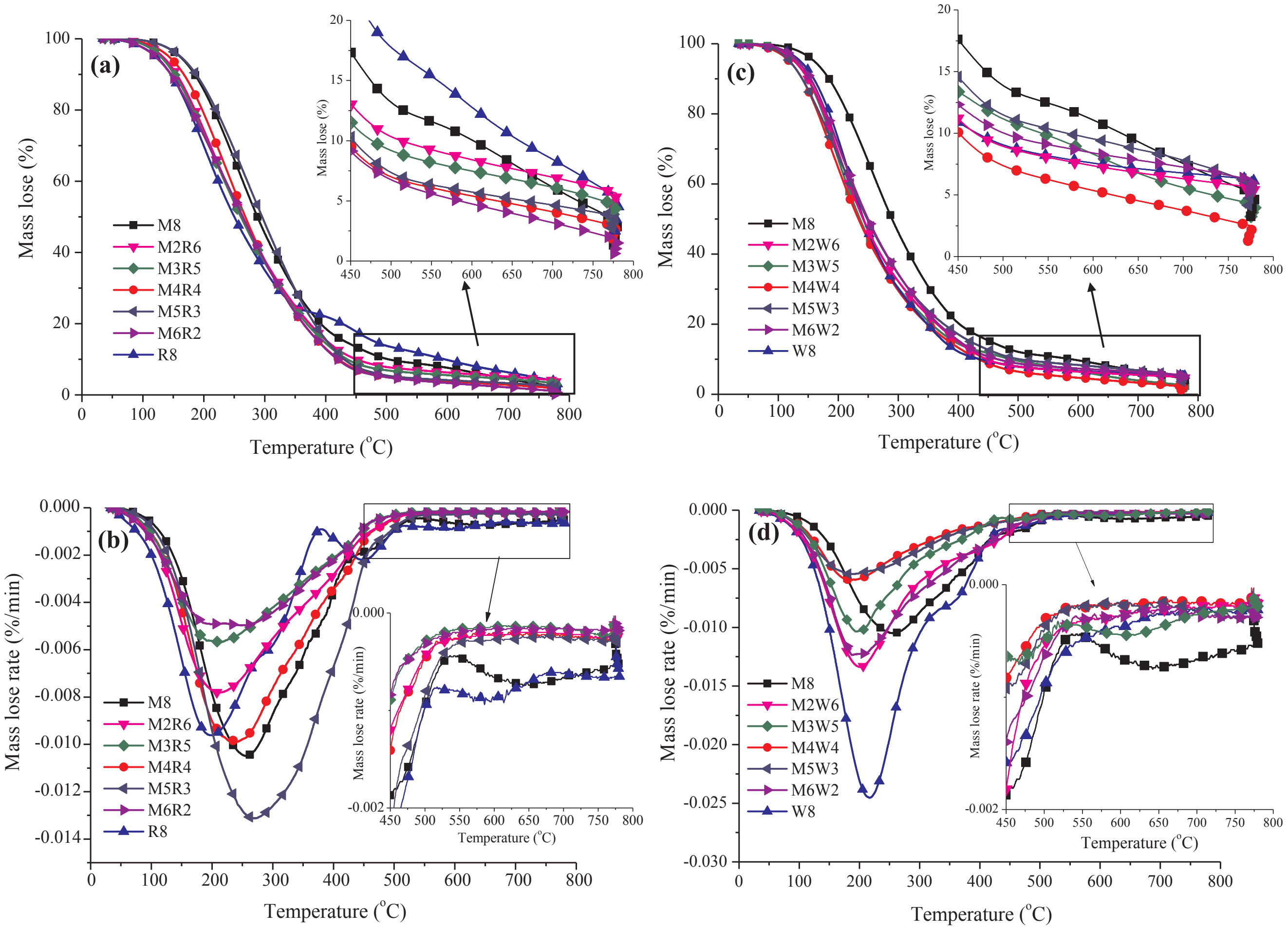


Fig. 4. Histograms of MSCN coefficients of original and distorted videos in pixel domain and tensor decomposition domain of tensor χ. (a) original and various distorted videos. (b) H.264 distorted videos under different distortion degrees. (c) JP2K distorted videos under different distortion degrees.

The parameters of AGGD (γ, σ_l^2, σ_r^2) can also be gotten by L-moments estimation. In addition, the mean of distribution is another parameters of AGGD and is given by

$$\mu = (\beta_r - \beta_l)\frac{\gamma(2/\alpha)}{\gamma(1/\alpha)} \tag{12}$$

The parameters of AGGD are computed along horizontal, vertical, main-diagonal and secondary-diagonal, respectively, so that total 16 parameters are gotten. All the parameters of GGD and AGGD are partial features of VQA predictor.

Local spatial and spectral entropy features are employed in image quality assessment, and these features have been proved to be effective for IQA [13]. Both spatial and spectral entropy features are computed on 8 × 8 blocks in this paper. Similar to MSCN coefficients, the histograms of spatial and spectral entropy also have regularity. The definition of spatial entropy is defined by

$$E_s = -\sum_{x} p(x)\log_2 p(x) \tag{13}$$

where x is the pixel value in the 8 × 8 blocks, and $p(x)$ corresponds to empirical probability density. 8 × 8 DCT coefficient blocks is used for computing spectral entropy. DCT coefficients are normalized to produce spectral probability

$$p(i,j) = \frac{C(i,j)^2}{\sum_{i}\sum_{j} C(i,j)^2} \tag{14}$$

where $1 \le i \le 8$, $1 \le j \le 8$, but the DC coefficient in DCT domain is excluded. Then the spectral entropy is given by

$$E_f = -\sum_{i}\sum_{j} p(i,j)\log_2 p(i,j) \tag{15}$$

Fig. 5 shows the histograms of spatial entropy in pixel domain and tensor decomposition domain, while Fig. 6 gives histograms of spectral entropy in the two domains. Figs. 5(a) and 6(a) are histograms with respect to original videos and their distorted versions under different distortion types, represented with different colors. The test video sequences are src1, src2, src3, src4, src5 and src6, respectively. Figs. 5(b) and 6(b) show the influence of different degrees of H.264 distortion on distributions of the spatial and spectral entropy in pixel domain and tensor decomposition domain, and different colors represent distorted videos with different MOS values. The used test video sequences include src1, src2, src7 and src8. Similarly, Figs. 5(c) and 6(c) show the situation of JP2K distortion with different distortion degrees, and the test video sequences are src1, src3, src9, src10 and src11.

From Figs. 5 and 6, we can find that different distortion types and degrees result in different distributions of spatial and spectral entropy. In the case of different distortion types and degrees, the spatial entropy of both sub-band $\boldsymbol{c}_1$ and pixel domain data did not show significant differences. Especially for H.264 distorted video, the spatial entropy distribution with respect to different distortion degrees are basically coincident, which means that $\boldsymbol{c}_1$ and pixel domain data are less useful for measuring video quality. By contrast, the spatial entropy histogram

