Information Sciences 2017

A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain

Jiachen Yang [a], Huanling Wang [a], Wen Lu [b,*], Baihua Li [c], Atta Badii [d], Qinggang Meng [c]

[a] School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300072, China
[b] School of Electronic Engineering, Xidian University, Xi'an 710071, Shanxi, China
[c] Department of Computer Science, School of Science, Loughborough University, UK
[d] Department of Computer Science, School of Mathematical, Physical and Computational Sciences (SMPCS), University of Reading, UK

A B S T R A C T

Most of the existing 3D video quality assessment (3D-VQA/SVQA) methods only consider spatial information by directly using an image quality evaluation method. In addition, a few take the motion information of adjacent frames into consideration. In practice, one may assume that a single data-view is unlikely to be sufficient for effectively learning the video quality. Therefore, integration of multi-view information is both valuable and necessary. In this paper, we propose an effective multi-view feature learning metric for blind stereoscopic video quality assessment (BSVQA), which jointly focuses on spatial information, temporal information and inter-frame spatio-temporal information. In our study, a set of local binary patterns (LBP) statistical features extracted from a computed frame curvelet representation are used as spatial and spatio-temporal description, and the local flow statistical features based on the estimation of optical flow are used to describe the temporal distortion. Subsequently, a support vector regression (SVR) is utilized to map the feature vectors of each single view to subjective quality scores. Finally, the scores of multiple views are pooled into the final score according to their contribution rate. Experimental results demonstrate that the proposed metric significantly outperforms the existing metrics and can achieve higher consistency with subjective quality assessment.

1. Introduction

Recently, the rapid technological developments of 3D image and video processing have led to an explosion in demand for 3D content [20]. 3D can bring consumers stereo perception and immersive viewing experiences, and can be displayed not only in movie theaters, but also on home electronics, e.g., television, smartphones, and tablets. Human eyes are the final receiver of stereo videos, and the user's Quality of Experience (QoE) ultimately represents the quality of videos. At this point, stereo video quality plays an important role in assessing 3D video coding, transmission and applications. Thus the research of stereo video quality evaluation is of particular significance.

E-mail addresses: yangjiachen@tju.edu.cn (J. Yang), wanghl07@tju.edu.cn (H. Wang), luwen@mail.xidian.edu.cn, luwen.xidian@gmail.com (W. Lu), B.Li@lboro.ac.uk (B. Li), atta.badii@reading.ac.uk (A. Badii), q.meng@lboro.ac.uk (Q. Meng).

The research on SVQA includes full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. The FR and RR need to utilize the pristine videos or their partial information, while NR need not refer to original videos [38]. The disadvantage of FR or RR methods is that the reference videos are not always accessible in most practical situations. As a solution, NR methods that do not require reference to the original videos is receiving growing attention. NR-SVQA research aims to develop a perceptual model to evaluate the quality of distorted stereo videos automatically and accurately without access to the non-distorted reference videos, but it is much more challenging owing to a lack of relevant statistical and perceptual models. Accordingly, there are fewer blind 2D-VQA algorithms and blind SVQA algorithms.

There have been several proposed approaches to generate an effective and accurate quality metric for stereoscopic videos. However due to the lack of generality, robustness and practicability, there is currently no widely accepted solution. The earliest pioneers who attempted to evaluate the quality of stereo videos would typically apply 2D image quality assessment (2D-IQA) or 2D-VQA methods in their metrics. In [13,43], 2D-IQA measures, including PSNR, SSIM, and 2D-VQA measures VQM [32], were applied to the left and right view of 3D videos separately and then averaged to obtain a 3D quality score. Experimental results showed that VQM performs better than PSNR and SSIM. In [37], a systematic prediction-bias-inspired SVQA metric was proposed and a subjective stereo video database was studied. The results indicated that the proposed model accounting for the prediction bias leads to significant performance boost compared with the methods that directly averaging 2D video quality of both views to predict 3D video quality.

