Cogn Comput (2010) 2:120–131
DOI 10.1007/s12559-010-9040-9

# A Video Quality Assessment Metric Based on Human Visual System

**Wen Lu · Xuelong Li · Xinbo Gao · Wenjian Tang · Jing Li · Dacheng Tao**

Published online: 5 May 2010


**Abstract** It is important for practical application to design an effective and efficient metric for video quality. The most reliable way is by subjective evaluation. Thus, to design an objective metric by simulating *human visual system* (HVS) is quite reasonable and available. In this paper, the video quality assessment metric based on visual perception is proposed. Three-dimensional wavelet is utilized to decompose video and then extract features to mimic the multichannel structure of HVS. *Spatio-temporal contrast sensitivity function* (S-T CSF) is employed to weight coefficient obtained by three-dimensional wavelet to simulate nonlinearity feature of the human eyes. Perceptual threshold is exploited to obtain visual sensitive coefficients after S-T CSF filtered. Visual sensitive coefficients are normalized representation and then visual sensitive errors are calculated between reference and distorted video. Finally, temporal perceptual mechanism is applied to count values of video quality for reducing computational cost. Experimental results prove the proposed method outperforms the most existing methods and is comparable to LHS and PVQM.

**Keywords** Video quality assessment · Human visual system · Three-dimensional wavelet · Contrast sensitivity · Temporal perceptual mechanism

W. Lu · X. Gao · W. Tang · J. Li
School of Electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, People's Republic of China
e-mail: luwen.xidian@gmail.com

X. Gao
e-mail: xbgao.xidian@gmail.com

W. Tang
e-mail: tangwenjian1987@gmail.com

J. Li
e-mail: rain.lee.922@gmail.com

X. Li (✉)
Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, People's Republic of China
e-mail: xuelong_li@opt.ac.cn

D. Tao
School of Computer Engineering, Nanyang Technological University, Singapore, Singapore
e-mail: dacheng.tao@gmail.com

## Introduction

Video quality assessment is playing a key role in visual information processing [1, 2]. With the development of information technology, the digital videos are exploited in many practical applications and play an increasingly important part in our daily life, such as video conference, video on demand, and video surveillance, etc. However, during acquisition, compression, processing, transmission, and reproduction, the video data is subject to various kinds of distortions such as mosquito noise, blocking artifacts, color bleeding, false contouring, and packet losses, which will degrade the quality of the digital video and result in the complaint of the customers, especially in modern video coding [4–7]. Thus, there is demand for a video system to control and quantify the circumstances of the video quality degradations. Moreover, to adjust and enhance the quality of the video by objective methods is desirable. Hence, an effective and efficient video quality metric [1] is of great urgency for this target.

Considering that human eyes are the ultimate receivers of a video sequence, the best way to evaluate the quality of

a video sequence is by subjective experiment. It requires a number of observers to watch series of video sequences and score them. However, according to the ITU-R BT500.11 [2], the requirements for subjective experiment are extremely strict, to realize it will face many problems such as the selection of viewers, the expensive cost, and most of all, the difficulty in real application. To resolve this dilemma, researchers have developed a series of objective quality assessment methods [3, 8], which aim to evaluate distortions of video sequences automatically and accurately.

The ultimate goal of the objective video quality metrics is trying to emulate the process of *human visual system* (HVS). According to the philosophy that the metrics are constructed, the video quality metrics can be classified as "bottom-up" approach and "top–down" approach [3]. "Bottom-up" means to emulate the quality assessment metrics from the functionality of each component in the HVS. For example, the multichannel decomposition, the temporal mechanism, contrast sensitivity function, pattern gain control, and spatio-temporal pooling. For the "top–down" approach, it is more like a "black box", what we are concerned with is the input and the corresponding output of the box rather than the underlying mechanism. So, the top–down approaches are trying to emulate the overall functionality of HVS. However, it is difficult to design this method because of the complexity of HVS. Theoretically, to understand and emulate the component functionality of HVS is a rational, feasible, and effective approach to gauge video quality.

The traditional PSNR [9] is the most widely used video quality assessment metric. However, it doesn't provide a satisfactory reflection of the human visual perception. In recent years, most of the researchers devoted to research the "bottom-up" methods. Sarnoff corporation designed a perceptual quality metric based on *just noticeable difference* (JND) in 2000 [10]. Winkler [11] proposed a method called the *perceptual distortion metric* (PDM), which employs the steerable pyramid transformation to emulate the multichannel mechanism. Watson proposed the *digital video quality* (DVQ) [12], which is implemented in the DCT domain. This model provided a method for resolving the dilemma of accuracy of perceptual prediction and huge computational cost. Moreover, some researchers preferred to the "top–down" methods for their simple computation and feasibility of practical application. Hoekstra's *perceptual video quality measure* (PVQM) [13] integrates the marginal characteristic of the luminance image, normalized color error and temporal change to evaluate the video quality. In 1999, Wolf and Pinson developed the *video quality metric* (VQM) [14] by integrating the spatial features, color features, local contrast features, and motion information. Wang [15] extended the SSIM to video quality assessment by weighing the SSIM in each color channel and frame according to the local luminance and motion information. Mei [16] proposed a novel spatio-temporal quality assessment scheme in terms of low-level content features for home videos. In addition, some researchers have developed a number of the weighting functions based on visual attention, which can be applied to the existing *video quality assessment* (VQA) metrics. Lu [17] et al. proposed a visual attention's modulator after-effects model by integrating the local perceptual stimuli from color contrast, texture contrast, motion, as well as cognitive features. Li [18] developed a statistical weighting model based on visual speed perception. Both of the weighting models have achieved better performance when applied to PSNR or SSIM than the original ones. Indubitably, these metrics have shown higher consistency than the traditional PSNR. But there are some limitations: (1) As the video is a 3-D signal, all of the metrics process the 3-D information either in temporal filtering or motion information, none of them has really fully utilized the 3-D features; (2) Most of the methods still need huge computational cost, especially in multichannel decomposition and motion estimation; (3) Most of the methods implement the spatio-temporal pooling by averaging or Minkowski pooling, however, both of the methods are too simple to reflect the visual perception on video quality assessment. So, there is necessity for researchers to develop a model on visual perception that can deal with all of the issues above.

In this paper, to extend the application of visual perception, to further improve the performance of video quality assessment, a novel metric of video quality assessment is proposed, which is pooling the *three–dimensional wavelet* (3-D Wavelet) for simulating multichannel characteristic of HVS, the *Spatio-temporal contrast sensitivity function* (S-T CSF) to reflect a fact that human eyes have different sensitivity to contrast with different spatiotemporal frequency, the temporal mechanism is employed to describe the phenomenon that how to perceive the quality of all video frames when watching them frame by frame. According to above description, it can design a method by simulating human visual system to capture the variation of visual perception. In this framework, three-dimensional wavelet is utilized to decompose video and then extract features, and then S-T CSF is employed to weight coefficient obtained by 3-D Wavelet to simulate nonlinearity feature of the human eyes. After that, perceptual threshold is exploited to obtain visual sensitive coefficients after sptio-temporal contrast sensitivity filtered, then visual sensitive coefficients are normalized representation and then visual sensitive error coefficients are calculated between reference and distorted video. Finally, temporal perceptual mechanism is applied to count values of the video quality for reducing computational cost.