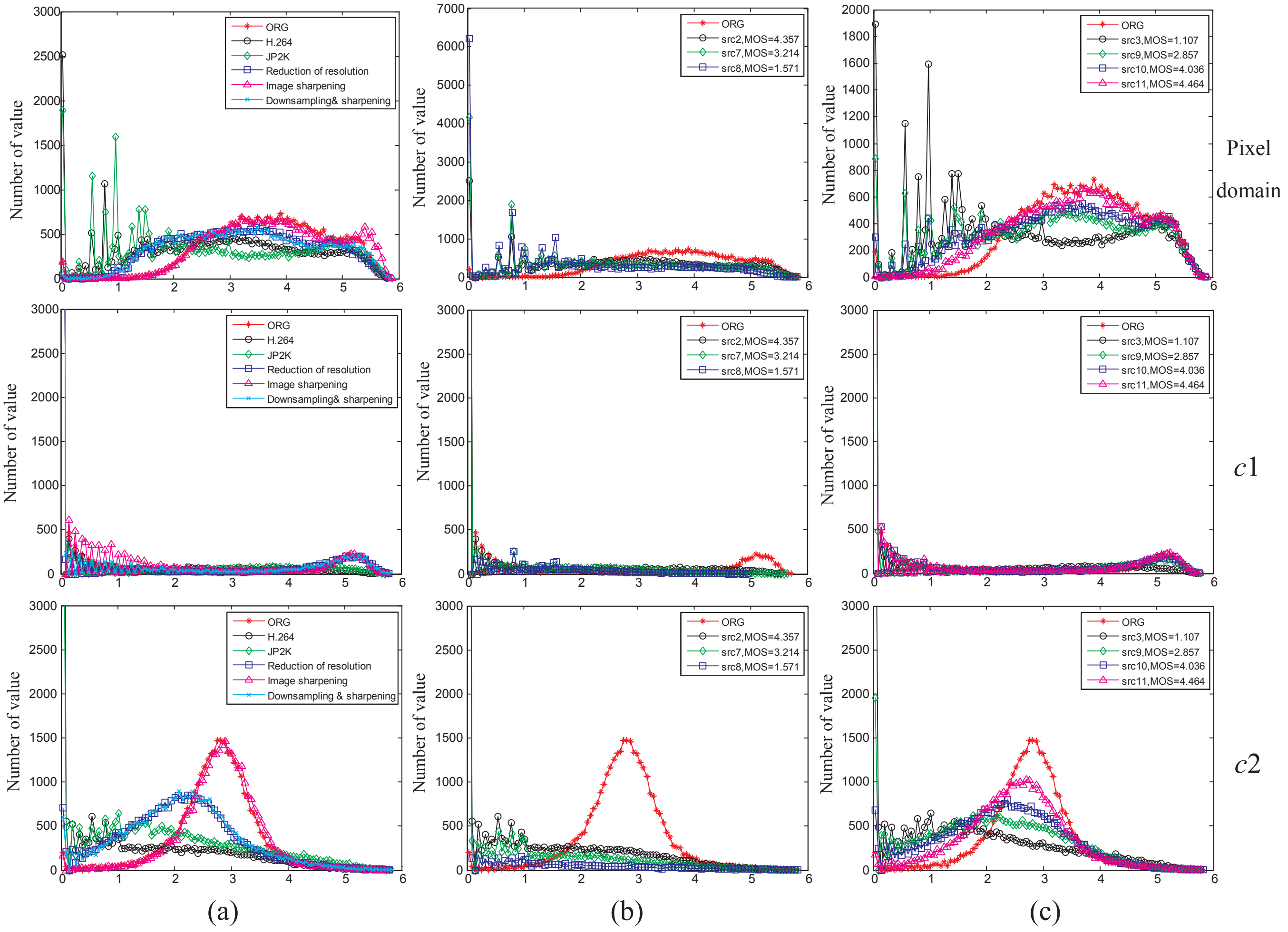


Fig. 5. Histograms of spatial entropy of original and distorted videos in pixel domain and tensor decomposition domain of tensor χ. (a) original and various distorted videos. (b) H.264 distorted videos under different distortion degrees. (c) JP2K distorted videos under different distortion degrees.

of sub-band $\boldsymbol{c}_2$ can distinguish the video's distortion type and degree through difference in skewness of the distribution. Unlike distribution of spatial entropy, after we compute spectral entropy in $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and pixel domain, all distributions of spectral entropy produce a large difference under different distortion types, as shown in Fig. 6(a). So, spectral entropy can be used for measuring video quality. But the use of spectral entropy has the same problem as MSCN coefficient histogram for H.264 distortion type. Distributions of spectral entropy of sub-band $\boldsymbol{c}_1$ and pixel domain data under different H.264 distortion degrees are basically coincident, and can not be used to characterize the distortion degree. However, as we can see from Fig. 6(b), unlike pixel domain and sub-band $\boldsymbol{c}_1$, the distributions of spectral entropy of $\boldsymbol{c}_2$ (motion information map) under different H.264 distortion degrees have different skewness and mean values. Thus, the value of skewness and mean of spectral entropy distribution of the motion information map can also be used as the video features sensitive to distortions. It should be pointed out that even the distortion degree of H.264 is quite different, the differences in the skewness and mean of spatial entropy distribution are still small. So the skewness and mean of spatial/spectral entropy distributions of the motion information map are still insufficient to accurately reflect the distortion degree of a video. Therefore, in this paper, the kurtosis of the distributions is also taken as a kind of perceptual distortion feature, so as to improve the correlation between the features and video quality.

NAMA3DS1-COSPAD1 database is used in our research. Because in the database, three types of distortions called resolution reduction, sharpening and their combination only have 10 videos for each type, the number of the samples is too small, additionally, these three types of distortions are related to clarity of an image, so we treat them as the same type of distortion, called D&S (downsamping &sharpening) in short.

Depth perception is important for stereo video. Spearman rank order correlation coefficient (SROCC) evaluates prediction monotonicity, while Pearson linear correlation coefficient (PLCC) and Root Mean Square Error (RMSE) indicate the consistency of the subjective and objective assessment of the quality of stereo videos. The given SROCC, PLCC and RMSE values in Table 2 shows the effectiveness of various features extracted in tensor decomposition domain of Z which relate to the time-varying information in the views as well as depth information between views on evaluating quality of stereo video. It is seen that the spatial entropy and parameters obtained by GGD and AGGD fitting with MSCN coefficients are relatively suitable for D&S distortion evaluation, and the spatial entropy is poor in H.264 and JP2K distortion evaluation. By contrast, spectral entropy achieves better comprehensive performance compared with the spatial entropy and parameters obtained by GGD and AGGD fitting with MSCN coefficients. Therefore, we only use spectral entropy of 3rd-order tensor Z in tensor decomposition domain for evaluating quality of stereo video. Since these features with respect to depth perception are obtained by tensor decomposition in this paper, only the results in tensor decomposition domain are shown in Fig. 7. From Fig. 7, it is seen that the distribution of spectral entropy of sub-band $\boldsymbol{c}_1$ also can not be used for distinguishing different distortion types. But there are significant differences in the skewness and mean of spectral entropy histogram of $\boldsymbol{c}_2$ for

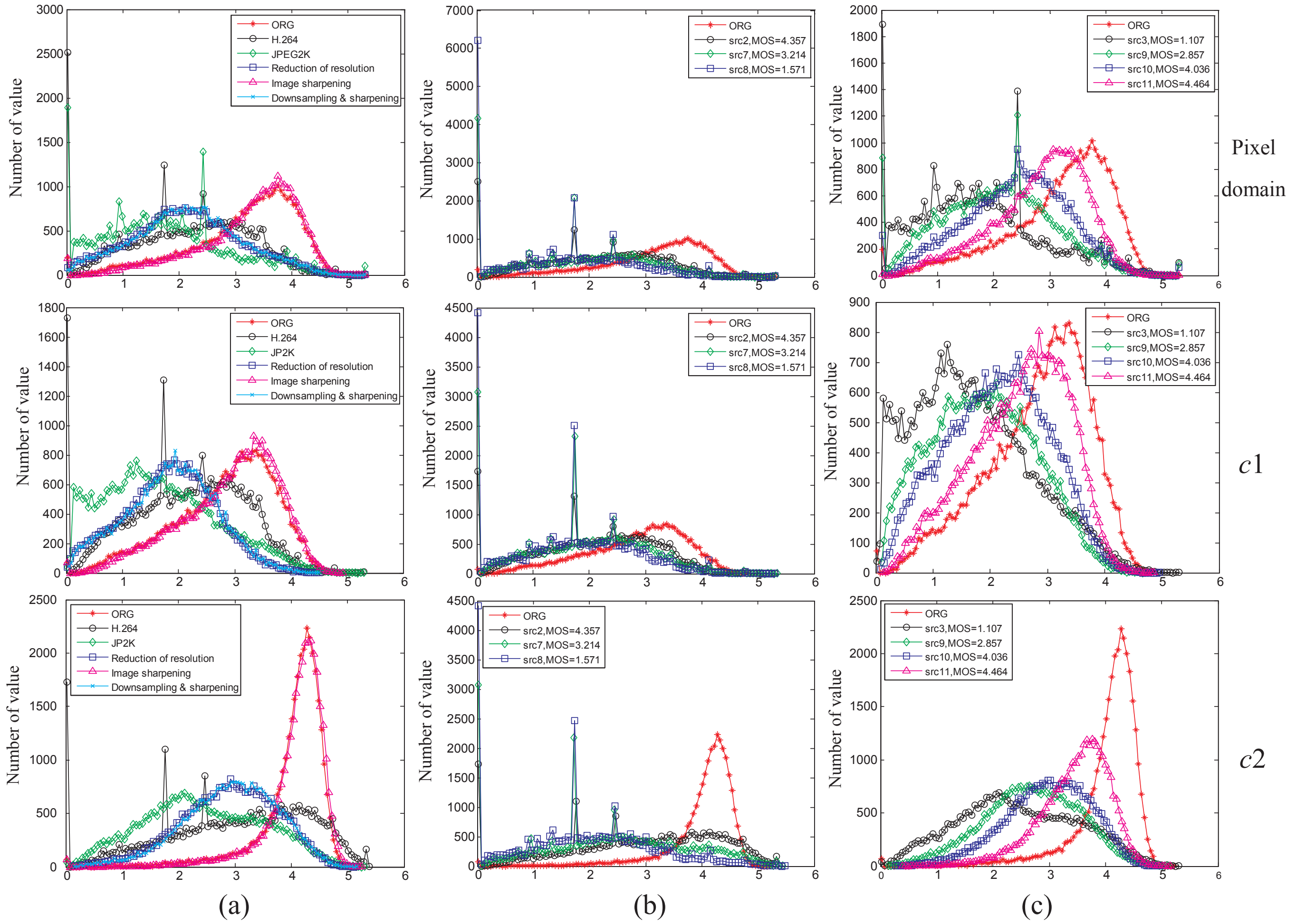


Fig. 6. Histograms of spectral entropy of original and distorted videos in pixel domain and tensor decomposition domain of tensor χ. (a) original and various distorted videos. (b) H.264 distorted videos under different distortion degrees. (c) JP2K distorted videos under different distortion degrees.

different distortion types and degrees. Thus the skewness and mean of spectral entropy histogram of $\boldsymbol{c}_2$ are used to measure the quality of stereo video.

