Perceptual Distortion Metric for Stereo Video Quality Evaluation

ZHONGJIE ZHU, YUER WANG
Ningbo Key Lab. of DSP
Zhejiang Wanli University
No 8, South Qianhu Road, Ningbo
CHINA
Zhongjiezhu@hotmail.com

Abstract:- Stereo video is regarded as an important developing trend of video technology and there is an increasing need to develop efficient and perceptually consistent methods for stereo video quality evaluation in the fields of stereo video signal processing. In this paper, a perceptual metric for stereo video quality evaluation is proposed based on the state-of-the-art physiological and psychological achievements on human visual system (HVS). Several main HVS properties related to stereo video are analyzed and a multi-channel vision model based on 3D wavelet decomposition is proposed. Simulations are performed and experimental results reveal that, compared with the traditional objective metrics such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE), the proposed metric is more perceptually consistent.

Key-Words: -Human visual system, stereo video, image quality evaluation, 3D wavelet decomposition

1 Introduction

Image quality evaluation is a key technology in the field of video signal processing. In most scenarios human being is the final receiver, so perfect image quality evaluation should use human-based subjective methods. However, subjective methods are usually very complex and computationally expensive [1][2]. In most of current applications, the widely used measures for evaluating image quality are based on the objective methods such as MSE and PSNR. Those methods are usually simple and flexible, but they are hard to be consistent with human perception due to their lacking of consideration of HVS characteristics. Therefore, there is an urgent need to develop computable and HVS-consistent objective method that is capable of providing accurate and efficient image quality evaluation. This topic has attracted great interest and many researchers have been devoting themselves to this field. So far, quite a few achievements have been made. In [3], a perceptual distortion metric for the evaluation of video quality is presented. It is based on a model of the human visual system that takes into account color perception, multi-channel architecture of temporal and spatial mechanisms, spatio-temporal contrast sensitivity, pattern masking as well as channel interactions. In [4], a computationally efficient video distortion metric has been proposed, which can operate in full- or reduced-reference mode as required. The metric is based on a model of the human visual system implemented using the wavelet transform and separable filters. In [5], an objective HVS-based no-reference metric has been proposed for video quality assessment for digitally coded videos containing natural scenes. Other newly proposed metrics that incorporated HVS characteristics can refer to [6]-[8]. However, due to the extreme complication of HVS, many of its properties are still not well understood even today. This topic should be further studied.

Stereo video is regarded as an important developing trend of video technology and it has many potential applications in education, entertainment, medical surgery, video conference and so on. However, there still exist many challenges in stereo video technology that should be further studied, one of which is the efficient and accurate stereo video quality evaluation. Current metrics for stereo video quality evaluation utilizes the same traditional methods such as PSNR, MSE and so on, which operate solely on a pixel-by-pixel basis and do not incorporate HVS properties. To the best of our knowledge, so far there is no exclusive perceptual metric that has been proposed for stereo video quality evaluation in literature. It is more difficult to propose a perceptual metric for stereo video than for single view video due to the more complicated HVS properties relevant to stereo video.

In this paper, based on state-of-the-art physiological and psychological achievements on human visual system, a perceptual metric for stereo video quality evaluation is proposed after analyzing the main properties of human visual system relevant to stereo video.

The paper is structured as follows: Several main HVS characteristics related to stereo video are firstly analyzed in section II. Section III describes the whole process of the perceptual metric. Experiments have been done to evaluate the performance of the proposed metric in section IV. Section V concludes the paper.

2 HVS Characteristics Related to Stereo Video

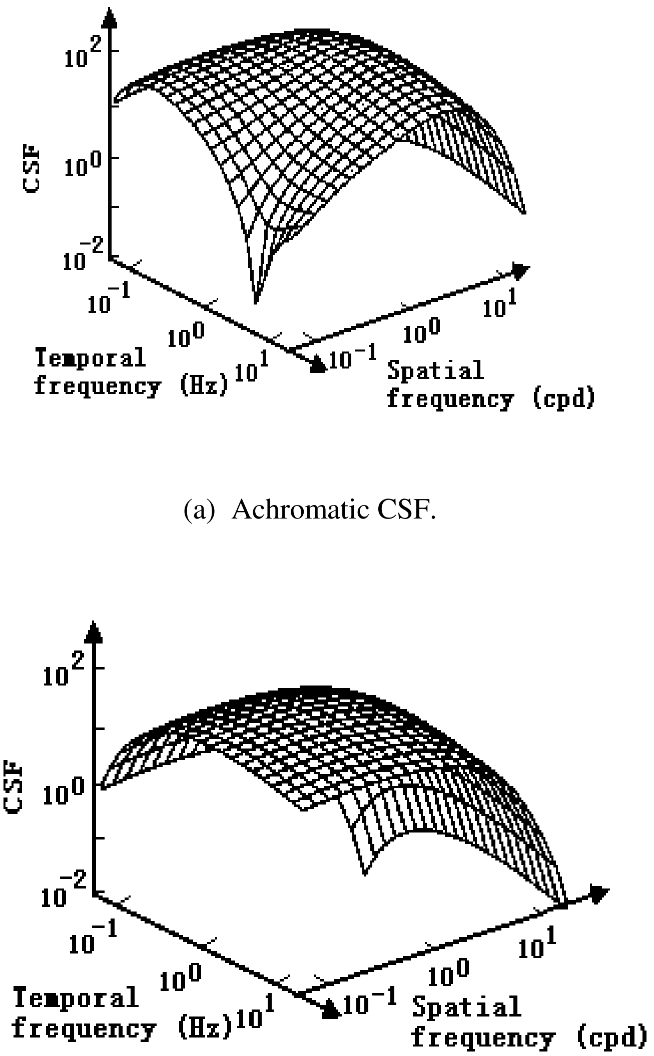
Many scholars are devoted to HVS characteristics and have made great achievements. However, due to the extreme complexity of HVS, to acquire accurate computational model for HVS is very difficult and it is impossible to incorporate the whole HVS features into quality evaluation metric. This section discusses several important phenomena of visual perception that are considered in the metric to be proposed in this work, including Contrast Sensitivity Function, Mmulti-channel and Masking, Depth Perception and so on.