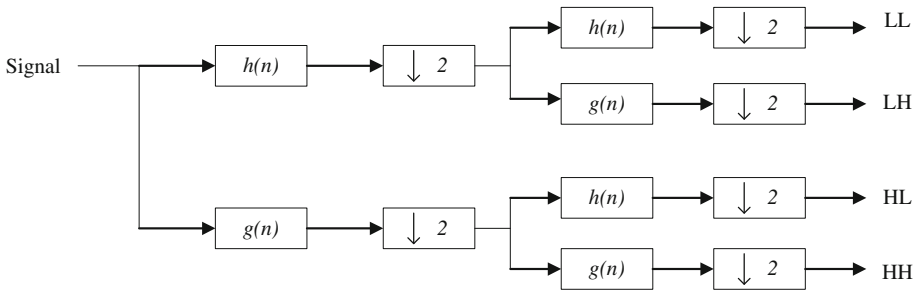


**Fig. 1** The flowchart of 2-D Wavelet decomposition

The rest of this paper is organized as follows: in Section "The Foundation of the Proposed Method", 3-D wavelet for multichannel structure of HVS, spatio-temporal contrast sensitivity function, and temporal perceptual mechanism are provided for better understanding. Section "The Proposed Methods" presents the 3-D wavelet decomposition, S-T CSF, perceptual threshold, normalized representation, and error pooling as proposed methods. Experimental results are presented in Section "Performance Evaluation", and Section "Conclusion" provides the overall summary.

## The Foundation of the Proposed Method

### 3-D Wavelet for Multichannel Structure of HVS

Wavelet transform is a local transformation of spatial and temporal frequencies and able to extract the information from the signal effectively. The multidimensional and multifrequency features make wavelet transform having the function of local analyzing and extracting the feature from the signal effectively, so it named as "mathematical microscope". Figure 1 shows the decomposition of the 2-D wavelet, where $h$ is a low-pass filter and $g$ is a high-pass filter.

In Fig. 2, Lena is decomposed by 2-D Wavelet transform. In the top-left is the low-frequency subband, which includes the whole structural information of the image. And the other six-ones $LH_x$, $HL_x$, and $HH_x$ subbands (the $x$ presents 1 or 2) are high-frequency subbands, corresponding to the specific information of the image. It is observed that the wavelet transform is well adapted to approximate the multichannel structure of HVS and the logarithmic characteristics of frequency perception. Meanwhile, it can eliminate the statistical and the visual redundancy, which is in accordance with the characteristic of the sparse representation in HVS.

3-D wavelet transform is the extension of 2-D wavelet transforms [20–22], equal to 1-D wavelet transform filtering along the three directions of the 3-D signal. Figure 3 shows 3-D wavelet transform-based image decomposition. For video sequences, 2-D wavelet transform is first applied to every frame, and then temporal decomposition is implemented along the time axis. Figure 3 shows the decomposition diagram by 3-D wavelet transform. After one level wavelet transform, it will generate one low-frequency subband ($LLL_1$) and seven high-frequency subbands ($LLH_1$, $HLL_1$, $HLH_1$, $LHL_1$, $LHH_1$, $HHL_1$, and $HHH_1$). If we continue to decompose the low-frequency subband, the low- and high-frequency 3-D subbands in different scales will be obtained.

Human visual system is a very complex system and not fully understood yet. A large number of physiological and psychophysical experiments show physiological evidence and phenomenon. Some neurons in the primary visual cortex are tuned to visual stimuli with specific frequencies, spatial locations, and orientations. This phenomenon is called the multichannel characteristic of HVS [19]. In signal processing, wavelet is considered as a filter to mimic

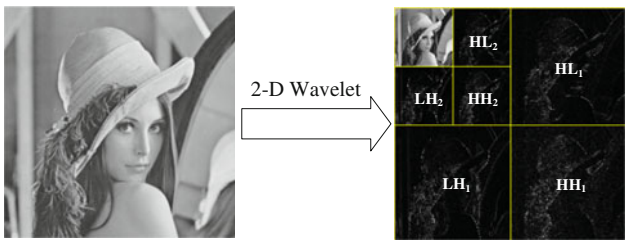


**Fig. 2** 2-D wavelet transform-based image decomposition

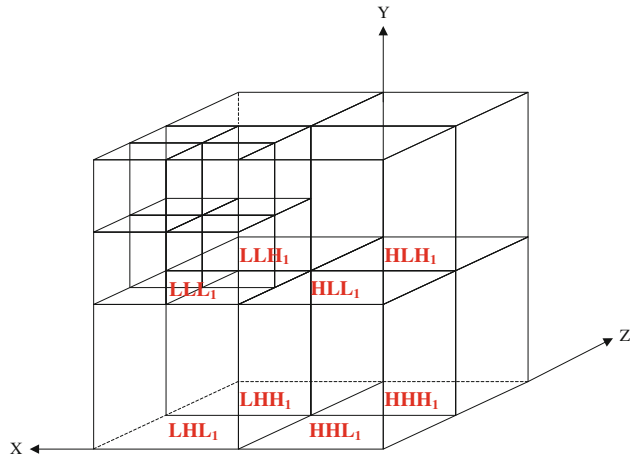


**Fig. 3** A schematic plot of 3-D wavelet decomposition

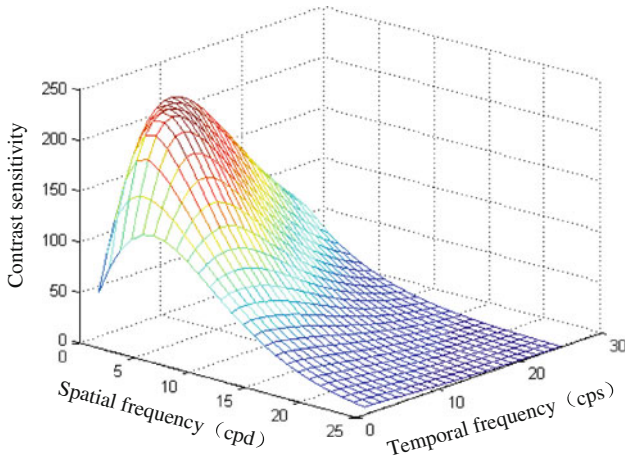


**Fig. 4** The surface of an S-T CSF

multichannel structure of HVS. Since motion is the difference between videos and still images, and the motion is described to use the three-dimensional character of videos. Thus, *three-dimensional wavelet* (3-D Wavelet) is employed to extract the visual features of videos and emulate the multichannel characteristic of HVS.