2.4. Selection of motion information map for feature extraction

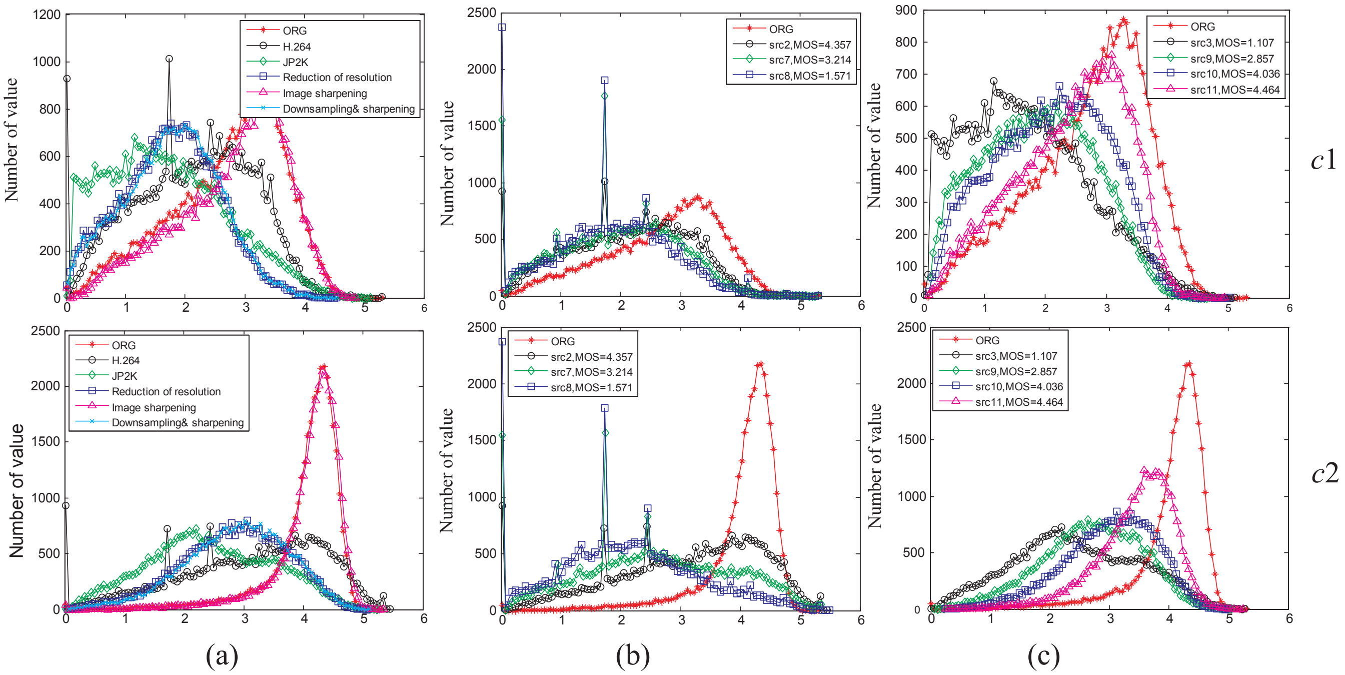
As mentioned above, in this paper four types of features, denoted as $\boldsymbol{F}[f_1,f_2,f_3,f_4]$, are used to evaluate the quality of stereo video. f_1, f_2 and f_3 represent GGD and AGGD parameters of MSCN coefficients, spatial entropy, spectral entropy respectively, and these three kinds of features are all extracted from tensor χ^L and χ^R ; f_4 represent features extracted from tensor Z which relates to depth perception. The tensor decomposition is achieved by N-mode SVD decomposition, and its operation is similar to PCA. Therefore, the more front a sub-band ranks in set $\boldsymbol{C}$, the more video information is included in the sub-bands. In this sub-section, we will discuss whether the latter motion information map has great impact on video quality evaluation.

Since a GOF tensor consists of 25 frames, that is, $S = 25$ in this paper, there are total 25 sub-bands after tensor decomposition. Fig. 8(a) shows all energies of the 25 sub-bands in a set $\boldsymbol{C}$, and Fig. 8(b) excludes the energy of $\boldsymbol{c}_1$. Fig. 8(c)–(f) also give the $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, $\boldsymbol{c}_3$ and $\boldsymbol{c}_{25}$ sub-bands after tensor decomposition on $\{\chi_1^L, \chi_2^L, \cdots, \chi_S^L\}$ of a GOF. It is seen that $\boldsymbol{c}_1$ contains primary video information compared with other twenty-four sub-bands. However, since VQA should pay more attention to time-varying contents of a video, while $\boldsymbol{c}_1$ mainly contains the background information of the video, our focus here is concentrated on $\boldsymbol{c}_2 \sim \boldsymbol{c}_{25}$ sub-bands. According to Fig. 8(b), we can get a conclusion that compared with $\boldsymbol{c}_2$, the later a sub-band ranks, the less information it will contain. This means that the front sub-bands already contain main time-varying information of a video.

Fig. 9 shows the effects of sub-bands $\boldsymbol{c}_2 \sim \boldsymbol{c}_9$ in the set $\boldsymbol{C}$ on assessing quality of the two views of stereo video when used for feature extraction. The ordinate of Fig. 9 is the PLCC value, which indicates the consistency of the subjective and objective assessment of the quality of

Table 2
Performance of features extracted in tensor decomposition domain of Z on evaluating quality of stereo video.

Features	SROCC				PLCC				RMSE			
	H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall
Spatial entropy	0.4857	0.6628	0.8281	0.5709	0.6519	0.7428	**0.9873**	0.6124	0.7845	0.7653	**0.1472**	0.8712
Spectral entropy	0.7647	**0.8729**	0.7537	0.7807	**0.8670**	**0.9638**	0.9759	**0.8083**	**0.4837**	**0.3105**	0.1944	**0.6519**
GGD + AGGD	**0.7674**	0.6547	**0.8503**	**0.7889**	0.7746	0.9416	0.9834	0.7844	0.6124	0.3660	0.1635	0.6889



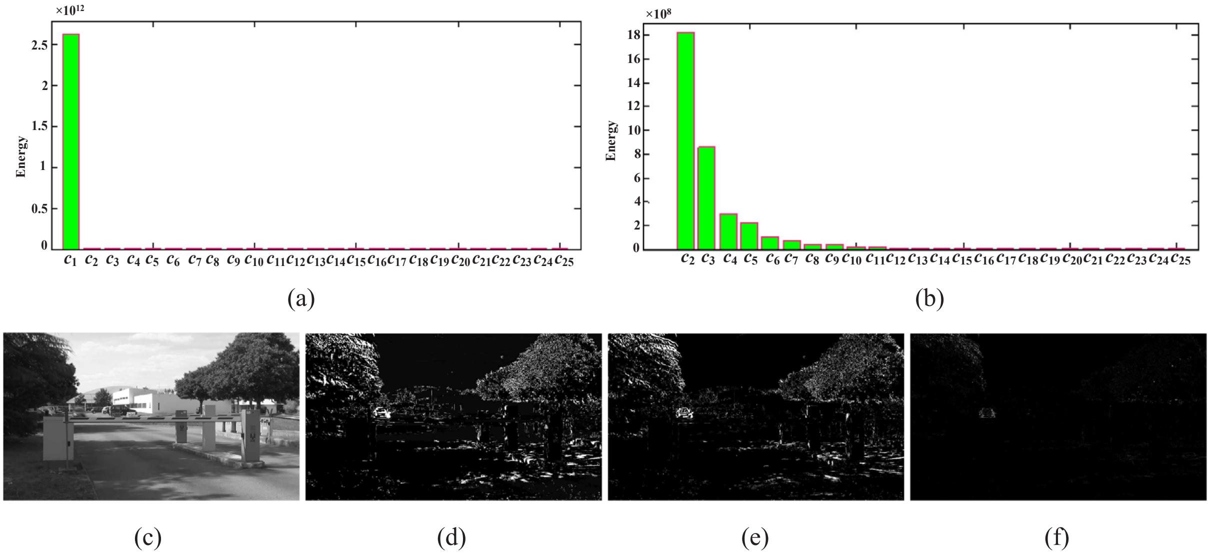
(a) (b) (c)