Compared with the traditional 2D videos, 3D videos contain additional depth information which is created by the difference between two views and may enhance or degrade the overall 3D viewing experience depending on the effect of the image/video processing system [39]. Thus the delivery of appropriate perceptual quality of stereo videos is much more challenging than that of 2D videos. Taking into consideration that stereo videos contain additional depth signals, some proposed methods have assessed the overall 3D video viewing experience mainly depending on the effect of two views and depth information. Hewage et al. [14] proposed a 3D video quality metric by evaluating two views and analyzing the edges and contours of the depth map. A similar RR-SVQA metric was proposed by Malekmohamadi et al. [25], they extracted side information from edge properties and gray level co-occurrence matrices from color and depth sections. Furthermore, Zhu et al. [49] also presented a SVQA method by taking two quality metrics into account. The quality of left-right views was based on significant pixels and a just noticeable distortion model, and the depth perception quality was based on a three-dimensional wavelet transform. However, the lack of accuracy in these metrics is mainly due to the fact that the human visual system (HVS) is a complex advanced system and it is difficult to model it in 3D by only analyzing pixels and depth.

Recently, an effective and promising method to quantify QoE of 3D is to take HVS models or HVS characteristics into consideration to devise more reliable measures of perceived quality. Given that HVS can perceive the difference between two retinal images to create a synthetic image with depth perception, the authors in [6] presented the 'cyclopean image' model based on binocular rivalry; they integrated the left and right images into a cyclopean image to simulate brain perception, and then used the 2D-IQA metric on the cyclopean image to derive a quality value. Yu et al. [48] presented a RR-SVQA method by modelling fusion and rivalry in the process of binocular perception, so stereo video frames were divided into binocular fusion portion and binocular rivalry portion. And experimental results demonstrated that the proposed method was highly consistent with the subjective perception. In [10], Galkandageit et al. introduced a novel HVS model taking into consideration the phenomena of binocular suppression and recurrent excitation. They used the proposed HVS model and an optimized temporal pooling strategy to obtain the final video quality. The experimental results showed its robustness.

Since videos contain large amount of data and require a long processing time, few researchers have studied the video quality assessment in the past few years; this has seriously hindered the development of stereoscopic video quality evaluation. In recent years, with the rapid development of computer technology, the continuous improvement of GPU performance and the establishment of public databases, more and more researchers are beginning to focus on the evaluation of stereo video quality.

In this paper, we introduce a blind multi-view feature learning algorithm for quality evaluation of stereo videos. In particular, we assume that the degradation of video quality mainly results in distortions in the (i) spatial, (ii) temporal and (iii) spatio-temporal interaction domains. Accordingly, the effort of this work is devoted to modelling and quantifying the interactions of the above three elements by training different regression models and developing weighting schemes. The contributions of this paper are fourfold:

(1) A new NR-SVQA method for extracting a small number of interpretable features relevant to perceptual quality has been proposed. To the best of our knowledge, this is the earliest ever proposal to deploy a NR-SVQA model.
(2) Videos often contain multi-view content and have different representations. The conventional methods of training all the features at once may not take into account all the information contained in the video and the compatibility between them. To tackle the above problems, we model the spatial, temporal as well as spatio-temporal video attributes separately, and analyze the essentiality and relationship between the three.
(3) Optical flow statistics can be affected by distortions, therefore temporal distortion is estimated based on the optical flow method. Furthermore, the experimental results show that the proposed temporal features can greatly improve the performance.
(4) Through a comprehensive validation, we show that our metric correlates well with subjective observations not only for symmetrically distorted stereoscopic videos, but also for asymmetrically distorted stereoscopic videos. This demonstrates that our approach can be deployed as a general quality evaluator for various stereoscopic video applications.

The remainder of this paper is organized as follows. Section 2 briefly introduces the motivations and framework of our metric. Section 3 presents the detailed feature implementation of our proposed method. The performance of our framework is presented in Section 4, using the publicly available SVQA database built by Qi et al. [33] and the NAMA3DS1- COSPAD1 database [35]. We conclude the paper in Section 5 with a discussion of future work.