### Spatio-Temporal Contrast Sensitivity Function

According to the psychophysical experiments, HVS is sensitive to the contrast and can only detect a signal whose contrast is above a certain threshold with respect to the spatial and temporal frequencies of stimuli [23]. The reciprocal of the threshold is called contrast sensitivity. In the circumstances of viewing video sequences, *spatio-temporal contrast sensitivity function* (S-T CSF) is sometimes referred to as the spatial acuity of HVS depending on the velocity of the image traveling across the retina, where the retinal image velocity implicitly denotes the temporal frequency [24]. The S-T CSF is shown in Fig. 4. The horizontal coordinates represents the spatial and temporal frequency, with the unit of *cycles per degree* (cpd) and *cycles per second* (cps). The vertical coordinate represent the contrast sensitivity. The S-T CSF is utilized in this paper as weight for the coefficients by 3-D wavelet transform.

### Temporal Perceptual Mechanism

Temporal perceptual mechanism explains the temporal pooling effects when human eyes watch the video frame by frame. It contains the short-term and long-term mechanisms [25]. Due to short-term human memory [26, 27], the influence of a strong stimulus persists for a short while and fades out gradually. When two stimuli occur within an interval shorter than the memory duration, responses to these two stimuli may merge. However, the merging effect is limited to masking of undistorted frames by impaired ones but not vice versa. This is called "smoothing effect". As the long-term mechanism is too complex to emulate precisely, most of researchers simplify the temporal pooling model just by short-term mechanism [12, 28–30]. In video processing, a series of video frame includes some distorted and undistorted frames, but the people would subjectively believe that all video frames are distorted. This phenomenon is called short-term mechanism. Due to the existence of this phenomenon, undistorted frames will be masked by distorted frames, that is, the distorted frames are dominated in the human perception. In this paper, we use the short-term mechanism for less computational cost. In this paper, short-term mechanism is utilized to reduce computational cost.

## The Proposed Methods

The essence of VQA is to calculate the difference of certain features between the reference and the distorted videos according to visual perception. Thus, how to extract and

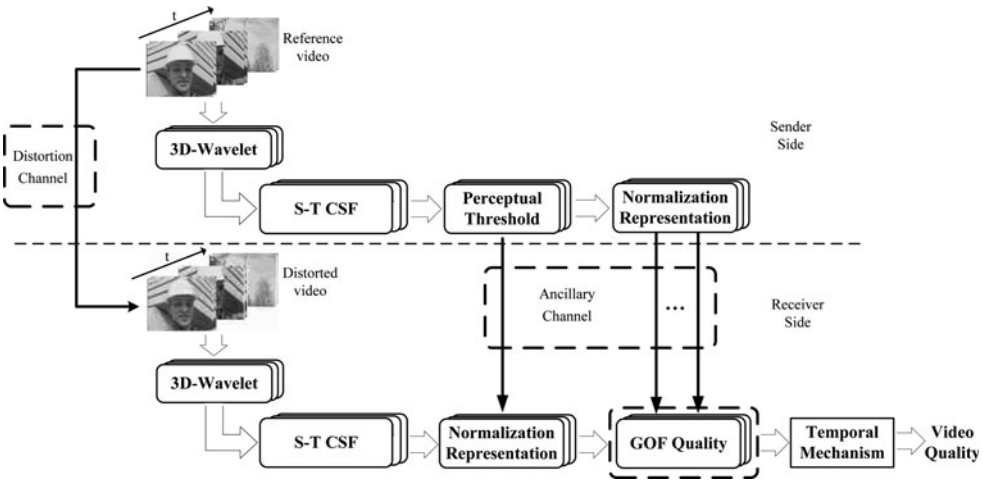


**Fig. 5** Visual perception-based video quality assessment scheme (sender side is applied to extract the normalized representation of the reference video. Receiver side is applied to extract the normalized representation of the distorted video. Ancillary channel is applied to transmit the extracted normalized histogram.)

process these features effectively are the keys of VQA. In this paper, 3-D wavelet transform is utilized to represent the content of the videos with different spatial frequency, temporal frequency and orientations, which is in accordance with the multichannel characteristics of human visual system. 3-D wavelet is applied to decompose *Group of Frames* [20, [31] (GOF) into three levels, and then the S-T CSF is applied as weight scheme to the corresponding coefficients obtained by 3-D wavelet. A perceptual threshold can be computed according to the characteristics of human visual perception and employed to each GOF for normalization histogram to obtain the varying of the visual sensitive coefficients in each sub-band. At last, the predicted video quality can be calculated by pooling all the qualities of the GOFs according to the short-term mechanisms of human visual system. Figure 5 shows the block diagram for our proposed metric.

### 3-D Wavelet Decomposition

In the proposed method, both of the reference and distorted videos are divided into a number of GOFs, and each GOF includes eight frames. For representing the content of videos and extracting feature information, 3-D wavelet is utilized to decompose each GOF into three levels, and then 21 high and 1 low frequency subbands are obtained, as shown in Fig. 6. In this method, we chose Daubechies "db1" as the filter of the wavelet transformation. After the decomposition, all the subbands are processed as follows.

### S-T CSF

After the 3-D wavelet decomposition, each GOF is decomposed into 22 different spatio-temporal subbands. However, the visual perceptual sensitivity to signals with different spatio-temporal frequency is different. In order to emulate this psychophysical phenomenon and make the coefficient with the same sensitivity for human eyes, the coefficients in different subbands should be weighted by S-T CSF, which is also a preparation for the next step.

In the literature [23], Kelly proposed an S-T CSF model, it can be expressed as

$$C(\alpha,\mu) = \left(6.1 + 7.3|\log(\mu/3)|^3\right)\mu\alpha^2 \times \exp(-2\alpha(\mu+2)/45.9) \quad (1)$$

where $\alpha = 2f$, $\mu = 2\pi f_t/\alpha$, and $f = \sqrt{f_x^2 + f_y^2}$ is the spatial frequency with units of cycles/degree. And $f_t$ is the temporal frequency with units of cycles/second. $f_x$ and $f_y$ are horizontal and vertical spatial frequency, respectively. We can normalize the spatial frequency by

$$f(\text{cycles/degree}) = f_n(\text{cycles/pixel}) \times f_s(\text{pixels/degree}) \quad (2)$$

where $f_n$ is the spatial frequency after normalizing with units of cycles/pixel. $f_s$ is the sampling frequency with units of pixels/degree, it is defined as

$$f_s = \frac{2v\tan(0.5^\circ)r}{0.0254} \quad (3)$$

where $v$ is the viewing distance with units of meter and $r$ is the resolution power of the display with units of pixels/inch. In this method, $v$ is 0.8 m and $r$ is 61 pixels/inch. If the sampling frequency of videos meets the Nyquist sampling theorem, then $f_n$ is 0.5, so $f$ changes from 0 to $f_s/2$.