2.1 Contrast Sensitivity Function

The response of the human visual system depends much less on the absolute luminance than on the relation of its local variations to the surrounding luminance [9]. Contrast is a measure of this relative variation of luminance，and there exists a threshold contrast for an observer to detect a change in intensity. The inverse of the contrast threshold is usually defined as contrast sensitivity which mostly depends on the stimulus characteristics such as color, spatial and temporal frequency. Fig. 1 shows an achromatic and a chromatic Contrast Sensitivity Functions (CSF) given by Burbeck and Kelly [10]-[11], from which it can be observed the achromatic contrast sensitivity is generally higher than chromatic, but it decreases at low spatial and temporal frequencies. CSF is one of the most important HVS properties considered by most existing perceptual video quality evaluation metrics.

2.2 Masking

Masking effect can be explained as the interaction among stimuli. The detection threshold of a stimulus may vary due to the existence of another. This kind of varying can be either positive or negative. Due to masking effect, similar artifacts may be disturbing in certain regions of an image while they are hardly noticeable elsewhere. Hence, in video quality assessment the masking effect can be used to deal with different scenarios with different tips to acquire perceptually consistent result.



(a) Achromatic CSF.

(b) Chromatic CSF.

Fig. 1. Typical spatio-temporal contrast sensitivity functions.

2.3 Multi-channel Mechanism

The multi-channel theory of human vision states that different stimuli are processed in different channels in the human visual system. Both the physiological and psychophysical experiments carried out on perception gave the evidence of the band-pass nature of the cortical cells' response in the spectral domain and the human brain seems therefore to possess a collection of separate mechanisms, each being more sensitive to a portion of the frequency domain. This suggests a filter bank approach to the modeling of vision. Filter bank is usually used to approximate the various mechanisms of vision and decompose the visual data in a collection of signals that are band-limited in orientation, spatial frequency and temporal frequency. The spatial frequency domain is usually divided in four to eight bands in a logarithmic partition and there exists about the same number of orientation bands. The temporal frequency axis seems to be covered by two to three channels. Recent studies tend to confirm the common concept of the existence of only two temporal mechanisms (transient and sustained). Typical spatial and temporal filter banks are illustrated in Fig. 2 [12]-[13], where 17 spatial channels and two temporal channels are shown.

2.4 Depth Perception

In normal conditions, people see a real world scene with two eyes in two slightly different directions, which lead to a horizontal shift between corresponding points in the two dimensional projections of the three dimensional real world scene on the left and right retina. The shift is known as the retinal disparity. And because of the existing of disparity, the relative depth information in the scene being viewed can be deduced by human brain. As a result, people can have three dimensional feelings.

It is well known that the ability of human visual system to discriminate depth variation is not limitless but within a scope. In this paper, we define the minimum relative depth that the human visual system can discriminate as the depth perception threshold, which can be computed by

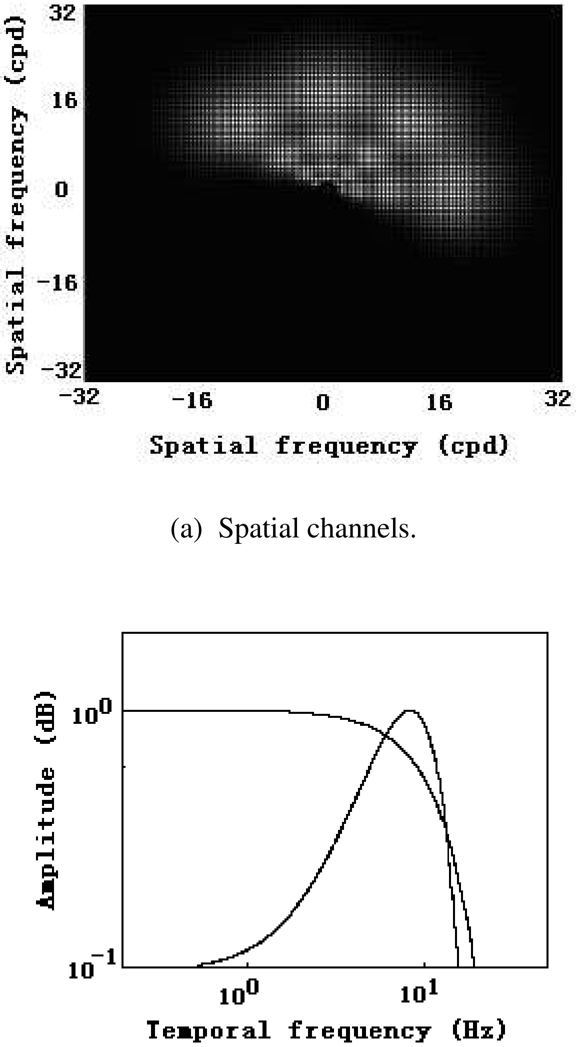
$$\Delta Z = \Delta \partial Z^2 / B \, , \qquad (1)$$

where ΔZ denotes the depth perception threshold, Z is the absolute distance between the real world point and camera focal point, B is the human baseline between two eyes, and $\Delta\partial$ is the stereo acuity.