2. Motivations and the proposed algorithm framework

2.1. Motivations

Unlike 3D images, stereo videos do not just carry information over the spatial domain. Therefore for stereo video quality assessment, it is not very effective to directly use the stereo image quality evaluation method which considers only spatial information into stereo video quality assessment. Traditionally, video is naturally represented by multi-view features including spatial and temporal domain information. Different views characterize different data properties and different features can complement one another. In the past few years, multi-view feature learning [40,41] has received growing attention and has been studied extensively. Xu et al. [42] and Yan et al. [46] adopted multi-view features learning technology to realize video classification. Ding et al. [7] extracted Multi-Directional Multi-Level Dual-Cross Patterns (MDML-DCPs) from images to realize face recognition. All experiments demonstrated that multi-view features can provide more characteristics and can significantly improve the learning performance.

In our study, one underlying assumption is that the perceived stereo video quality not only depends on the spatial information (frame level information) but also can be affected by the temporal information and the inter-frame information along the temporal axis (we refer to this as spatio-temporal information). Therefore, exploring the temporal characteristics and the spatio-temporal characteristics is also of great significance for SVQA. In our implementation, we perform the multi-view feature learning from the spatial view, the temporal view and the spatio-temporal view.

By contrast, there are even fewer blind SVQA algorithms than blind SIQA algorithms; this is due to a lack of relevant statistical and perceptual models. Feature detection is a fundamental and important problem in computer vision and image processing. For quality evaluation, the ideal features should behave in a manner that depends on the degree of distortion and on the perception of distortion. The NVS (natural video statistics) features in [28,34] delivers excellent video quality predictive power. During recent years, extensive investigation of feature detection methods has been carried out [8,22]. Thereinto texture structure is a general concept that can be applied to image recognition and image/video quality assessment [19,21] and can deliver prominent performance. Existing studies show that image texture carries essential visual information and that HVS adaptably extracts structural and texture information for image perception and understanding. Therefore, proper extraction and description of image texture information plays a significant role in perceptual quality assessment. At present, there are a large number of algorithms to describe texture. For instance, the wavelet transform and the curvelet transform are the most adequate techniques to extract texture. These transforms decompose the original image into subbands that preserve high and low frequency information, and enable the image to be represented by multiple scales in a way that is quite close to what takes place in the HVS. Compared with curvelet transform, wavelets only work well in one-dimension. In addition, curvelet transform has been proved to be a powerful tool for multi-resolution analysis of images. Curvelet can provide a sparse representation of the objects exhibiting 'curve punctuated smoothness' [4].

In addition, LBP is also an effective texture description operator with many significant advantages. For instance, it generates histograms that are very useful in representing texture features that are invariant to orientation (e.g., rotations) and brightness levels. Traditionally, the LBP features are extracted from the raw pixel values. However, curvelet transform and LBP appear to be more effective for texture analysis. The combination of LBP and curvelet transform will better describe the structure of curvelet coefficients, and the literature survey has revealed that the combination of curvelet transforms with LBP yields more effective feature vectors than using the curvelet transform or the LBP alone [29]. The curvelet-based LBP texture operator is an effective feature extractor, which can describe texture through only a few parameters.

As well as the perception of spatial information, the temporal information also plays a very important role in human perception of moving image sequences; this is often measured by the motion information including the motion intensity and the motion direction. Born and Bradley [2] claimed that visual area V5 of HVS plays a major role in the perception of motion and the guidance of eye movements. Therefore, much research effort has been devoted to the study of the effect of distortion on temporal motion characteristics. The success of VQA algorithms should depend on their abilities to model motion perception in the HVS. Manasa et al. [26] estimated the temporal distortion by the use of the mean, the standard deviation, the coefficient of variation, and the minimum eigenvalue of local optical flow statistics. In addition, experiments demonstrated that local flow statistics are effective features for assessing temporal quality. Further, optical flow can represent motion information at its finest resolution, which motivates us to work with the optical flow for motion processing.