The temporal frequency of the $k$-th frame $f_t(k)$ is defined as [31]

$$f_t(k) = \frac{k}{2m} \times f_r \quad 0 \le k \le m-1 \quad (4)$$

where $m$ is the total number of the frames, $f_r$ is the frame rate. We set $m$ as 8 in this paper.

### Perceptual Threshold

According to the characteristics of the HVS [19], the human eyes are more sensitive to the coefficients with larger magnitude in subbands. As shown in the Fig. 7a, the original "barbara" contains 348160 coefficients in subbands, but a image without any appreciable distortion can be reconstructed only by maximal 100,000 coefficients in subbands, as shown in Fig. 7b. Figure 7c and 7d show the

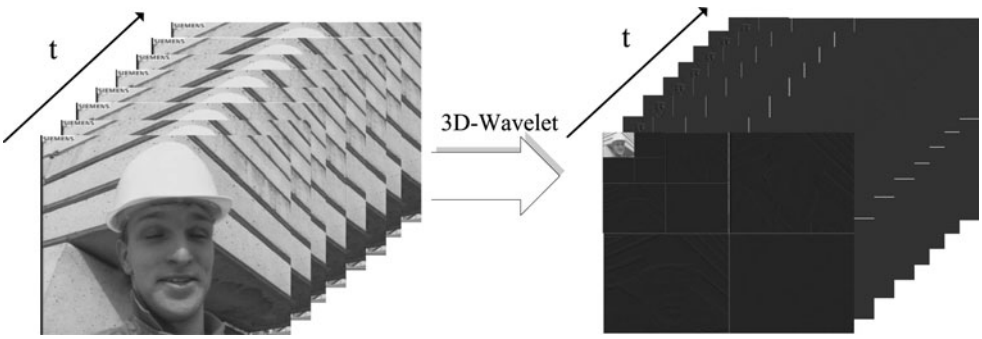


**Fig. 6** 3-D wavelet transform-based GOF decomposition


**Fig. 7** Barbara image decayed with the decrease in transform coefficient. **a** The original image with 348160 coefficients in subbands; **b** Reconstructed image with maximal 100,000 coefficients in subbands; **c** Reconstructed image with maximal 10,000 coefficients in subbands; **d** Reconstructed image with maximal 3,000 coefficients in subbands

reconstructed images with maximal 10,000 coefficients and 3,000 ones, respectively.

In the proposed method, the GOF is decomposed into several subbands, which have different temporal and spatial low–high frequencies. After being filtered by S-T CSF, all the coefficients have similar perceptual importance in different frequency subbands. Then perceptual threshold is used to remove visually insensitive coefficients, the rest of coefficients are called visual sensitive coefficients which reflects the visual quality of the reconstructed GOF. Here, the perceptual threshold is defined as

$$T = \sum_{i=1}^{3} \omega_i \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{i,j} - \bar{x}_i\right)^2} \tag{5}$$

where $x_{i,j}$ is the $j$-th coefficient of the $i$-th subband in the finest scale and $\bar{x}_i$ is the mean value of the $i$-th subband coefficients, $N_i$ is the amount of coefficients of the $i$-th subband, and $\omega_i$ is the weight factor of the $i$-th subband. Figure 8 presents the diagram of a GOF which is decomposed into three levels by 3-D wavelet. We choose the three gray subbands to compute the perceptual threshold as shown in Fig. 8. Not only could this reduce the calculated amount, it also comprises information of temporal and spatial frequencies, and the three subbands selected are symmetrical distributed, which makes the calculation of the perceptual threshold more reasonable.

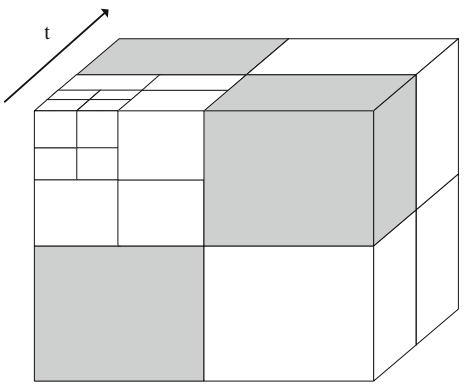


**Fig. 8** An example of GOF decomposition

## Normalized Representation

By using threshold, the number of visual sensitive coefficients in the $n$-th selected subband can be counted, which is defined as $C_T(n)$. It means the number of coefficients in the

$n$-th selected subband that are larger than $T$ obtained from Eq. 5. The number of coefficients in the $n$-th selected subband is therefore, for a given GOF, we can obtain the normalized histogram with $L$ bins ($L$ subbands are selected) for representation and the $n$-th entry is given by

$$P(n) = \frac{C_T(T)}{C(n)}. \tag{6}$$

Error Pooling

According to Eq. 5, we can obtain the normalized histogram for reference and distorted GOF as $P_R(n)$ and $P_{\mathrm{D}}(n)$, respectively. The distorted GOF quality can be calculated by formulae 3–7

$$Q = \frac{1}{1 + \log_2\left(\frac{S}{Q_0} + 1\right)} \tag{7}$$

where $S = \sum_{}^{L} |P_R(n) - P_D(n)|$ is the city-block distance between $P_R(n)$ and $P_{\mathrm{D}}(n)$, and $Q_0$ is a constant used to control the scale of the distortion measure. In this framework, we set $Q_0$ as 0.1.

Video Quality Considering Temporal Perceptual Mechanism

After processing all GOF, a set of GOF qualities are obtained. Due to short-term human memory [26, 27], a strong stimulation could persist for a short time, and fade out gradually. In terms of VQA, when short durations of unimpaired frames interleave with distorted ones, all the frames appear distorted subjectively. So, when viewers watching the videos, unimpaired frames will be masked by the distorted ones because of the phenomenon, it means the distorted frames are much more important than the unimpaired ones, and the process is irreversible.

In order to emulate the physiology and psychophysical characteristics of visual perception, we process all the GOF qualities as follows

$$Q'_{i+1} = \begin{cases} Q_i, & \text{if} \quad Q_{i+1} - Q_i > D \\ Q_{i+1}, & \text{others} \end{cases} \tag{8}$$

where $Q_i$ is the $i$-th GOF quality before processing, $Q'_i$ is the $i$-th GOF quality after processing, $D$ is a perceptual threshold, we set $D$ as 0.1 in the paper, which is a empirical value. If the current GOF quality is higher than the previous one and the variation is above the threshold $D$, then we chose the previous GOF quality as the current one, as shown in Eq. 8. In other words, the bad GOF could mask the good GOF as a visual stimulation before fading out. At last, we define the metrics of the distorted video quality as

$$\mathrm{SCORE} = \mathrm{median}\{Q'_1, Q'_2, Q'_3 \ldots Q'_n\} \tag{9}$$

where $n$ is the number of the GOF in the video.