From equation (1), it can be observed that the depth perception threshold depends not only on stereo acuity and human baseline between two eyes, but also on the point's spatial position in the scene.

3 Perceptual Evaluation Metric Based on 3D Wavelet Decomposition

Current metrics for stereo video quality evaluation use the same traditional methods as for single view video such as PSNR, MSE and so on. Due to lack of consideration of HVS characteristics, these methods are hard to acquire consistent results with human perception. In this paper, a perceptual distortion metric for stereo video quality evaluation is proposed, which incorporates the main properties of human visual system. The block diagram of the metric is shown in Fig. 3.



(a) Spatial channels.

(b) Temporal channels.

Fig. 2. Typical spatial and temporal channels.

Computation of the metric consists of the following several steps: perceptual decomposition, contrast conversion and masking, pooling and quality mapping. Each step is briefly introduced below.

3.1 Perceptual Decomposition

The aim of perceptual decomposition, including spatial decomposition, temporal decomposition and view decomposition, is to decompose the stereo video signal into different subsets corresponding to different visual channels.

As mentioned above, the temporal mechanism usually consists of two channels, transient channel and sustained channel. Let $f_i(x, y, v)$ denotes the original stereo video sequence, where i is the frame number, $v \in \{r, l\}$ is the view number. Then the process of temporal decomposition can be expressed as

$$g_i(x, y, v) = f_i(x, y, v) * h(n), \qquad (2)$$

where $h(n)$ is the impulse response function of channel, $g_i(x, y, v)$ is the filtered output result.

After temporal decomposition, the above output is further subject to spatial decomposition. Traditional spatial decomposition is performed by using a filter bank at different frequencies and different orientations. This kind of implementation is computationally complex. Due to its similar property to filter bank and its efficient implementation, recently the wavelet transform technology is alternatively used to perform spatial decomposition. Since the stereo video has two view channels, the traditional 2D wavelet transform cannot be directly used here. Therefore, a 3D wavelet decomposition technology based on disparity compensation view filtering (DCVF) [14] technology is employed in this paper. The DCVF based 3D wavelet decomposition is illustrated in Fig.4, where $g_i(x, y, l)$ and $g_i(x, y, r)$ are the temporal filtered results of $f_i(x, y, l)$ and $f_i(x, y, r)$ respectively, DC means disparity estimation and compensation, and IDC means the inverse of DC.

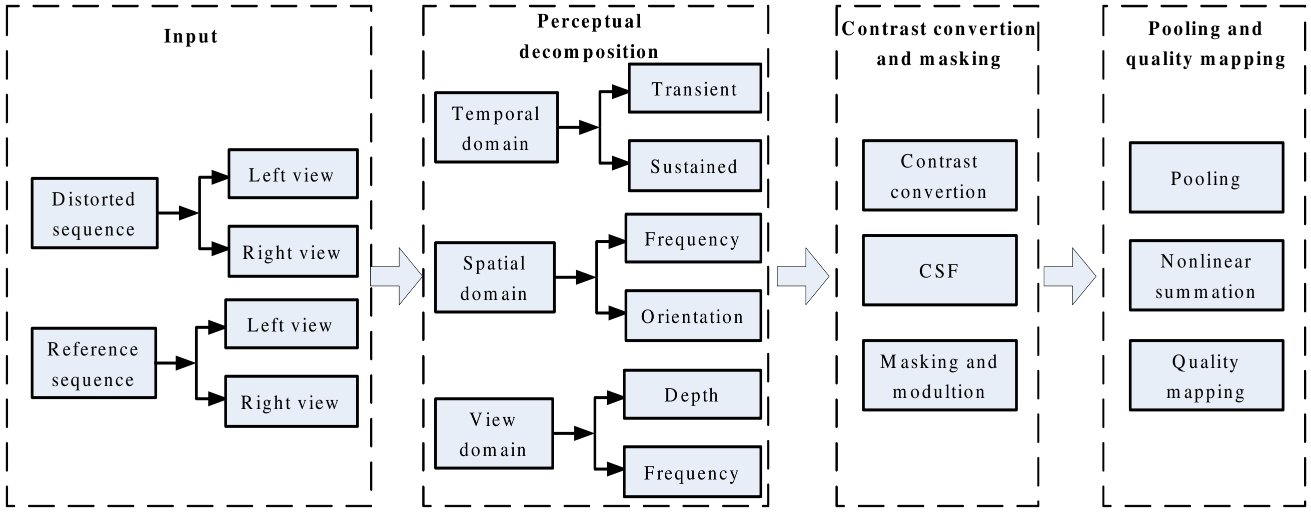


Fig. 3. Block diagram of the proposed perceptual distortion metric.