2.2. The proposed-algorithmic framework

Based on the above analysis and discussion, the model we propose predicts the overall quality of stereo videos by integrating the contributions of multi-view information, i.e., spatial quality, temporal quality, and spatio-temporal quality. The framework of the proposed SVQA metric is shown in Fig. 1. Essentially this consists of four parts: (1) quality calculation of spatial domain by extracting visual perception features from the binocular summation channel and binocular difference

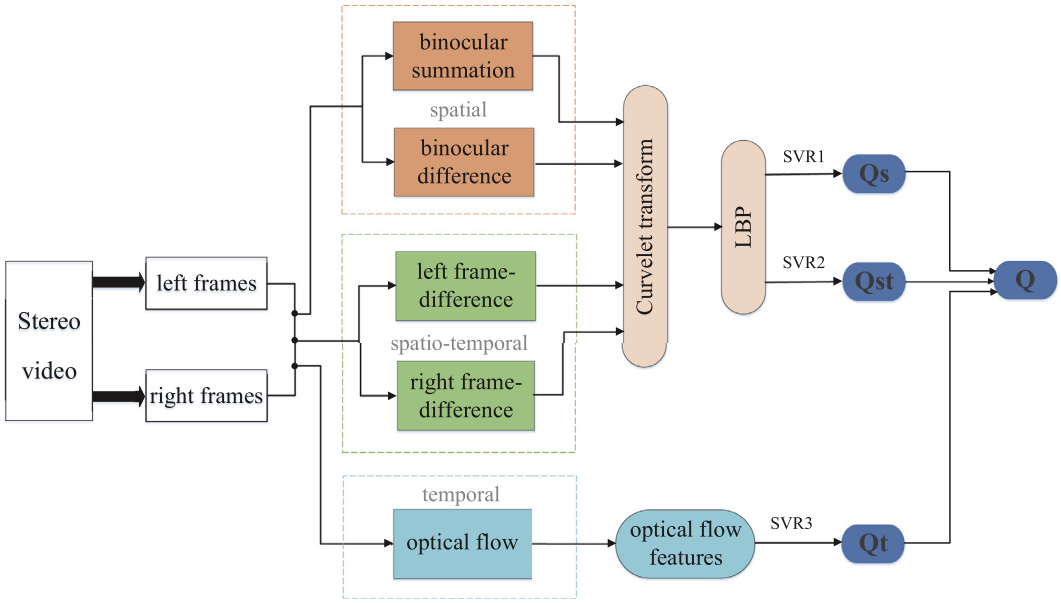

Fig. 1. Framework of the proposed methodology.

channel of stereo video; (2) quality calculation of spatio-temporal domain based on the difference of adjacent frames; (3) quality calculation of temporal domain by extracting optical flow features, which reflects the effect of distortion on temporal variability of video; and finally (4) pooling the spatial quality, the temporal quality, and the spatio-temporal quality into the ultimate quality perception metric for stereo video.

Next, the elaborate description of the proposed method is divided into four subsections. The binocular summation channel and binocular difference channel that can represent spatial binocular characteristics are presented in Section 2.2.1. Section 2.2.2 is devoted to the evaluation of the spatio-temporal distortions at eye fixation level. The evaluation of the temporal distortion on the whole video sequence is described in Section 2.2.3. Finally, the Section 2.2.4 describes the overall stereo video quality.

2.2.1. Binocular summation channel and binocular difference channel

We look out the world through two separate eyes, but our brains combine the images from the left and right eyes into a single fused percept. In general, according to human vision system characteristics, binocular vision can operate in several different 'modes' [6,16,45]. Our previous work [44,45,47] has proposed the binocular summation and difference channels, and has shown that it may help the brain to encode binocular information in an efficient fashion. Moreover, compared with other binocular models, our proposed model has the merit of low-complexity and fast computation speed, which indicates that it is suitable for stereo video processing with large amount of data. Thus, we will use the binocular summation and difference channels to represent the spatial domain frame-by-frame. Given the left view L and right view R, the binocular summation signal S and binocular difference signal D can be calculated as follows:

$$\begin{aligned} S &= \frac{L}{2} + \frac{R}{2} \\ D &= |L - R| \end{aligned} \tag{1}$$

The summation and the difference images between two views are shown in Fig. 2. The summation image looks like a 3D image which we watch while not wearing a stereoscope, while the difference image emphasizes contour information associated with depth sensation. In [12], Sid et al. also conducted an investigation into summation and difference signals and proposed that signals of neurons in primary visual cortex can equally be expressed as a weighted combination of summation and difference signals. In addition, recently May and Zhao [27] stated that applying a larger gain to the difference channel would make more sense in efficiently encoding images as perceived by the two human eyes respectively. However this has not been demonstrated, but our experiments have provided support for it.