Fig. 7. Histograms of spectral entropy of original and distorted videos in tensor decomposition domain of tensor Z. (a) original and various distorted videos. (b) H.264 distorted videos under different distortion degrees. (c) JP2K distorted videos under different distortion degrees.

stereo videos. The PLCC value more approaches to 1, the better the performance of the corresponding feature is. From Fig. 9 it is seen that for features f_1 and f_3, $\boldsymbol{c}_2$ achieves obviously better results than the other seven sub-bands when evaluating videos with H.264 distortion, while for JP2K and D&S distortions, the differences among $\boldsymbol{c}_2 \sim \boldsymbol{c}_9$ are small, additionally, features f_1 and f_3 are more suitable for evaluating videos with JP2K and D&S distortions than for H.264 distortion. Feature f_2 is poor in evaluating H.264 distortion, the PLCC values are between 0.55 and 0.65, while for evaluating JP2K distortion features f_2 extracted from $\boldsymbol{c}_2$, $\boldsymbol{c}_3$, $\boldsymbol{c}_5$ and $\boldsymbol{c}_9$ sub-bands are relatively better than the others. On the whole, in regard to the effects of features f_1, f_2 and f_3 on evaluating distortions, the assessment accuracy of D&S distortion is the highest, followed by JP2K distortion, and the assessment accuracy of H.264 distortion is relatively poor.

Fig. 10 further gives the performances of features f_1, f_2 and f_3 extracted from multiple sub-bands, the groups are from one sub-band $\{\boldsymbol{c}_2\}$, two sub-bands $\{\boldsymbol{c}_2,\boldsymbol{c}_3\}$, ..., to eight sub-bands $\{\boldsymbol{c}_2,\boldsymbol{c}_3,\boldsymbol{c}_4,\boldsymbol{c}_5,\boldsymbol{c}_6,\boldsymbol{c}_7,\boldsymbol{c}_8,\boldsymbol{c}_9\}$. The abscissa represents the number of sub-bands included in a group, and the ordinate is the PLCC value. The performances of f_1 and f_2 features show declining trend along with the increase of the number of sub-bands for H.264 distortion, but when f_3 is used for H.264 distortion evaluation, the groups $\{\boldsymbol{c}_2,\boldsymbol{c}_3\}$ and $\{\boldsymbol{c}_2,\boldsymbol{c}_3, \boldsymbol{c}_4\}$ are relatively better. For JP2K and D&S distortions, there is little difference among different combinations of sub-bands.

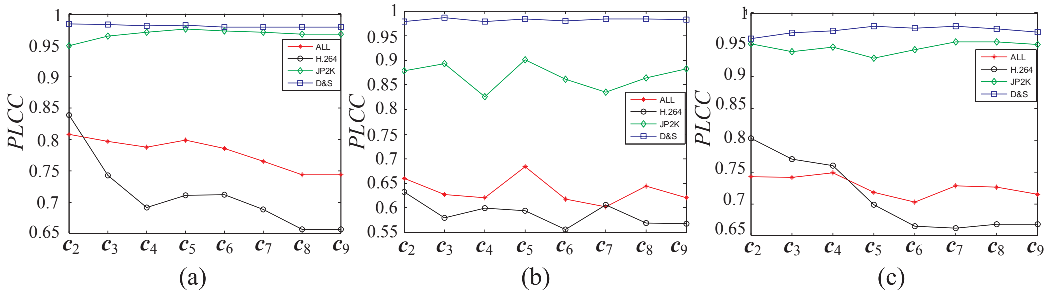
Fig. 11 shows how single sub-band and multiple sub-bands influence the performance of feature f_4. As we can see from Fig. 11(a), the curve with respect to H.264 distortion fluctuates greatly with different single sub-bands. The results of the first three sub-bands are better than those of the other five sub-bands. Fig. 11(b) is the results of multiple sub-band combinations. There is no obvious fluctuation in Fig. 11(b), and all of the PLCC values are more than 0.8. Therefore, the front sub-bands are more suitable for feature extraction when feature f_4 is



(a) (b)

(c) (d) (e) (f)

Fig. 8. Energy of sub-bands in set $\boldsymbol{C}$. (a) $\boldsymbol{c}_1 \sim \boldsymbol{c}_{25}$. (b) $\boldsymbol{c}_2 \sim \boldsymbol{c}_{25}$. (c) $\boldsymbol{c}_1$. (d) $\boldsymbol{c}_2$. (e) $\boldsymbol{c}_3$. (f) $\boldsymbol{c}_{25}$.



(a) (b) (c)

Fig. 9. Performance of f_1, f_2 and f_3 extracted from sub-bands $\boldsymbol{c}_2 \sim \boldsymbol{c}_9$ when used for VQA. (a) f_1. (b) f_2. (c) f_3.

considered. Additionally, Fig. 12 gives evaluation performance of features extracted from each pair of sub-bands for all distortions. It is seen that there is difference among different pairs of sub-bands, however, the fluctuation is low.

As mentioned above, the selection of motion information map has great impact on H.264 distortion evaluation, but the influence on JP2K and D&S distortion evaluation is relatively small. In order to ensure the overall performance of video quality evaluation, and to ensure that the number of extracted features should not be too much to bring over-fitting, $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ are used as motion information map for feature extraction in this paper. In addition to the sub-bands, the extracted features will also affect the evaluation performance. Through the experiments, it is found that we can achieve a good performance for the proposed method when we get features in the following way: 1) The MSCN coefficients of $\boldsymbol{c}_2$ associated with the two views of stereo video are fitted by GGD and AGGD models to obtain 18 features $f_1\{m^t_{l/r1,1}, m^t_{l/r1,2}, \ldots, m^t_{l/r1,18}\}$; 2) The spatial entropy and spectral entropy distributions are calculated for $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ with respect to the two views of stereo video respectively, and their mean, skewness and kurtosis are calculated respectively, then we get 2×3 spatial entropy features $f_2\{m^t_{l/r2,1}, m^t_{l/r2,2}, \ldots, m^t_{l/r2,6}\}$ and 2×3 spectral entropy features $f_3\{m^t_{l/r3,1}, m^t_{l/r3,2}, \ldots, m^t_{l/r3,6}\}$; 3) Calculate the distribution of spectral entropy for $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ related to depth perception and further obtain the mean, skewness and kurtosis of the distribution, so as to obtain 2×3 features as the depth perception feature $f_4\{m^t_{4,1}, m^t_{4,2}, \ldots, m^t_{4,6}\}$.

2.5. Feature fusion and quality evaluation

Since the above features are extracted per S frames, it is necessary to pool these features further for the whole stereo video. As mentioned above, for each view of a GOF of a stereo video, there are 18 features about the GGD and AGGD models of MSCN coefficients of $\boldsymbol{c}_2$. To pool the corresponding features of the left and right views of a GOF, the maximum value of each pair of the 18 kinds of features with respect to the left and right views is calculated, then the mean value of the feature over the whole video is calculated, thus 18 features of the whole video are obtained as f_1 of the stereo video. Eq. (16) gives the formula of calculating the ith feature in f_1 of the whole stereo video.

$$f_{1,i} = \frac{\sum_{t=1}^{T} \max(m^t_{l1,i}, m^t_{r1,i})}{T} \tag{16}$$

where T is the number of GOFs in the whole video, t is the index of GOF, and $1 \le i \le 18$.