After perceptual decomposition, each stereo pair, the simultaneously acquired two images in the left and the right views at the same time, is decomposed into several subsets as shown in Fig. 5, where H_t denotes the high frequency band after temporal filtering including the left high frequency band $H_t^l(x, y)$ and the right high frequency band $H_t^r(x, y)$, H_v is the high frequency band after view filtering and D is the acquired disparity information after disparity estimation.

3.2 Modulation and Masking

After perceptual decomposition, each stereo pair is decomposed into several subsets of coefficients:

$$f_i(x, y, v) = \{S_{k,\theta}, H_t, H_v, D(x, y)\}\ . \qquad (3)$$

Different subsets are assumed to be processed in different channels. The final overall output of HVS is the fusing of the result of every channel. Assume $S_{k,\theta}$ is the subset at scale k and phase θ after wavelet decomposition. According to HVS theory, human perceived visual quality relates not only to the contrast but also to the frequency of the stimulus as well as the masking effect. Hence, for each subset $S_{k,\theta}$, its output through HVS system can be formulated as

$$r_{k,\theta}(x,y) = K_s csf_k^{p-q} \frac{c_{k,\theta}^{p}(x,y)}{\Delta + \sum_{\theta} c_{k,\theta}^{q}(x,y)}, \quad (4)$$

where csf_k denotes the contrast sensitivity of scale k, $c_{k,\theta}(x,y)$ is the contrast of $S_{k,\theta}$, Δ is a non-zero constant selected to prevent division by zero, K_s is the gain control factor, p and q are called excitatory and inhibitory exponents respectively. In most cases, p is fixed at 2 and q can be chosen between 1 and 3[15]. csf_k and $c_{k,\theta}(x,y)$ are computed by

$$c_{k,\theta}(x,y) = \frac{S_{k,\theta}(x,y)}{\bar{l}_{k-2}(x,y)}, \quad (5)$$

$$csf_k = 2.6(0.192 + 0.114 f_k)e^{-(0.114 f_k)^{1.1}} \quad (6)$$

where $\bar{l}_{k-2}(x,y)$ is the mean low-pass response under scale $k-2$, f_k denotes the spatial frequency of scale k in unit of cycle per degree (cpd).

For stereo video, disparity is the key factor to deduce depth information and perceive stereo sense. Similar to CSF, we define a Depth Sensitivity Function (DSF) computed by

$$dsf = \frac{1}{\Delta Z} = \frac{B}{\Delta \partial Z^2}. \quad (7)$$

where ΔZ denotes the depth perception threshold, Z is the absolute distance between the real world point and camera focal point, B is human baseline between two eyes, $\Delta\partial$ is the stereo acuity.

The depth information can be deduced in terms of disparity $D(x,y)$, camera focus F and the baseline B. When considering the parallel configuration stereo imaging system, depth information can be calculated by

$$Z = -\frac{FB}{d_{x(l\to r)}}, \quad (8)$$

where $d_{x(l\to r)}$ is the horizontal disparity which can be acquired through disparity estimation.

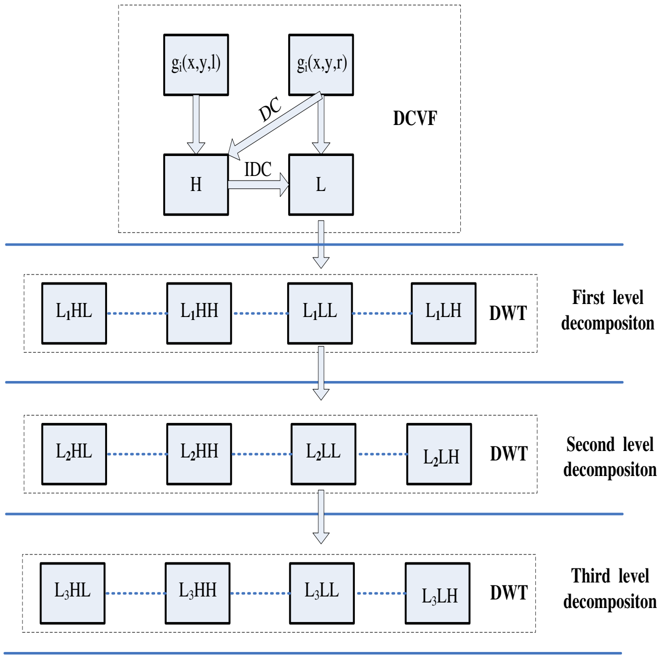


Fig. 4. DCVF based wavelet decomposition.

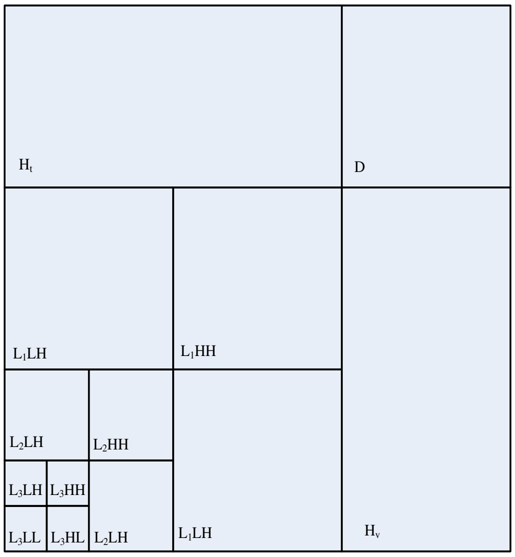


Fig. 5. Data structure of decomposed stereo pair.