## Experimental Results

In this section, we design three experiments to evaluate the performance of the proposed method: the consistency experiment, the sensitivity experiment and the rationality experiment.

The Consistency Experiment

This experiment is based on the full set of *video quality expert group* (VQEG) [14] 625/50-Hz test sequences, with the resolution of 720 × 576 and 4:2:2 sampling format, the number of a test sequence frame is 200. The 625/50 Hz database includes 10 original sequences and each of them is evaluated under 16 (HRC1-HRC16) conditions. VQEG [14] suggested using three criteria as evaluation criteria: the *Pearson linear correlation coefficient* (CC), the *Spearman rank-order correlation coefficient* (ROCC), and the *outlier ratio* (OR).

The performance of the proposed method is compared with some classical VQA methods shown in Table 1. It should be noted that the proposed is a reference reduced approach, which uses only a threshold value and 21 features of high frequency subbands decomposed by 3-D wavelet from a GOF. So, the transmitted data rate is (21 + 1):(720 × 576 × 8) = 1:138,240. From Table 1 we can see, the results of PVQM and LHS are better than those of the proposed. However, PVQM is a *full reference* (FR) metric, so it needs all data (1:1) of the video to evaluate the video quality, all data of video are difficult to obtain in practical application such as video communication. LHS is a *reduced reference* (RR) method, but the data need to

**Table 1** The performance of PSNR, JND, PDM, DVQ, PVQM, SSVM, PM, VQM, LHS, and the proposed on the VQEG 625/50 Hz database

| Method | Type | Data rate | CC | ROCC | OR |
|---|---|---|---|---|---|
| PSNR [9] | FR | 1:1 | 0.826 | 0.810 | 0.728 |
| JND [10] | FR | 1:1 | 0.759 | 0.753 | 0.750 |
| PDM [11] | FR | 1:1 | 0.684 | 0.718 | 0.689 |
| DVQ [12] | FR | 1:1 | 0.780 | 0.771 | 0.633 |
| PVQM [13] | FR | 1:1 | 0.864 | 0.866 | 0.594 |
| SSVM [15] | FR | 1:1 | 0.842 | 0.814 | 0.693 |
| PM [32] | FR | 1:1 | 0.840 | 0.830 | – |
| VQM [14] | RR | 1:64 | 0.760 | 0.785 | 0.767 |
| LHS [33] | RR | 1:1024 | 0.850 | 0.860 | – |
| Proposed | RR | 1:138240 | 0.844 | 0.778 | 0.624 |

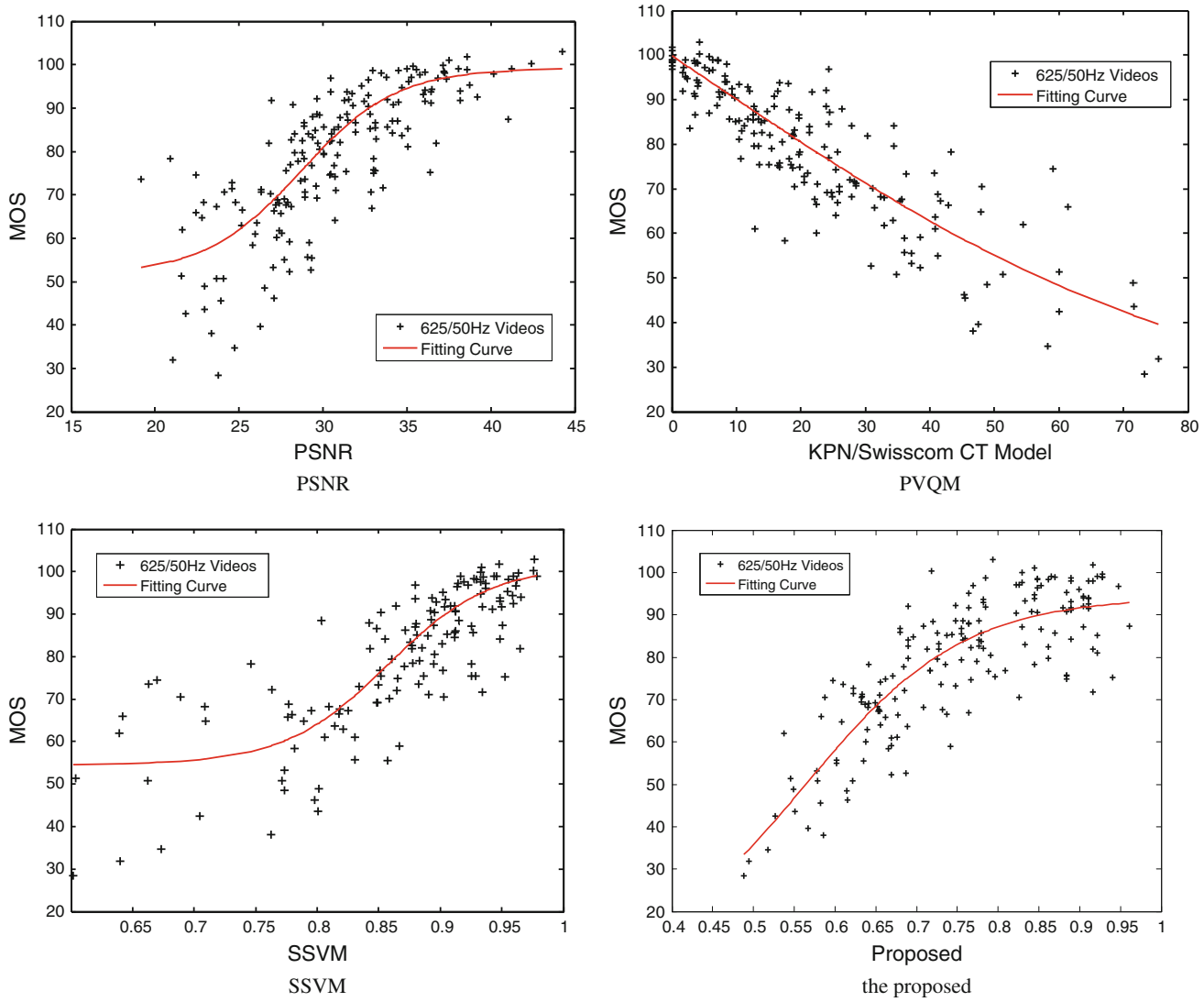


**Fig. 9** Scatter plots of MOS versus different VQA methods on the VQEG 625/50 Hz database

transmit are still much more than ours (1:1,024), which limits its application. In a word, the proposed method has greater room improved than the existing methods, so the objective assessment results have good consistency with subjective perception values. Moreover, it is a method that can meet the requirements of real-time application. Figure 9 presents the scatter plots of MOS versus the predicted score by PSNR, PVQM, SSVM, and the proposed after the nonlinear mapping.