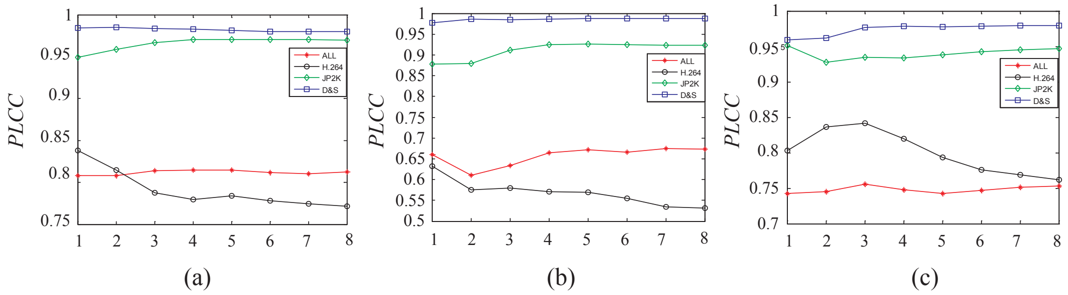
Similarly, for the ith spatial and spectral entropy features, we first obtain minimum of the pair of the left and right views and then compute the mean along the time dimension. The formula of calculating the ith features in f_2 and f_3 of the whole stereo video are as follows where $1 \le i \le 6$.

$$f_{2,i} = \frac{\sum_{t=1}^{T} \min(m^t_{l2,i}, m^t_{r2,i})}{T} \tag{17}$$

$$f_{3,i} = \frac{\sum_{t=1}^{T} \min(m^t_{l3,i}, m^t_{r3,i})}{T} \tag{18}$$

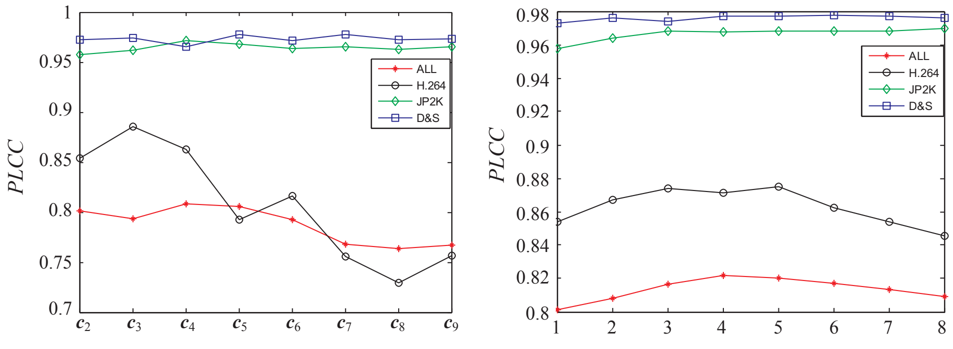
For the 6 features related to depth perception, the maximum and minimum of each of the 6 features along the time dimension are calculated, so that 12 depth perception features are finally obtained for the whole stereo video.

Therefore, there are total $18 + 6 + 6 + 12 = 42$ features represented as $\boldsymbol{F}[f_1, f_2, f_3, f_4]$ extracted from a stereo video for VQA. After the feature extraction, random forest is used to model the relationship between the features and subjective scores of the quality of stereo video. Random forest is an improvement to the decision tree algorithm. In the process of random forest training, it is assumed that the sample set P contains q samples, and there are B features associated with the samples. Then random sampling with replacement is used to form a training set P_i which also contains q samples, $1 \le i \le k$, and k is the number of the training sets. After that, k decision trees will be generated with the k training sets, and during the growth of the decision tree, only b ($b < B$) features are randomly selected from the total B features at current none-leaf nodes, each decision tree grows completely without the pruning. Finally, the test samples are then tested using each of the k decision trees to determine the regression results. In a random forest, the introduction of two randomities makes the random forest difficult



(a) (b) (c)

Fig. 10. Performance of f_1, f_2 and f_3 extracted from multiple sub-bands when used for VQA. (a) f_1. (b) f_2. (c) f_3.



(a) (b)

Fig. 11. Influence of selected sub-bands on evaluation performance of feature f_4. (a) Performance of f_4 extracted from single sub-band. (b) Performance of f_4 extracted from multiple sub-bands.

	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_2	0.7492	0.6349	0.6876	0.6876	0.7489	0.7479	0.7479
c_3		0.5327	0.7479	0.5121	0.7457	0.7046	0.7687
c_4			0.6904	0.4838	0.6813	0.5508	0.6948
c_5				0.6850	0.7164	0.7531	0.6835
c_6					0.6791	0.5416	0.6828
c_7						0.6965	0.7361
c_8							0.7664

(a)

	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_2	0.6105	0.6389	0.6549	0.6387	0.6339	0.6565	0.6339
c_3		0.6430	0.6867	0.6712	0.6690	0.6900	0.6819
c_4			0.6896	0.6616	0.6594	0.6795	0.6601
c_5				0.6600	0.6451	0.6608	0.6511
c_6					0.6173	0.6527	0.6314
c_7						0.6457	0.6231
c_8							0.6336

(b)

	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_2	0.7450	0.7596	0.7337	0.7288	0.7433	0.7464	0.7452
c_3		0.7432	0.7216	0.7212	0.7434	0.7422	0.7462
c_4			0.7298	0.7281	0.7439	0.7484	0.7546
c_5				0.6987	0.7192	0.7267	0.7305
c_6					0.7045	0.7121	0.7193
c_7						0.7094	0.7146
c_8							0.7078

(c)

	c_3	c_4	c_5	c_6	c_7	c_8	c_9
c_2	0.8092	0.8173	0.8254	0.8153	0.8109	0.8098	0.8125
c_3		0.8136	0.8128	0.8039	0.7959	0.7965	0.8010
c_4			0.8199	0.8111	0.8066	0.8037	0.8094
c_5				0.8028	0.7944	0.7891	0.7959
c_6					0.7913	0.7853	0.7902
c_7						0.7709	0.7766
c_8							0.7741

(d)

Fig. 12. Evaluation performance of features extracted from each pair of sub-bands. (a) PLCC of f_1. (b) PLCC of f_2. (c) PLCC of f_3. (d) PLCC of f_4.

Table 3
List of hypothetical reference conditions (HRCs) of NAMA3DS1-COSPAD1 dataset [23].

HRC	Impairments & Degradations	
	Type	Parameters
0	None – Reference sequence	
1	Video coding (H.264)	QP 32
2	Video coding (H.264)	QP 38
3	Video coding (H.264)	QP 44
4	Still image coding (JPEG2k)	2 Mb/s
5	Still image coding (JPEG2k)	8 Mb/s
6	Still image coding (JPEG2k)	16 Mb/s
7	Still image coding (JPEG2k)	32 Mb/s
8	Reduction of resolution	↓4 downsampling
9	Image sharpening	Edge enhancement
10	Downsampling & sharpening	HRC 8 + HRC 9

to fall into the overfitting. It improves the accuracy of the algorithm without increasing the computational cost.

3. Experimental results and discussions

We verified the performance of the proposed MNSVQM method on the stereo video database NAMA3DS1 -COSPAD1 [23]. There are total 100 symmetric distorted stereo videos derived from 10 original videos with five types of distortions including H.264/AVC compression, JPEG 2000 compression (JP2K), reduction of resolution, sharpening, and combination of resolution reduction and sharpening. Nine of the 10 original videos compose of 400 frames and the last one has 325 frames. Table 3 gives the details of the 10 distortions in database NAMA3DS1-COSPAD1. The sequences feature 1920 × 1080 progressive Full HD resolution per view and 25 frames per second. Two of these sequences have two scenes (with one scene cut), while the others have just one scene. The MOS value of every pair of stereo video is accessible in the database. The range of MOS is from 0 to 5, and the higher the MOS value is, the better the subjective quality of stereo video will be.

PLCC and SROCC are calculated to evaluate the performance of the proposed objective metric. The range of SROCC is [−1, 1], and the closer the absolute value of SROCC approaches to 1, the better the performance of the objective metric will be.

3.1. Analysis and selection of S value

As mentioned above, the advantage of feature extraction in tensor decomposition domain is that the extracted features can distinguish not only the image distortion type and but also the distortion degree. However, considering the computational complexity, N-mode SVD is implemented on GOF tensor instead of the whole video in this paper. Thus, the number of frames in a GOF, that is, the parameter S, is important for the proposed motion feature based no reference stereo video quality metric.