In a standard stereo pair, the left and the right images should meet

$$f_i(x, y, l) = f_i(x + d_{x(l\to r)}, y + d_{y(l\to r)}, r)\ , \qquad (9)$$

where $d_{x(l\to r)}$ and $d_{y(l\to r)}$ are the horizontal and the vertical disparities of pixel (x, y) respectively.

In real distorted stereo pairs, the above relation may not come to existence, which will affect the overall stereo visual quality. Herein, the perceptual response to depth of HVS is defined by

$$\begin{aligned} r_z(x, y) = {} & K_z dsf\ {}^g c_z^i(x, y) + (1 - K_z) dsf\ \left| H_t^l(x, y) \right. \\ & \left. + H_v(x, y) - H_t^r(x, y) \right| , \qquad (10) \end{aligned}$$

where $c_z(x, y)$ is the depth contrast, $H_t^l(x, y)$ and $H_t^r(x, y)$ are the high frequency bands after temporal filtering, K_z is weight coefficient which can be chosen from 0.5 to 0.8.

3.3 Summation and Mapping

Suppose $f_i^r(x, y, v)$ and $f_i^d(x, y, v)$ are the reference stereo pair and the distorted stereo pair respectively. Let $r_{k,\theta}^r(x, y)$, $r_z^r(x, y)$ and $r_{k,\theta}^d(x, y)$, $r_z^d(x, y)$, computed by equations (4) and (10), denote the visual responses corresponding to $f_i^r(x, y, v)$ and respectively $f_i^d(x, y, v)$. Then the overall visual distortion e_i between reference stereo pair $f_i^r(x, y, v)$ and the distorted stereo pair $f_i^d(x, y, v)$ can be measured by

$$\begin{aligned} e_i = {} & \sum_k A_k \Big(\sum_{\theta, x, y} \left| r_{k,\theta}^r(x, y) - r_{k,\theta}^d(x, y) \right|^4 \Big)^{1/4} \\ & + B_z \Big(\sum_{x, y} \left| r_z^r(x, y) - r_z^d(x, y) \right|^4 \Big)^{1/4} , \end{aligned} \qquad (11)$$

where A_k and B_z are weight coefficients that are determined experimentally. After the distortion having been calculated, it can be easily further mapped to visual quality[1]-[4].

4 Experimental Results

To evaluate the performance of the proposed metric, experiments are implemented. Several standard test sequences are used including *Train and tunnel*, *Im* and *Sergio* stereo sequences. For each original test sequence, some distorted versions are produced by randomly adding Gaussian noise, low-pass filtering and compressing at different levels. For each distorted sequence, its quality is evaluated by using the proposed metric, PSNR metric and mean opinion score (MOS), respectively.

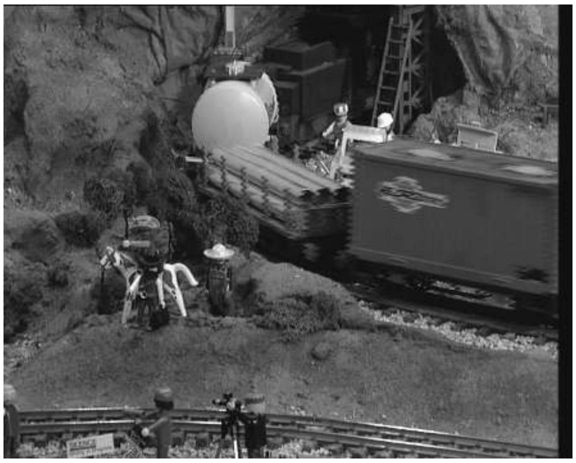
To implement subjective evaluation, our test procedure is set up similar to the double stimulus impairment scale (DSIS) method formalized in ITU-R Recommendation BT.500-10[16]. In our experiments, for each standard sequence, 16 distorted versions are produced by randomly adding Gaussian noise, low-pass filtering and compressing at different bit-rates. Both the reference and the distorted sequences are presented to the observers for only once at a viewing distance 5 times the screen height under the same lighting and viewing conditions. For each distorted sequence, each observer is asked to answer “how close it visually resembles the original reference” and a score between 0 and 5 is subsequently asked to give.

The evaluation results of every metric are normalized to the scope from 0 to 1 according to following principle:

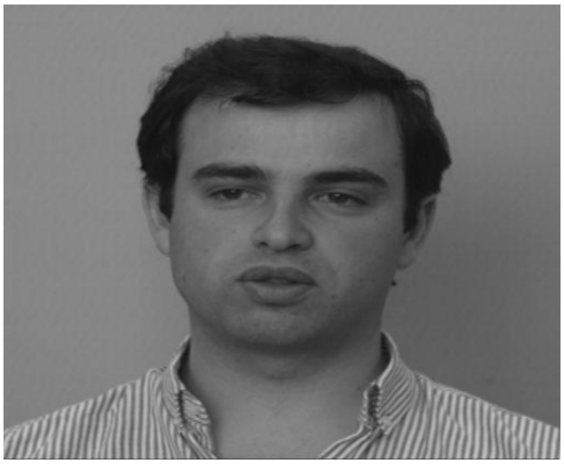
$$S_n = \frac{S_t - S_{\min}}{S_{\max} - S_{\min}} , \qquad (12)$$

where S_t denotes the calculated value of each metric, S_n denotes the normalized value. $S_{\min}$ and $S_{\max}$ are pre-defined possible minimum and maximum values.