## The Sensitivity Experiment

In this subsection, we build a series of sequences by adding different spatial frequency noise to the standard test sequence "Foreman", and test them by PSNR which is one of the most popular objective quality metrics and the proposed method, respectively. Figure 10 shows three kinds of these noise sequences frames, from observer's view, these sequences have very different levels of distortion, while the PSNR is the same relative to the standard test sequence. Table 2 shows the objective assessment results of the proposed method, PSNR and MSE. It is obvious that not only the objective assessment results of the proposed method have good consistency with subjective perception values but also can well capture the visual perceptual variation of different distorted videos.

## The Rationality Experiment

To verify the rationality of the proposed method, we imitate three distorted videos, the first is the distortions caused by frame dropping, the second is the distortions caused by MPEG-4, and the third is the distortions caused by H.264.

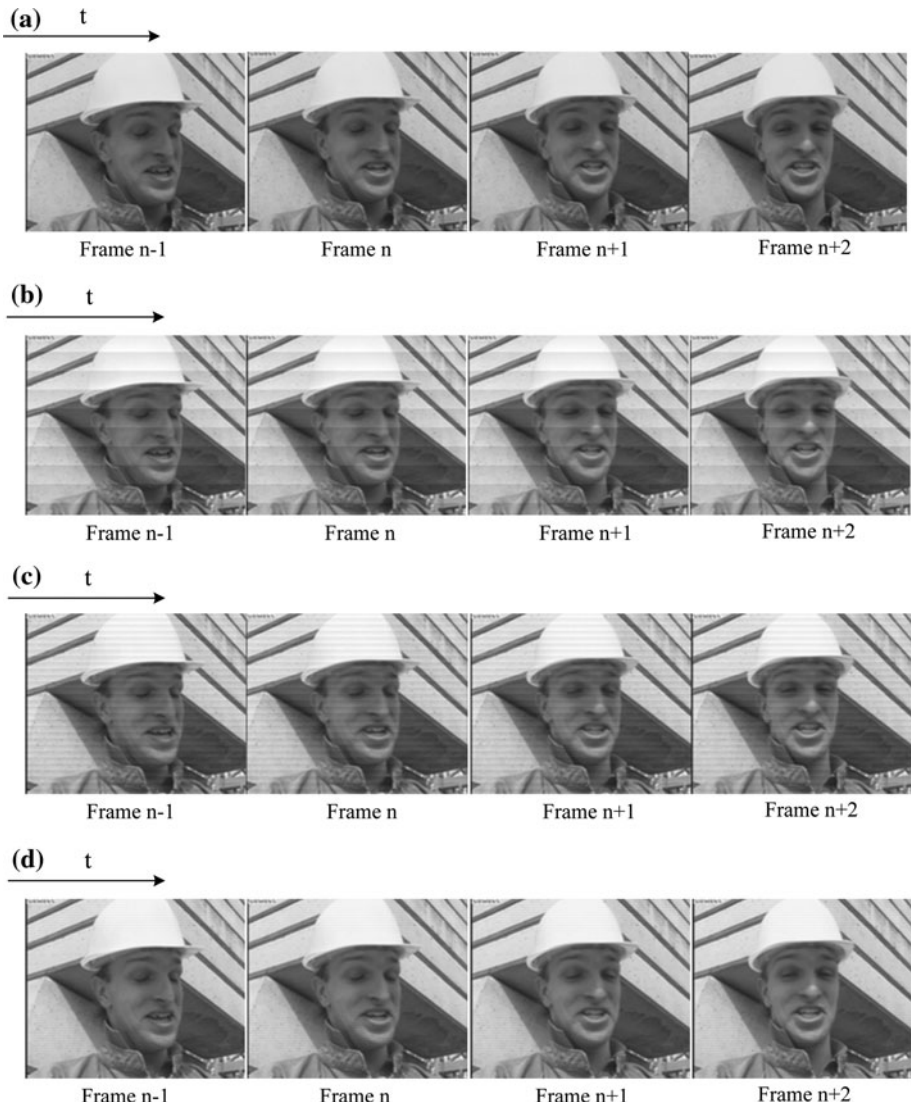


**Fig. 10** The sequences with the same PSNR but different perceived quality. **a** Shows the original sequence; **b**, **c** and **d** show the distorted sequences added different noise with increasing spatial frequency

**Table 2** The values of MSE, PSNR, and the proposed method in Fig. 10

| Method | (b) | (c) | (d) |
|---|---|---|---|
| MSE | 125.78 | 125.78 | 125.78 |
| PSNR | 27.13 | 27.13 | 27.13 |
| Proposed | 0.71 | 0.85 | 0.97 |

### *The Test of the Frame Dropping*

Frame dropping is widely occurred during the transmission. As we known, the more number of the frame dropping is, the worse visual quality of the video is. In this experiment, we build a series of sequences by dropping $N$ frames from the test sequences randomly, and replace the dropped frames with the previous frame, then test these sequences by the proposed method, as shown in Fig. 11. It