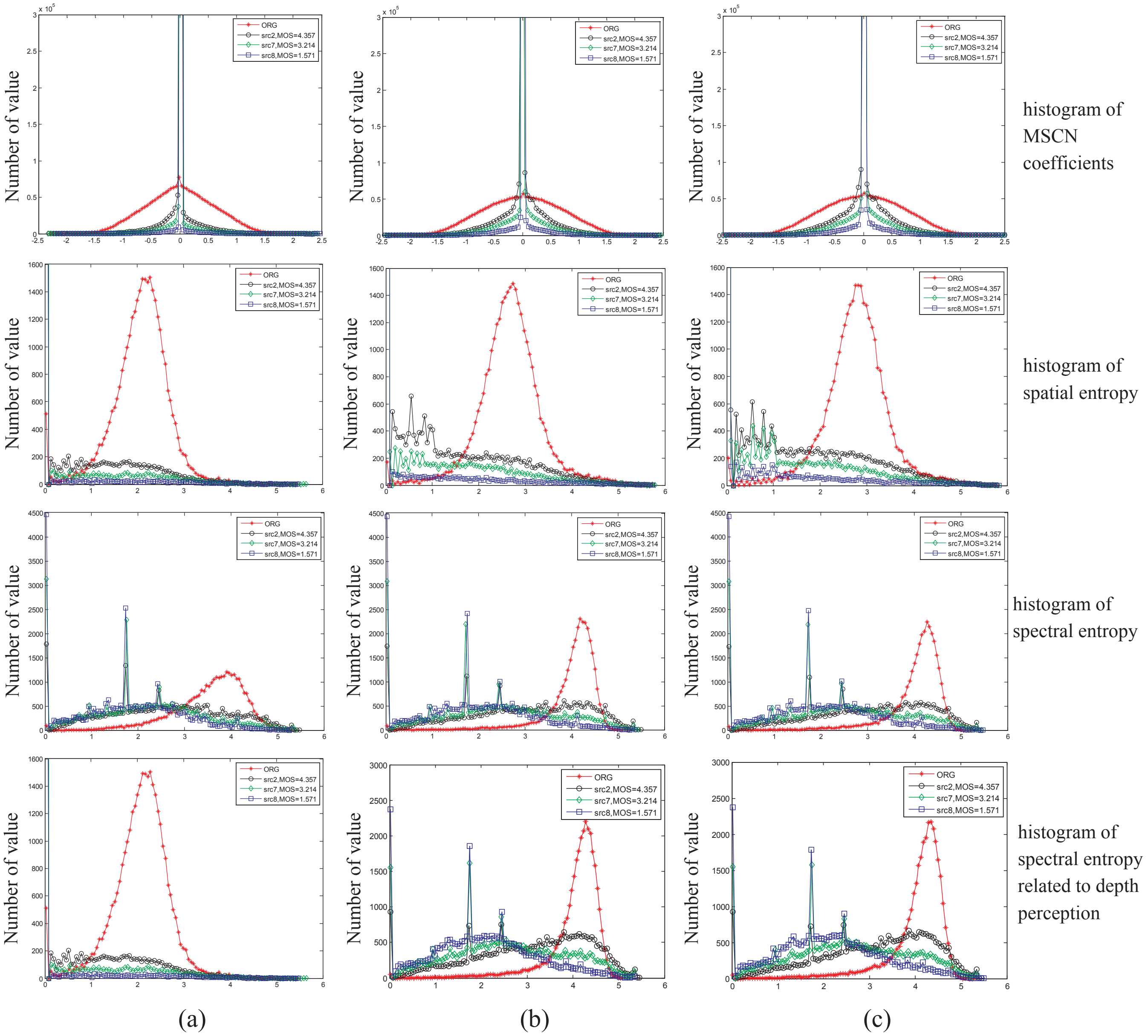


Fig. 13. Effect of S on evaluating quality of stereo video with H.264 distortion. (a) $S = 4$. (b) $S = 16$. (c) $S = 25$.

Since the most obvious difference between feature extraction in tensor decomposition domain and pixel domain is the capability of distinguishing distortion degree of H.264 distortion, the MSCN coefficients, spatial entropy and spectral entropy with respect to different distortion degrees of H.264 distortion under different S values are compared and the results are shown in Fig. 13. Because the frame rate of the videos used in this paper is 25fps, the result of $S = 25$ is compared with that of $S = 16$ and $S = 4$. From Fig. 13, it can be seen that the distributions of $S = 16$ and $S = 25$ are similar, but are quite different to distribution of $S = 4$. For example, when MSCN coefficient distribution is considered, discrimination of $S = 16$ and $S = 25$ for MOS = 1.571 and MOS = 3.214 is more obvious than that of $S = 4$. The situation of spatial entropy is similar to that of MSCN coefficient. For spectral entropy, when $S = 16$ or $S = 25$, skewness of MOS = 4.375 is obvious different from that of MOS = 1.571 and MOS = 3.214, but the three curves almost overlap with each other when $S = 4$. The effect of the three S values on the final depth perception features is not very differentiable. On the whole, the choice of $S = 16$ or $S = 25$ will lead to better results.

Table 4
Effect of S selection on processing time and prediction accuracy.

S value	Process time (h)	SROCC				PLCC				RMSE			
		H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall
$S = 4$	2.5557	0.3954	0.8810	0.7714	0.7272	0.5821	0.9689	0.9846	0.7596	0.8367	0.2891	**0.1562**	0.7216
$S = 16$	1.6989	0.6000	**0.9048**	**0.7945**	0.7810	0.7313	**0.9718**	0.9844	0.7991	0.6648	**0.2725**	0.1564	0.6644
$S = 25$	**1.6488**	**0.7714**	0.8982	0.7775	**0.8394**	**0.8850**	0.9706	**0.9850**	**0.8611**	**0.4675**	0.2769	**0.1562**	**0.5634**

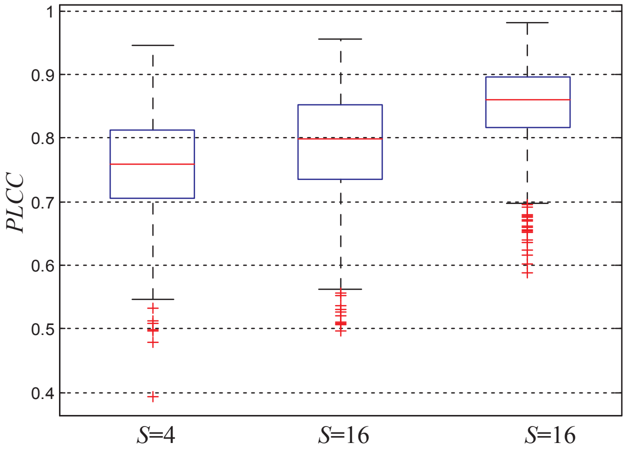


Fig. 14. PLCC distribution with respect to different S.

Table 4 gives the comparison of processing time with respect to the three different S values where the stereo video src2 and src12 in the NAMA3DS1-COSPAD1 database are used as the test video. The resolution of the stereo video is 1920 × 1080 and there are total 400 frames in a video. The computer used is with Intel(R) Core(TM) i7-3770 CPU @3.40 GHz, 8G RAM, Windows 7 64-bit, MATLAB® R2014a. The unit in Table 4 is hour (h). It can be seen that $S = 4$ is most time-consuming, while $S = 16$ and $S = 25$ are comparative, and $S = 25$ is relatively the most time-saving one among the three cases.

Table 4 also shows the effect of S selection from the perspective of evaluation performance. PLCC and RMSE indicate the accuracy of the predication, while SROCC demonstrates prediction monotonicity, and the best ones are in bold. All experiments were conducted with 80% of the videos for training and the remaining 20% for testing. The random forest in this model was composed of 1380 decision trees. The number of training sets was 42, and the number of construction decision trees was 6. To ensure that objective metric was robust across all contents and distortion severities, all the tests were conducted 1000 times with random permutations. The median SROCC, PLCC and RMSE values over the 1000 trials are shown in Table 4, and the corresponding distribution ranges of PLCC are depicted in Fig. 14, the ordinate of which represents the PLCC value.

The results in Table 4 show that the S value has significant effect on the evaluation performance of H.264 distortion. Although the performance of $S = 16$ and $S = 25$ in Fig. 13 appears to be similar, the results in Table 4 show that there is obvious difference between the two situations. The main reason is that Fig. 13 lists only the results of a tensor motion information map ($\boldsymbol{c}_2$ in one GOF) rather than the characteristics histogram distribution of all the motion information map of a video. Fig. 14 gives the boxplot of PLCC with respect to different S. In the figure, the top and bottom of the blue boxes indicates the 75th percentile and 25th percentile of the datum, while the red line in the blue boxes indicates the 50th percentile of the datum (the median). Therefore, the more flat the box is, the more concentrated the datum are. Further, it can be seen from the plot of Fig. 14 that when $S = 25$, all the maximum value, the minimum value and the median value are higher than those of the other two cases. At the same time, the difference value between the maximum value and the minimum value is smaller than the other two situations. And when $S = 25$, the data in 25% and 75% is also more concentrated. From the results shown in Table 4 and Fig. 14, it can be seen that the larger the S value is, the better the evaluation performance is. However, due to the limitation of the data size of the tensor process by N-mode SVD, the S value should not be too large. So, S is set to 25 in this paper, that is, a GOF is composed of the signal within one second.

Table 6
PLCC results of the combination of the features in tensor decomposition domain and pixel domain.