(a) Train and tunnel stereo sequence

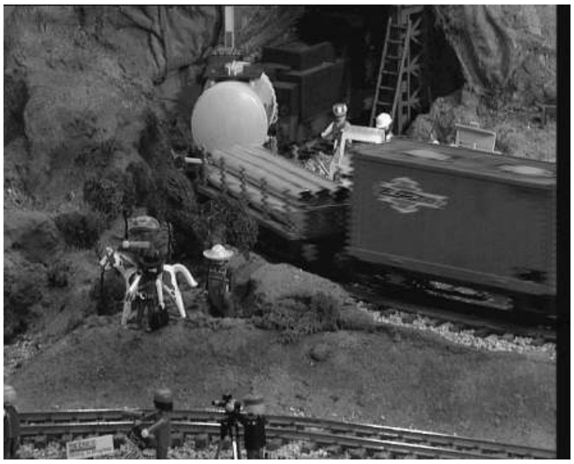


(b) Im stereo sequence

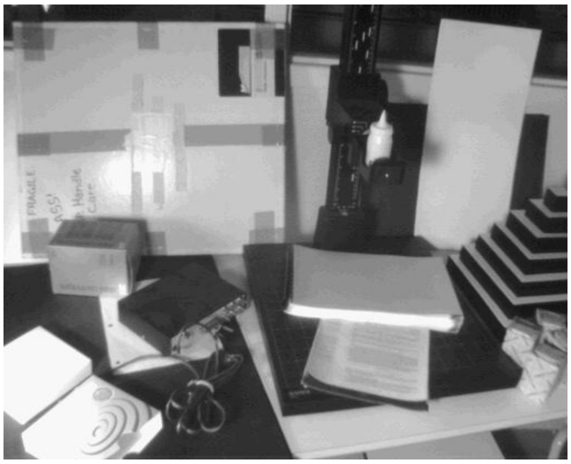


(c) Sergio stereo sequence

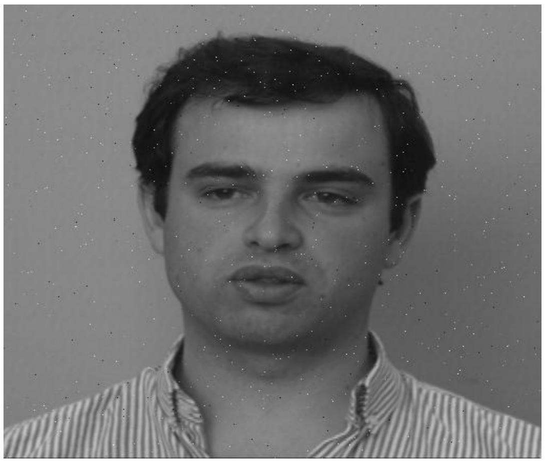
Fig. 6. Original test sequences.



(a) Evaluation results of a low-pass filtered sequence (Train and tunnel), where MOS score is 0.75, PSNR score is 0.48, and the score of proposed metric is 0.74.



(b) Evaluation results of a compressed sequence (Im), where MOS score is 0.63, PSNR score is 0.47, and the score of proposed metric is 0.57.



(c) Evaluation results of a noise polluted sequence (Sergio), where MOS score is 0.43, PSNR score is 0.51, and the score of proposed metric is 0.38.

Fig. 7. Some evaluation results of distorted sequences with MOS, PSNR, and the proposed metric, respectively.

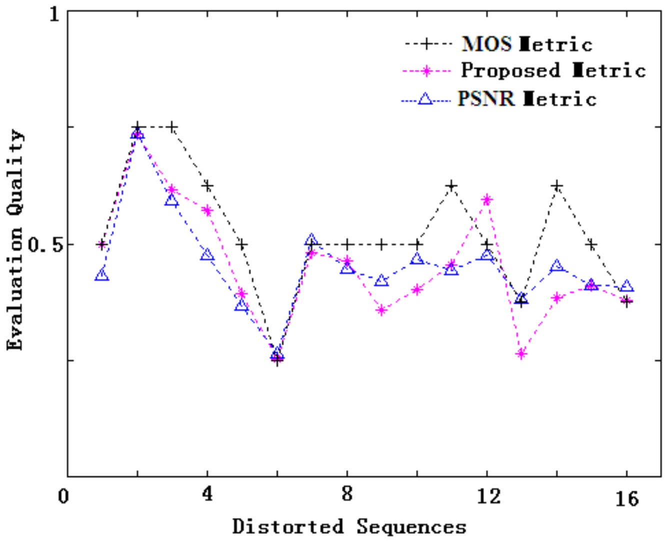
The correlations between each of the three objective metrics, that is, the proposed metric, PSNR metric, MSE metric, and the MOS method are calculated to justify which one of the three metrics is mostly consistent with HVS. The correlations are calculated by

$$R = \frac{\sum_{i}^{N} (p_i^k - \overline{p^k})(q_i - \overline{q})}{[\sum_{i}^{N} (p_i^k - \overline{p^k})^2 \sum_{i}^{N} (q_i - \overline{q})^2]^{1/2}}, \qquad (13)$$