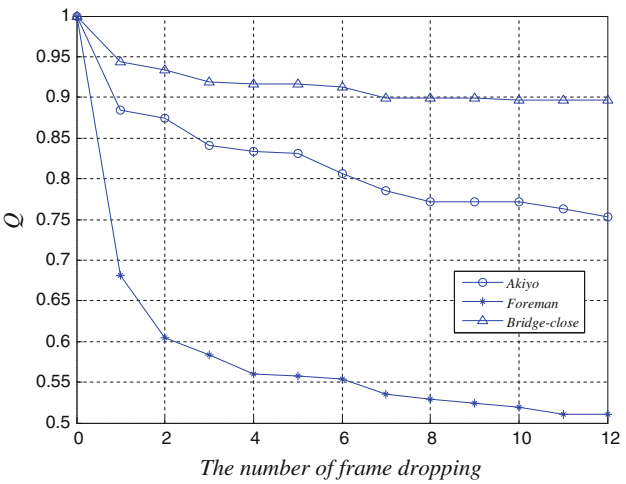


**Fig. 11** Trend plots with different number of frame dropping using the proposed method

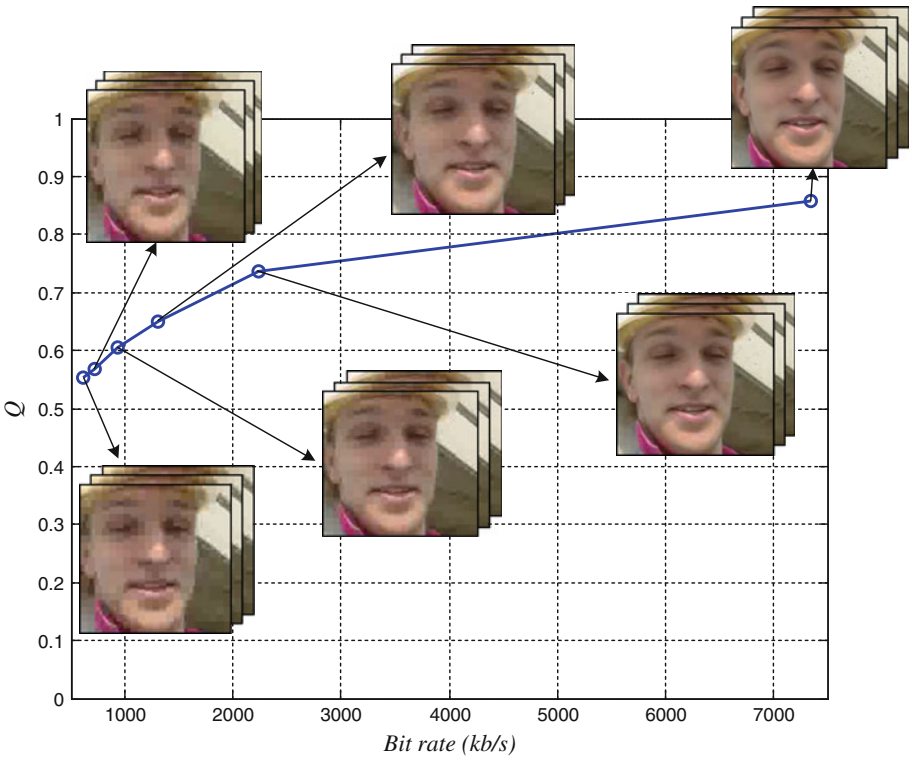


**Fig. 12** Trend plots with MPEG-4 compression at different bit rate using the proposed method

is found that the proposed method predict the tendency of video drop with the number of the frame dropping.

It is quite clear that distortion caused by frame dropping is context sensitive. The sequence "Foreman" has high-motion content, when the number of the frame dropping is excessive, the distortion will be very serious. While the sequence "Akiyo" has less motion information, the degree of distortion changes less when the number of the frame dropping increases. For the sequence "Bridge-close", human eye is not sensitive to the distortion caused by frame dropping. Figure 11 reflects this phenomenon correctly.

*The Test on MPEG-4 Videos*

In this experiment, we test a series of "Foreman" sequences compressed by MPEG-4 with different bit rates,

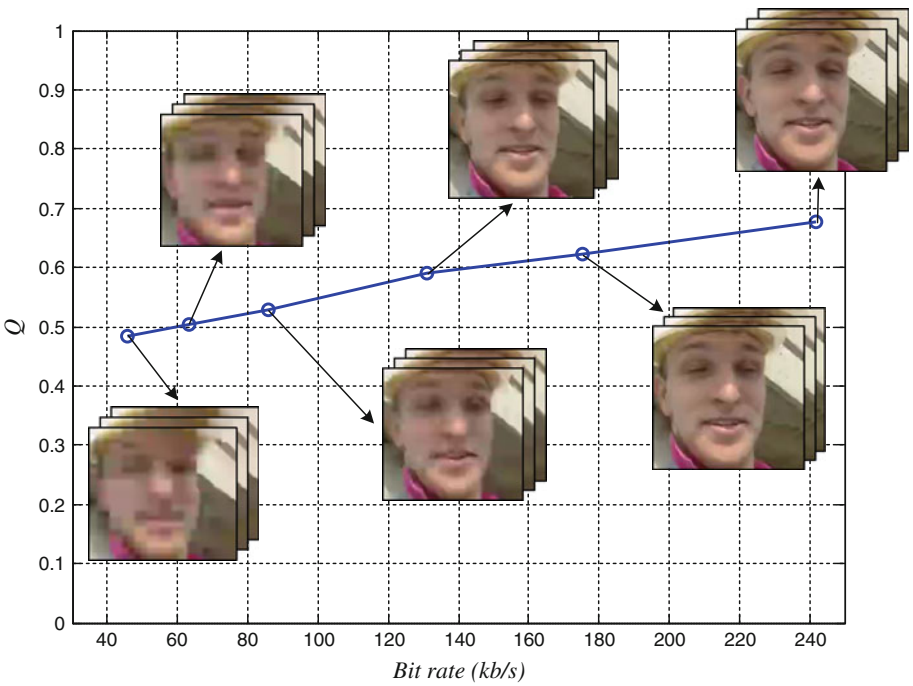


**Fig. 13** Trend plots with H.264 compression at different bit rate using the proposed method

which range from 618.5 to 7347.5 kb/s. Generally speaking, the higher bit rate is and the better visual quality of the video is. We test these distorted sequences by the proposed method, as shown in Fig. 12, the quality of the video becomes higher and higher as the bit rate increases, so it is proved that the proposed method can reflect the distortion degree of videos correctly.

*The Test on H.264 Videos*

We also built a series of "Foreman" sequences compressed by H.264 with different bit rates, which range from 45.5 to 1624.8 kb/s. We test these distorted videos by the proposed method, as shown in Fig. 13, the quality of the video becomes higher and higher as the bit rate increases, it is proved that the proposed method can well reflect the distorted degree of videos.

In this paper, we proposed a novel method for assessing video quality by mimicking human visual system. The experiments on the VQEG database demonstrate the advantage of our proposed method. (1) Firm effectiveness. The results show that our method performs consistently on different experiment, which outperforms most of the existing method and is comparable to LHS and PVQM. (2) Low data rate. Our proposed is a reduced reference method. The requirement of transmitted data rate is very low, only 1: 138,240. But the most of the existing method are full reference methods, and the transmitted data rate is 1:1. The better reduced reference metric VQM is 1:64 also.

## Conclusion

In this paper, a RR VQA metric based on human visual perception is proposed. It utilizes the 3-D wavelet to decompose video, which emulate the multichannel decomposition characteristic of HVS. Then, S-T CSF is employed to weigh the 3-D wavelet coefficients, making the coefficients have the same sensitivity to human eyes. Perceptual threshold is exploited to count the number of visual sensitive coefficients, then visual sensitive coefficients are normalized representation and then visual sensitive errors are transformed city-block distance between reference and distorted video. At last, the short-term memory is considered in temporal pooling and the video quality is obtained. The proposed method outperforms most of the existing VQA models. And the sensitivity experiment verifies that the objective assessment results can capture the visual perceptual variation of different distorted videos. The rationality experiment results show that the proposed method can reflect the degradation in transmission correctly.