	Features	H.264	JP2K	D&S	Overall
Tensor decomposition domain	(f_1,f_2)	0.8327	0.9624	0.9868	0.8301
	(f_1,f_3)	0.8330	0.9511	0.9893	0.8054
	(f_2,f_3)	0.7874	0.9449	0.9866	0.7599
	(f_1,f_2,f_3)	**0.8488**	**0.9718**	0.9871	**0.8445**
Pixel domain	(f_1,f_2)	0.6924	0.9445	0.9840	0.7817
	(f_1,f_3)	0.6954	0.9473	0.9835	0.7817
	(f_2,f_3)	0.6498	0.4767	**0.9908**	0.4743
	(f_1,f_2,f_3)	0.6926	0.9493	0.9858	0.7885

Table 7
PLCC results of the combination of depth perception feature f_4 in tensor decomposition domain.

Tensor decomposition domain	H.264	JP2K	D&S	Overall
(f_1,f_4)	0.8836	0.9664	0.9837	0.8597
(f_2,f_4)	0.8637	0.9638	0.9816	0.8111
(f_3,f_4)	**0.8944**	0.9625	0.9781	0.8130
(f_1,f_3,f_4)	0.8895	0.9680	0.9842	0.8600
(f_2,f_3,f_4)	0.8802	0.9621	0.9820	0.8128
(f_1,f_2,f_3,f_4)	0.8850	**0.9706**	**0.9850**	**0.8611**

3.2. Performance of the proposed objective metric

In order to analyze the contribution of each feature to the quality evaluation, Table 5 shows the comparison results of the features extracted from the motion information map in tensor decomposition domain with the same features directly extracted in pixel domain. Obviously, for the three features extracted in pixel domain, the evaluation performance of the f_1 feature is superior to the other two features. However, the corresponding PLCC value of the f_1 feature extracted in pixel domain has only 0.6926 for H.264 distortion, far below 0.8384, which is the result of f_1 feature extracted in tensor decomposition domain. There are greater gap in overall evaluation performance when using f_2 and f_3. The reason is that the feature extraction in pixel domain ignores the time and motion information of a video, while the motion information map obtained by tensor decomposition contains the main motion information of the video. Therefore, the features extracted in tensor decomposition domain are superior to features extracted in pixel domain when used for video quality assessment.

Table 5
PLCC results of different features for different distortions.

Feature	H.264		JP2K		D&S		Overall	
	Tensor decomposition domain	Pixel domain	Tensor decomposition domain	Pixel domain	Tensor decomposition domain	Pixel domain	Tensor decomposition domain	Pixel domain
f_1	0.8384	0.6926	0.9493	0.9349	0.9844	0.9805	0.8083	0.7698
f_2	0.5800	0.5861	0.8933	0.4333	0.9863	0.7031	0.6267	0.3288
f_3	0.8368	0.6754	0.9272	0.4099	0.9618	0.9888	0.7450	0.4977
f_4	0.8670		0.9638		0.9759		0.8083	

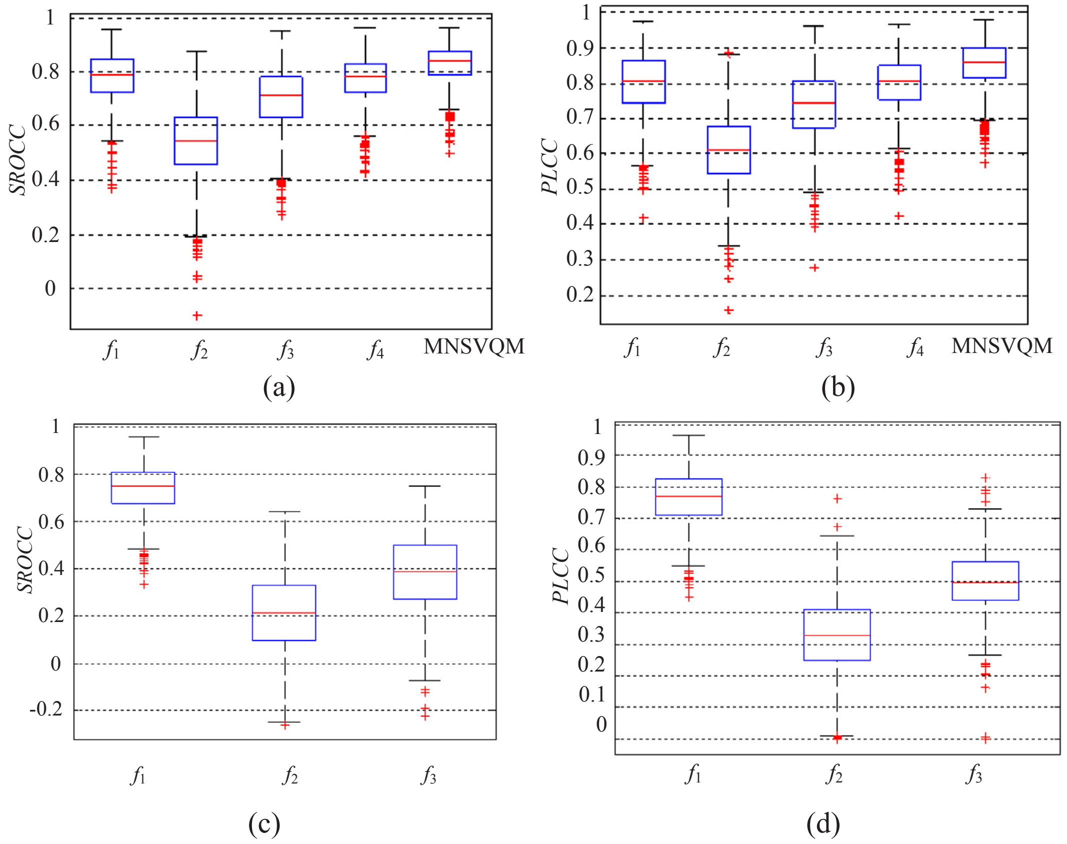


Fig. 15. Statistical results of the evaluation performance of each feature. (a) SROCC value in tensor decomposition domain. (b) PLCC value in tensor decomposition domain. (c) SROCC value in pixel domain, (d) PLCC value in pixel domain.

Table 8
Predication performance of different VQM methods on NAMA3DS1-COSPAD1 database.

Metrics	SROCC				PLCC				RMSE			
	H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall	H.264	JP2K	D&S	Overall
PSNR	0.5282	0.6087	0.7009	0.5415	0.5809	0.7176	0.6759	0.5868	0.9421	0.8677	0.6740	0.9198
SSIM	0.7804	**0.9186**	0.5273	0.7480	0.7616	0.9408	0.6014	0.7797	0.7500	0.4223	0.7306	0.7112
MS-SSIM	0.6469	0.9083	0.6746	0.7691	0.6812	0.9195	0.8085	0.7831	0.8474	0.4897	0.5382	0.7063
VIF	**0.8364**	0.8890	**0.8100**	**0.8408**	0.8069	0.8675	0.6751	0.8380	**0.3476**	0.2834	0.3177	0.5144
VQM	0.7715	0.8021	0.6499	0.7020	0.8097	0.8352	0.6714	0.7242	0.6793	0.6851	0.6777	0.7833
MNSVQM	0.7714	0.8982	0.7775	0.8394	**0.8850**	**0.9706**	**0.9850**	**0.8611**	0.4675	**0.2769**	**0.1562**	**0.5634**

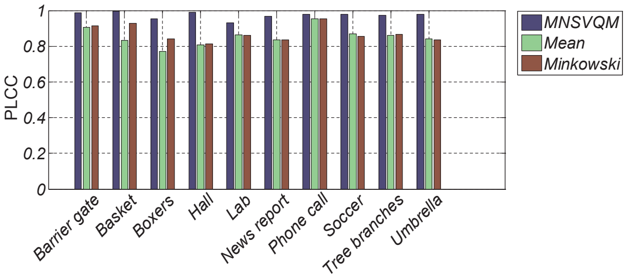


Fig. 16. Comparison results of MNSVQM and the 3D VQA methods BEVQM pooling by mean and Minkowski, respectively [7].

Table 9
Process time of different metrics.

Metrics	PSNR	SSIM	MS-SSIM	VIF	VQM	MNSVQM
Process time (h)	0.3814	0.4093	0.4312	1.1229	0.8693	1.6488

Tables 6 and 7 give the results of further consideration of the various combinations of features. Since pixel domain does not have the f_4 feature, the results of combinations of f_4 in tensor decomposition domain are separately shown in Table 7. As can be seen from Table 6, when features f_2 and f_3 are combined, the PLCC value for D&S distortion can reach 0.9908, which is the best for D&S distortion, however, this combination are poor in evaluating other types of distortions, and the PLCC value of overall evaluation performance is only 0.4743.

The evaluation performances of f_1, f_2 and f_3 in tensor decomposition domain are obviously better than that in pixel domain, and the combination of them can achieve better effects for all the distortion types. Moreover, as shown in Table 7, considering the depth perception feature f_4 can further improve the performance of the VQA. The overall performance of the video database with respect to the combination of f_1, f_2, f_3 and f_4 reaches 0.8611, which suggests that the combination of the four features in tensor decomposition domain can achieve a better performance for stereo VQA.