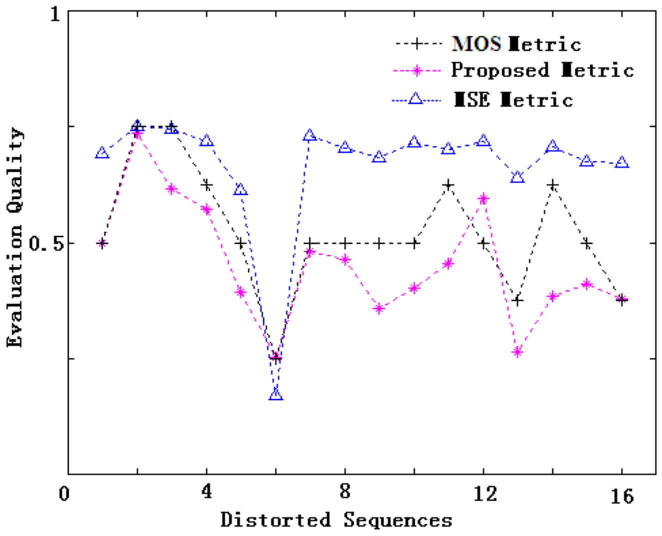
where $\{p_i^k\}, \{q_i\}$ are the subjective and the objective evaluation samples, N is the number of samples.

Partial results are given in Fig. 7, Fig.8 and Table 1, where Fig.7 gives the evaluation results of some distorted samples, Fig.8 are the comparative results of the three metrics, and Table 1 shows the correlations between the three objective metrics and the MOS method.

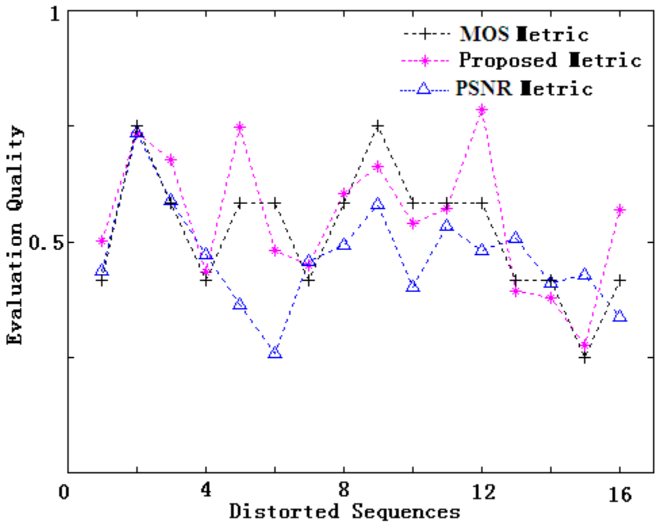
Results from Fig.7, Fig.8 demonstrate that the evaluation results of the proposed novel metric are more close to the MOS results than that of the PSNR or MSE. Table 1 also shows that the correlation between the proposed metric and MOS is high than that of PSNR or MSE. As a whole, the experimental results show that the proposed metric is more consistent with human perception.



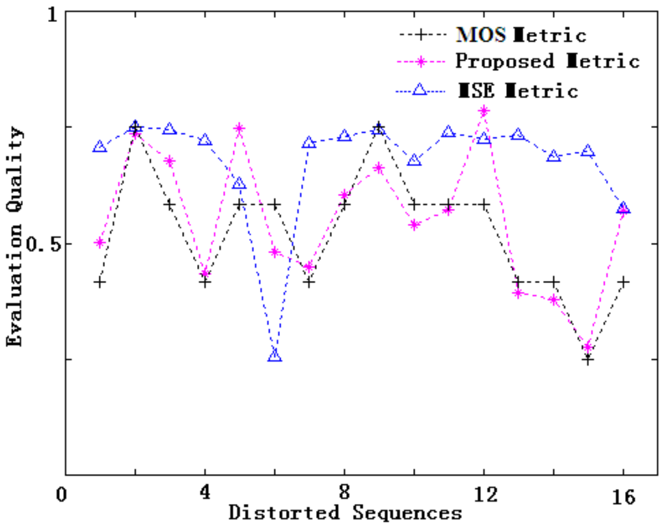
(a)



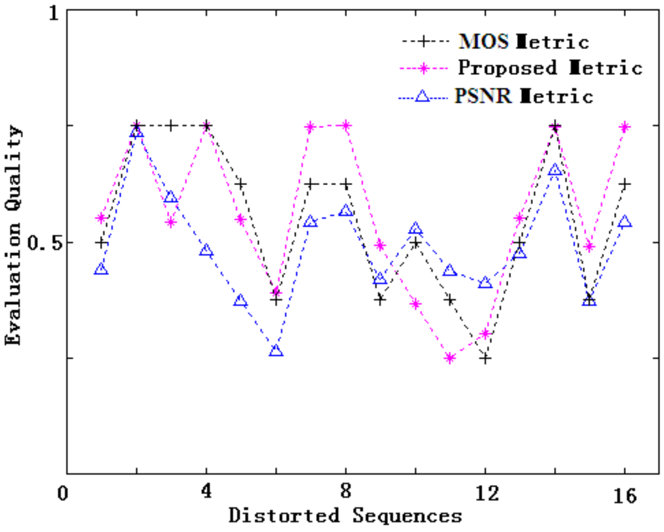
(b)