2.2.2. Spatio-temporal interaction distortions

We define the inter-frame difference along the temporal axis as the spatio-temporal information. Fig. 3 plots an example of the frame differences. It can be seen that the spatio-temporal interaction mainly includes the spatial information of the moving object. The interaction between motion and spatial change is of particular interest, and the statistics of frame-differences have previously been explored. Dong and Atick [9] found that frame-difference natural videos reliably obey a

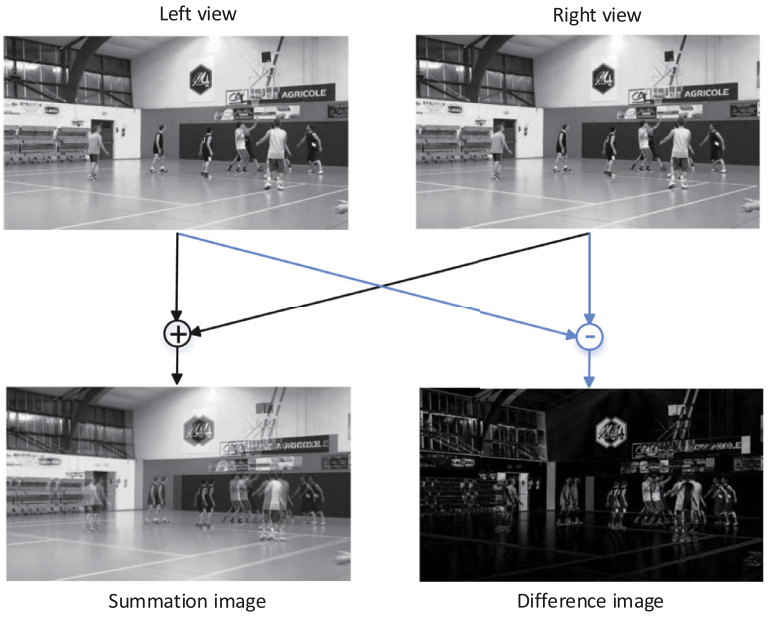

Fig. 2. 200th frame of the distorted Basket stereo video with a MOS of 3.36 and the corresponding summation and difference images.

Fig. 3. The spatio-temporal interaction images (MOS = 3.36). (a) The left frame-difference of Barrier gate sequence. (b) The right frame-difference of Barrier gate sequence.

space-time spectral model. Saad et al. [34] also found that the pristine and the distorted frame-difference videos obey a DCT coefficient statistics regularity. In our model, the frame-difference statistics are characterized by a coherency measure for which we define and use the parameters derived from the LBP coefficients on the curvelet transform.

2.2.3. Temporal distortions

Motion analysis is one of the main tasks of stereo video quality assessment. However, most existing VQA algorithms are able to capture spatial distortions that occur in video sequences, but do not do an adequate job in capturing temporal distortions, and most researchers directly use the motion magnitude as the temporal feature. In this paper, we measure the temporal information by the optical flow algorithm. The optical flow methods try to calculate the motion between two image frames which are taken at times t and $t + \Delta t$ at every voxel position. We make a brief overview of the optical flow algorithms used in our experimental study. In our studies, the Horn–Schunck method [15] was applied to obtain the motion vectors in the temporal domain.