Fig. 15 shows the statistical results of the performance of the overall evaluation of all distortion types using the various features extracted in the tensor decomposition domain, pixel domain and the combination of the four features in the tensor decomposition domain. The abscissa represents the feature used for the test, and the ordinate represents the statistical result of the SROCC or PLCC values of 1000 tests. From Fig. 15, it is found that features extracted in tensor decomposition domain achieve better performance than in pixel domain, features f_1 and f_4 have the greatest contribution to the proposed no reference stereo video quality evaluation method, and this conclusion is also proved by Table 5. The MNSVQM is the video quality assessment method proposed in this paper, which combines the four features together. In comparison with the results of MNSVQM in Table 7 (ie. the results of (f_1,f_2,f_3,f_4) in tensor decomposition domain), it can be seen that the proposed method is good at various distortion types. The evaluation performance of JP2K, reduction of resolution, sharpening and combination of resolution reduction and sharpening is better than that of H.264. Although the overall evaluation of a single feature is less than 0.81 (the overall result in Table 5), the overall result of the combination of the four features achieves 0.86, which means that the combination is successful for stereo VQA.

We compared the proposed no reference MNSVQM ($S = 25$) with some of full reference objective video quality metrics: PSNR, SSIM, MS-SSIM, VIF [24] and VQM [25]. Table 8 shows the SROCC, PLCC and RMSE values of these full reference methods and the proposed no reference MNSVQM on the stereo video databases, the best of each kind of index is in bold. It is seen that the proposed method achieves best PLCC and RMSE values, which means that the objective scores predicated by the proposed method is better consistent with the subjective quality assessment results. From the perspective of prediction monotonicity, the VIF performs better than the proposed method but the gaps are not big. We also compared our MNSVQM method with the full reference 3D VQA methods BEVQM pooling by mean and Minkowski, respectively [7]. The results are shown in Fig. 16, the abscissa of which represents the ten stereo video sequences in database NAMA3DS1-COSPAD1, and the ordinate is the PLCC value. The results confirmed that the proposed method is promising. In addition, Table 9 gives the process time of different metrics. In these experiments, the resolution of the stereo video is 1920 × 1080 and there are total 400 frames in a video. The computer used is with Intel(R) Core(TM) i7-3770 CPU @ 3.40 GHz, 8G RAM, Windows 7 64-bit, MATLAB® R2014a. The proposed no reference MNSVQM, which is the most time-consuming, costs 1.6488 h to assess the video, and the time spending for tensor decomposition is more than six times as that for feature extraction on average. The VIF ranks the second, followed by VQM, MS-SSIM, SSIM and PSNR. From this point of view, the proposed method has room to be improved.

4. Conclusion

In this paper, we presented a no reference video quality metric named MNSVQM. Tucker tensor decomposition is used to obtain the main motion information maps which contain the time-varying information of a stereo video. We extract video features from motion information maps of each view of stereo video. Additionally, depth perception features are extracted from motion information maps of tensors associated with depth perception. Then these features are pooled as a preparation for the next step. Random forest is adopted to model human visual perception so as to predict the quality of stereo video. The experimental results show that the proposed method can solve the problem that the feature extraction in pixel domain is ineffective for H.264 and other distortion types. The experimental results of the selection of the S which make up the tensor proves that the more video frames a tensor contains, the more time-varying information will be obtained, and more accurate prediction result will be achieved for the VQM. That is, the time-varying information that contains long time information is of great significance to evaluate the video quality. Results on NAMA3DS1-COSPAD1 database prove that the proposed MNSVQM method is promising in video quality prediction. However, the proposed MNSVQM method mainly considers the motion information of the stereo video, but ignores the effect of sub-band $\boldsymbol{c}_1$, which contains the most background information of the video. In the future, we will also investigate the features of sub-band $\boldsymbol{c}_1$ associated with the quality of stereo video so as to further improve the performance of the method.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant nos. U1301257 and 61671258, and 61620106012, the National High-tech R&D Program of China under Grant no. 2015AA015901, and the Natural Science Foundation of Zhejiang Province under Grant nos. LY15F010005 and LY16F010002. It is also sponsored by K.C. Wong Magna Fund in Ningbo University.

References

[1] S. Winkler, D. Min, Stereo/multiview picture quality: overview and recent advances, Signal Process. Image Commun. 28 (10) (2013) 1358–1373.
[2] P. Hanhart, E. Bosc, P. Le Callet, T. Ebrahimi, Free-viewpoint video sequences: a new challenge for objective quality metrics, 16th International Workshop on Multimedia Signal Processing (MMSP), Jakarta, Indonesia, 2014, pp. 79–92.
[3] D.K. Broberg, Infrastructures for home delivery, interfacing, captioning, and viewing of 3-D content, Proc. IEEE 99 (4) (2011) 684–693.
[4] A. Mittal, M.A. Saad, A.C. Bovik, A completely blind video integrity oracle, IEEE Trans. Image Process. 25 (1) (2016) 289–300.
[5] F. Zhang, D.R. Bull, A perception-based hybrid model for video quality assessment, IEEE Trans. Circuits Syst. Video Technol. 26 (6) (2016) 1017–1028.
[6] M. Yu, K. Zheng, G. Jiang, et al., Binocular perception based reduced-reference stereo video quality assessment method, J. Vis. Commun. Image Represent. 38 (2016) 246–255.
[7] C. Galkandage, J. Calic, S. Dogan, et al., Stereoscopic video quality assessment using binocular energy, IEEE J. Sel. Top. Signal Process. 11 (1) (2017) 102–112.
[8] W. Zhao, L. Ye, J.L. Wang, et al., No reference objective stereo video quality assessment based on visual attention and edge difference, IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2015, pp. 523–526.
[9] Y. Han, Z. Yuan, G.M. Muntean, No reference objective quality metric for stereoscopic 3D video, 2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Beijing, China, 2014, http://dx.doi.org/10.1109/BMSB.2014.6873539.
[10] Y. Han, Z. Yuan, G.M. Muntean, An innovative no reference metric for real-time 3D stereoscopic video quality assessment, IEEE Trans. Broadcast. 62 (3) (2016) 654–663.
[11] X. Xia, Z. Lu, L. Wang, et al., Blind video quality assessment using natural video spatio-temporal statistics, IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 2014, pp. 1–6.
[12] A. Mittal, A.K. Moorthy, A.C. Bovik, Making image quality assessment robust, The Forty Sixth Asilomar Conference on Signals, Systems and Computers, California, USA, 2012, pp. 1718–1722.
[13] L. Liu, B. Liu, H. Huang, et al., No reference image quality assessment based on spatial and spectral entropies, Signal Process. Image Commun. 29 (8) (2014) 856–863.
[14] M.A. Saad, A.C. Bovik, Blind quality assessment of videos using a model of natural scene statistics and motion coherency, The Forty Sixth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2012, pp. 332–336.
[15] R. Soundararajan, A.C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2012) 684–694.
[16] F. Cong, Q. Lin, L. Kuang, et al., Tensor decomposition of EEG signals: a brief review, J. Neurosci. Methods 248 (15) (2015) 59–69.
[17] G. Cui, L. Gu, Q. Zhao, et al., Bayesian CP factorization of incomplete tensor for EEG signal application, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vancouver, BC, Canada, 2016, pp. 2170–2173.
[18] O. Kaya, B. Uçar, High performance parallel algorithms for the tucker decomposition of sparse tensors, 2016 the 45th International Conference on Parallel Processing (ICPP), Philadelphia, PA, USA, 2016, pp. 103–112.
[19] L. Qiao, B. Zhang, L. Zhuang, et al., An efficient algorithm for tensor principal component analysis via proximal linearized alternating direction method of multipliers, International Conference on Advanced Cloud and Big Data (CBD), Chengdu, China, 2016, pp. 283–288.
[20] J. Zhang, C. Xu, P. Jing, et al., A tensor-driven temporal correlation model for video sequence classification, IEEE Signal Process Lett. 23 (9) (2016) 1246–1249.
[21] M. Lee, C. Choi, Incremental N-mode SVD for large-scale multilinear generative models, IEEE Trans. Image Process. 23 (10) (2014) 4255–4269.
[22] A. Mittal, A.K. Moorthy, A.C. Bovik, Blind/referenceless image spatial quality evaluator, The Forty Fifth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2011, pp. 723–727.

[23] M. Urvoy, M. Barkowsky, R. Cousseau, et al., NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences, The Fourth International Workshop on Quality of Multimedia Experience, Yarra Valley, VIC, Australia, 2012, pp. 109–114.

[24] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[25] M. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.