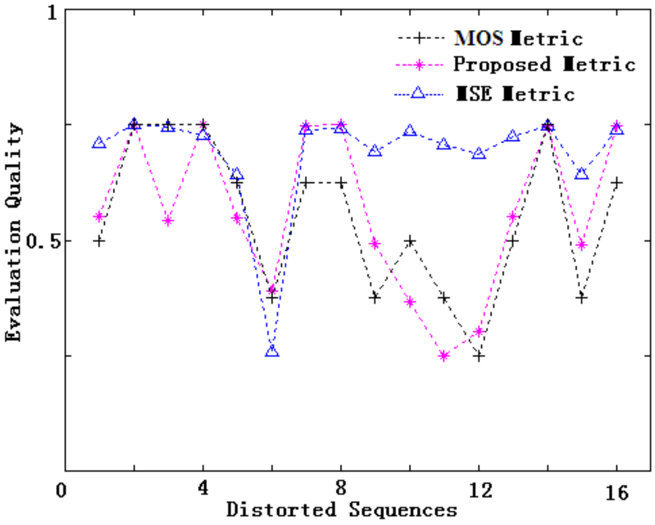
(c)



(d)



(e)



(f)

Fig. 8. Comparisons results of the three metrics, where (a) and (b) are the experimental results of *Im* sequence, (c) and (d) are the results of *Sergio* sequence, (e) and (f) are the results of *Train and tunnel* sequence.

Table 1. Comparison of correlations of three metrics

metrics	test sequences		
	Im	Sergio	Train
proposed metric	0.6732	0.6179	0.6628
PSNR metric	0.6353	0.4942	0.6164
MSE metric	0.4647	0.1594	0.3513

5 Conclusion

Development of perceptually consistent video quality metrics is a very difficult task and limited success has been achieved so far due to the complexity and the limited understanding to the process of the human visual system. In this paper, some main properties of human visual system such as the Contrast Sensitivity Function, Multi-channel and Masking, Depth Perception and so on are firstly analyzed. Then a perceptual model for stereo video quality evaluation is proposed, which mainly consists of three steps: wavelet-based perceptual decomposition, contrast conversion and masking, pooling and quality mapping.

Simulation has been performed and some experimental results are given, which reveal that, compared with traditional objective metrics such as PSNR and MSE, the proposed model is shown to be more consistent with subjective human perception. However, due to extremely complex of the HVS, it still needs to be further studied. There also exists a gap between the evaluation results of the proposed metric and the subjective perception. Hence the proposed metric shall be further improved.

6 Acknowledgment

This work was supported in part by the Natural Science Foundation of Zhejiang Province (No.Y107740);the Open Project Foundation of Ningbo Key Laboratory of DSP (No.2007A22002); the Natural Science Foundation of Ningbo(No.2008A610015).

References:

[1] M.Ahmet, S.Paul, "Image quality measures and their performance", IEEE Transactions on communications, pp. 2959-2965, 1995.

[2] H. Lee, D.Haynor, and Y. Kim, "Subjective evaluation of compressed image quality", Proceedings of SPIE, Image Capture, Formatting and Display, pp. 241-245, 1992.

[3] W. Stefan, "Vision models and quality metrics for image processing applications," PH.D Thesis, Swiss Federal Institute of Technology, Lausanne, Ecublens, Switzerland, 2000.

[4] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 2, pp. 260-273, 2006.

[5] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," Signal Processing Letters, vol. 12, no. 10, pp. 685-688, 2005.

[6] S. Yao, W. Lin, E. Ong, and Z. Lu, "A wavelet-based visible distortion measure for video quality evaluation," Proceedings of International Conference on Image Processing, pp. 2937-2940, 2006.

[7] R. Dosselmann, and X. D. Yang, "A prototype no-reference video quality system," Proc. of Canadian Conference on Computer and Robot Vision, pp. 411-417, 2007.

[8] M. Masry, and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," Signal Processing: Image Communication, vol. 19, pp. 133-146, 2004.

[9] F. W. Campbell, and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," Journal of Physiology, vol. 197, pp. 551-566, 1968.

[10] J. J. Koenderink, and A. J. Van-Doorn, "Spatiotemporal contrast detection threshold surface is bimodal," Optics Letters, vol. 4, no. 1, pp. 32-34, 1979.

[11] D. H. Kelly, "Spatiotemporal variation of chromatic and achromatic contrast thresholds," Journal of the Optical Society of America, vol. 73, no. 6, pp. 742-750, 1983.

[12] C. J. Van Den Branden Lambrecht, and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," Proc. of SPIE, San Jose, LA, vol. 2668, pp. 450-461, 1996.

[13] R. F. Hess, and R. J. Snowden, "Temporal properties of human visual filters: number shapes and spatial covariation," Vision Research, vol. 32, no. 1, pp. 47-59, 1992.

[14] W. X. Yang, Y. Lu, F. Wu, J. F. Cai, K. Ngi Ngan, and S. P. Li, "4D wavelet-based multi-view video coding," IEEE transaction on Circuits and Systems for Video Technology, vol. 16, no. 11, pp. 1385-1396, 2006.

[15] J. M. Foley, "Human luminance pattern-vision mechanisms: experiments require a new model," Journal of the optical Society of America, vol.11, no.6, pp. 1710-1719, 1994.

[16] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-10, Geneva, Switzerland, 2000.