Let $I(x, y, t)$ denote the image intensity in the point (x, y) at time t, and this point will move to $(x + \Delta x, y + \Delta y)$ at time $t + \Delta t$, denoted by $I(x + \Delta x, y + \Delta y, t + \Delta t)$. Let $v = (v_x, v_y)$ denote the optical flow between the two image frames, where v_x and v_y are the x and y components of the velocity. The well-known optical flow constraint equation can be written as the following:

$$I_x v_x + I_y v_y + I_t = 0 \tag{2}$$

The optical flow in a pristine natural video is generally smooth; when the distortion sets in, this smoothness is lost. Distortion affects the magnitude and direction of optical flow. Fig. 4(a) shows a non-distorted video frame, Fig. 4(b) and (c) show video frames with low perceptual quality, respectively. Fig. 4(d) shows a 32 × 32 optical flow patch from the reference video frame. Similarly, Fig. 4(e) and (f) show a 32 × 32 optical flow patch from the low quality frame at the same spatial location as the reference patch. While Fig. 4 is an illustrative example, we found the principle that distortion affects the intensity and direction of optical flow work consistently well over a large set of frames and videos.

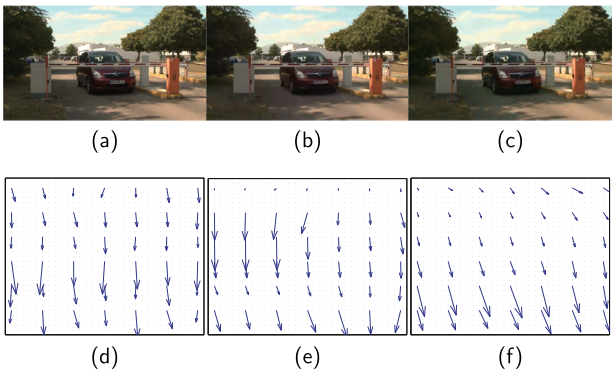

Fig. 4. Illustration of the effect of distortions on the optical flow. (a) 200th frame of the original Barrier gate sequence. (b) 200th frame of the Barrier gate sequence with a MOS of 2.85. (c) 200th frame of the Barrier gate sequence with a MOS of 1.82. (d) A 32 × 32 optical flow patch of (a). (e) A 32 × 32 optical flow patch of (b). (f) A 32 × 32 optical flow patch of (c).

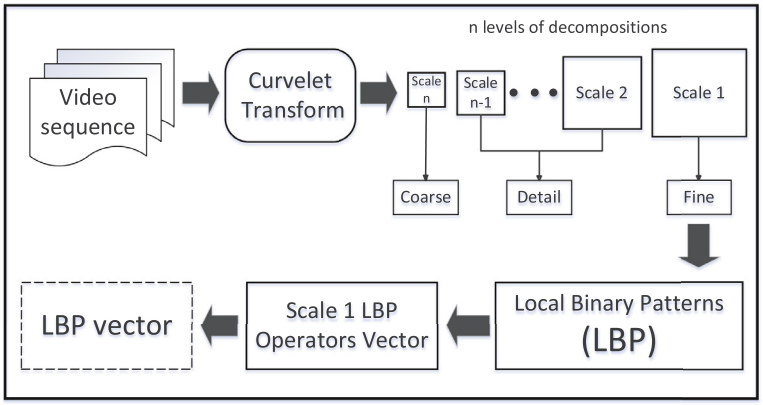

Fig. 5. Block diagram of the main steps for curvelet feature extraction.

2.2.4. The overall stereo video quality evaluation

In order to evaluate stereo video quality, these features extracted from each individual view are then used to drive a SVR through training, respectively. Finally, the quality of stereoscopic video is a relation of the scores of the three parts. The most general functional form of a model of first order that represents the overall stereo video quality (Q) as a function of the spatial quality (Q_s), spatial-temporal quality (Q_{st}) and temporal quality (Q_t) is:

$$Q = \alpha \cdot Q_s + \beta \cdot Q_{st} + \gamma \cdot Q_t + K \tag{3}$$

$$Q_s = \mu \cdot Q_{summation} + \eta \cdot Q_{difference} \tag{4}$$

where $\alpha + \beta + \gamma = 1$, $\mu + \eta = 1$, K is a constant offset, $Q_{summation}$ is the quality of binocular summation channel, and $Q_{difference}$ is the quality of binocular difference channel. This model takes into account the separate contribution from each view.