The proposed method has good consistency with subjective perception and can well reflect the visual quality of videos. However, the imitation of HVS is relatively simple. The characteristics of HVS deserve to be further investigated in the future, which may improve the VQA model in both theory and practice. In the next step, different wavelet transforms [34] (un-decimated wavelet transform and dynamic wavelet transform) and different wavelet bases [35] will be introduced to enhance the performance of VQA. Besides, the 3D wavelet transformation can be extended to 3D multiscale geometric transformation [36, 37]. And tensor-based approaches [39, 38] can be utilized to represent images in classification, which can be employed in VQA in the future work.

**Acknowledgements** We want to thank the helpful comments and suggestions from the anonymous reviewers. This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005), the Ph.D. Programs Foundation of Ministry of Education of China (No. 20090203110002), the Natural Science Basic Research Plane in Shaanxi Province of China (2009JM8004), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China.

## References

1. Wang Z, Sheikh RH, Bovik CA. Objective video quality assessment. In: Furht B, Marques O, editors. The handbook of video databases: design and applications. Florida: CRC Press; 2003. p. 1041–78.
2. ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures. 2002.
3. Wang Z, Bovik CA. Modern image quality assessment. New York: Morgan and Claypool Publishing Company; 2006.
4. Yuan Y, Evans A, Monro D. Low complexity separable matching pursuits. IEEE Int Conf Acoust Speech Signal Process. 2004; 3(17–21):725–8.
5. Yuan Y, Monro D. Improved matching pursuits image coding. IEEE Int Conf Acoust Speech Signal Process. 2005;2:201–4.
6. Monro D, Yuan Y. Bases for low complexity matching pursuits image coding. IEEE Int Conf Image Process. 2005;2(11–14): 249–52.
7. Yuan Y, Monro D. 3D wavelet video coding with replicated matching pursuits. IEEE Int Conf Image Process. 2005;1(11–14): 69–72.
8. Li X, Tao D, Gao X, Lu W. A natural image quality evaluation metric. Signal Process. 2009;89(4):548–55.
9. ATIS Technical Report T1.TR.PP.74, Objective video quality measurement using a peak-signal-to-noise-ratio (PSNR) full reference technique. 2001.
10. Sarnoff Corporation, J. Lubin. Sarnoff JND Vision Model. Contribution to IEEE G-2.1.6 compression and processing subcommittee. 1997.
11. Winkler S. A perceptual distortion metric for digital color video. In Proc SPIE, 3644: 175–184, San Jose, CA, Jan. 23–29, 1999.
12. Watson AB, Hu J, McGowan JF. DVQ: a digital video quality metric based on human vision. J Electron Imag. 2001;10(1):20–9.
13. Hekstra AP, Beerends JG, Ledermann D, de Caluwe FE, Kohler S, Koenen RH, Rihs S, Ehrsam M, Schlauss D. PVQM—a

perceptual video quality measure. Signal Process Image Commun. 2002;17(10):781–98.
14. VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I VQEG. 2000. 2000Available: http://www.vqeg.org/.
15. Wang Z, Lu L, Bovik AC. Video quality assessment based on structural distortion measurement. Signal Process Image Commun. 2004;19(2):121–32.
16. Mei T, Hua X-S, Zhu C-Z, Zhou H-Q, Li S. Home video visual quality assessment with spatio-temporal factors. IEEE Trans Circuits Syst Video Technol. 2007;17(6):699–706.
17. Lu Z, Lin W, Yang Xiaokang, Ong EP, Yao S. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. IEEE Trans Image Process. 2009;14(11): 1928–42.
18. Wang Z, Li Q. Video quality assessment using a statistical model of human visual speed perception. J Opt Soc Am A. 2007;24(12): B61–9.
19. Wandell BA. Foundations of vision. Sinauer Associates. 1995.
20. Moyano E, Quiles FJ, Garrido A, Orozco-Barbosa L, Duato J. Efficient 3-D wavelet transform decomposition for video compression. International Workshop on Digital and Computational Video, 118–125, Feb. 2001.
21. Wang C, Ma K-L. A statistical approach to volume data quality assessment. IEEE Trans Vis Comput Graphics. 2008;14(3):590–602.
22. Mallat SG. A theory for multiresolution decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11(7):674–93.
23. Kelly DH. Motion and vision II. Stabilized spatio-temporal surface. J Opt Soc Am. 1979;69(10):1340–9.
24. Jia Y, Lin W, Kassim AA. Estimating just-noticeable distortion for video. IEEE Trans Circuits Syst Video Technol. 2006;16(7):820–9.
25. Narita N. Subjective-evaluation method for quality of coded images. IEEE Trans Broadcasting. 1994;40(1):7–13.
26. Tan KT, Ghanbari M, Pearson DE. An objective measurement tool for MPEG video quality. Signal Process. 1998;70(3):279–94.
27. Ninassi A, Meur OL, Callet PL, Barba D. Considering temporal variations of spatial visual distortions in video quality assessment. IEEE J Sel Topics Signal Process. 2009;3(2):253–65.
28. Winkler S. A perceputal distortion metric for digital color video IV. Proc SPIE Human Vis Electron Imag. 1999;3644:175–84.
29. Van den Branden Lambrecht CJ. A working spatio-temporal model of the human visual system for image restoration and quality assessment applications. Proc IEEE Int Conf Acoust Speech Signal Process. 1996;4:2291–4.
30. Masry M, Hemami SS, Yegnaswamy S. A scalable waveletbased video distortion metric and applications. IEEE Trans Circuits Syst Video Technol. 2006;16(2):260–73.
31. Chou C-H, Chen C-W. A perceptually optimized 3-D subband codec for video communication over wireless channels. IEEE Trans Circuits Syst Video Technol. 1996;6(2):143–56.
32. Ong E, Lin W, Lu Z, Yao S, Etoh M. Visual distortion assessment with emphasis on spatially transitional regions. IEEE Trans Circuits Syst Video Technol. 2004;14(4):559–66.
33. Gunawan IP, Ghanbari M. Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration. IEEE Trans Circuits Syst Video Technol. 2008;18(1):71–83.
34. Pan Q, Zhang L, Dai G, Zhang H. Two de-noising methods by wavelet transform. IEEE Trans Signal Process. 1999;47(12): 3401–6.
35. Zhang L, Paul B, Wu X. Multiscale LMMSE-based image denoising with optimal wavelet selection. IEEE Trans Circuits Syst Video Technol. 2005;15(4):469–81.
36. Gao X, Lu W, Tao D, Li X. Image quality assessment based on multiscale geometric analysis. IEEE Trans Image Process. 2009;18(7):1409–23.
37. Lu W, Zeng K, Tao D, Yuan Y, Gao X. No-reference image quality assessment in contourlet domain. Neurocomputing. 2010;73(1):784–94.
38. Li X, Lin S, Yan S, Xu D. Discriminant locally linear embedding with high-order tensor data. IEEE Trans Syst Man Cybern Part B. 2008;38(2):342–52.
39. Tao D, Li X, Wu X, Maybank SJ. General tensor discriminant analysis and Gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell. 2007;29(10):1700–15.