3. NVS-model-based features

Feature extraction is a key issue in designing a NR-I/VQA model. In this section, we begin by describing the texture spatial and spatio-temporal features that is expressed in the curvelet domain. We then discuss the motion analysis process according to the coefficients of patch optical flow.

3.1. Spatial and spatio-temporal feature extraction

In our study, the steps involved in spatial and spatio-temporal feature extraction are schematised in Fig. 5, which are based on the application of LBP to curvelet transforms with the purpose of differentiating distortion.

Table 1
Structure of curvelet coefficients.

Name of layer	Scale index	Orientation number	Matrix form
Fine	C1	1	256×128
Detail	C2	32	$(67 \times 22) \times 4$ $(64 \times 22) \times 12$ $(44 \times 34) \times 4$ $(44 \times 32) \times 4$ $(43 \times 32) \times 8$
	C3	16	$(44 \times 18) \times 4$ $(42 \times 16) \times 4$ $(35 \times 22) \times 4$ $(32 \times 22) \times 4$
Coarse	C4	1	43×21

3.1.1. Curvelet transform

The curvelet transform represents an image at different scales and angles. Candes et al. [5] introduced a variation of the curvelet transform called fast discrete curvelet transform (FDCT). There are two implementations to perform FDCT. The first one is based on unequally spaced fast Fourier transforms (USFFT). The second one is based on the wrapping of specially selected Fourier samples. Compared with the USFFT, the wrapping method is faster and reduces redundant information; hence, in the present work we have chosen the wrapping method. Curvelets are parameterized not only by spatial position and scale, but also by orientation. The curvelet coefficients $c^D(j, l, k_1, k_2)$ at different scales and orientations can be derived by:

$$c^D(j,l,k_1,k_2) = \sum_{n=1}^{N}\sum_{m=1}^{M} f[m,n]\varphi^D_{jlk_1k_2}[m,n]. \tag{5}$$

where k_1 and k_2 denote coordinates in the spatial domain, and j and l are the scale and orientation parameters, respectively [1].

Several studies have focused on the extraction of features from curvelet transforms. However, the curse of dimensionality problem arises when using the curvelet coefficients. Obviously, the use of all the curvelet coefficients of all levels and all bands is not practical since this would lead to a curse of dimensionality problem. Therefore a reduction method is required to extract a reduced set of discriminative features.

In this study, we use FDCT via the wrapping method to extract textural descriptor features from each set of stereoscopic videos, and we use the curvelet transform scalar division principle $nscales = ceil(\log_2(\min(M, N)) - 3)$ (whereby M, N defines the size of the input image). For example, for a 1280×720 image, the scale is 7; and for a 1920×1080 image, the scale is 8. For simplicity, we have shown the structure of the curvelet coefficients of a 256×128 image as in Table 1.

The curvelet coefficients at coarse scales mainly embody the low frequency information, while the curvelet coefficients at fine scales mainly embody the high frequency information, especially including image edge and contour. Usually, distortions often affect the high frequency components of a video, whereas low frequency components are less affected. Accordingly, we only consider the curvelet coefficients at Fine scales in our model.

3.1.2. Local binary patterns

After the computation of the curvelet transform, we implement the LBP algorithm on the extracted curvelet coefficients. LBP has emerged as one of the most prominent and widely studied local texture descriptors [3,23,31]. Given an image pixel c, the basis LBP pattern at c is computed by comparing the gray value c of given pixel with the gray values of its p neighbours according to the following formula,

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{6}$$

where P is the number of neighbours, R is the radius of the neighbourhood, and $s(x)$ is a threshold function.

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{7}$$

Ojala et al. [30] have demonstrated that the traditional LBP does not provide very good discrimination, and have thus introduced an improvement to the non-uniform LBP, namely:

$$LBP^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1 & \text{otherwise,} \end{cases} \tag{8}$$

where

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} \left|s(g_p - g_c) - s(g_{p-1} - g_c)\right|. \tag{9}$$

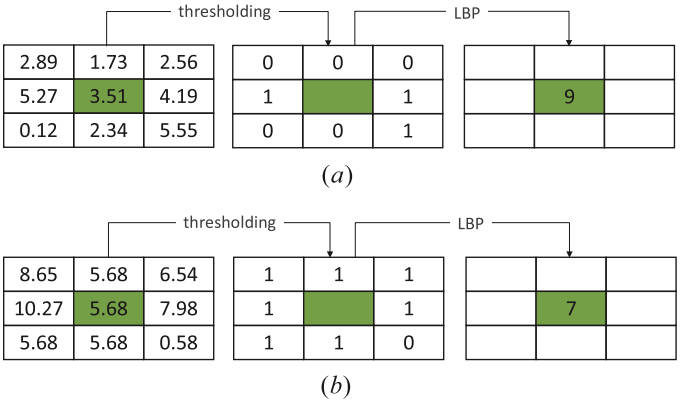

Fig. 6. LBP operator performed onto a 3 × 3 neighbourhood.

where U is the uniform measure, superscript ^{riu2} reflects the use of rotation invariant "uniform" patterns that have U value of at most 2. By definition, $LBP^{riu2}_{P,R}$ has $P+2$ $(0 \sim P+1)$ distinct output patterns in all. In general, given an $M\times$ N input matrix, a LBP pattern can be computed at each pixel c, such that a general matrix can be converted into a constant matrix whose value is between 0 and $P+1$. Fig. 6 shows two examples of circularly symmetric neighbourhood sets used to compute the LBP operator for parameters $P=8$ and $R=1$.

After the computation of $LBP^{riu2}_{P,R}$, we compute the number of times each pattern occurs, the bins of which are given by:

$$H(k)=\sum_{i=1}^{N}\sum_{j=1}^{M} f(LBP^{riu2}_{P,R}(i,j),k), \quad k\in[0,K] \tag{10}$$

where K is the maximal $LBP^{riu2}_{P,R}$ pattern, and $f(x, y)$ coincides with the threshold σ given by:

$$\sigma(x,y)=\begin{cases}1, & x=y\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

The final LBP features can be expressed as the frequency of each pattern:

$$g(i)=H(i)\Big/\sum_{k=0}^{P+1}H(k), i\in[0,K] \tag{12}$$

The application of LBP operators on curvelet coefficients is similar to LBP operation on gray scale images. We chose the parameters $P=8$ and $R=1$ in our study.

3.2. Temporal feature extraction

As just discussed in Section 2, naturalistic videos exhibit statistical regularities and strong correlations over both space and time. Thus, characterizing these regularities on pristine videos and distorted videos is an important step towards developing models of stereo video quality assessment. Therefore, we propose the following 10 temporal statistical features based on optical flow aiming at measuring the degree of temporal distortion.

For a velocity field $\mathbf{v}$, we calculate

$$v=\sqrt{{v_x}^2+{v_y}^2} \tag{13}$$

$$div(\mathbf{v})=I_x v_x+I_y v_y, \quad shA(\mathbf{v})=I_x v_x-I_y v_y \tag{14}$$

$$rot(\mathbf{v})=I_x v_y-I_y v_x, \quad shB(\mathbf{v})=I_x v_y+I_y v_x \tag{15}$$

For each optical flow matrix of two adjacent frames, we divide it into $K\times L$ non-overlapping patches. Then for each patch, we calculate v, $div(\mathbf{v})$, $rot(\mathbf{v})$, shA($\mathbf{v}$) and shB($\mathbf{v}$) and use them to define ten flow features as

$$\varphi[v],\ \varphi[div(\mathbf{v})],\ \varphi[rot(\mathbf{v})],\ \varphi[shA(\mathbf{v})],\ \varphi[shB(\mathbf{v})] \tag{16}$$

$$\Lambda[v],\ \Lambda[div(\mathbf{v})],\ \Lambda[rot(\mathbf{v})],\ \Lambda[shA(\mathbf{v})],\ \Lambda[shB(\mathbf{v})] \tag{17}$$

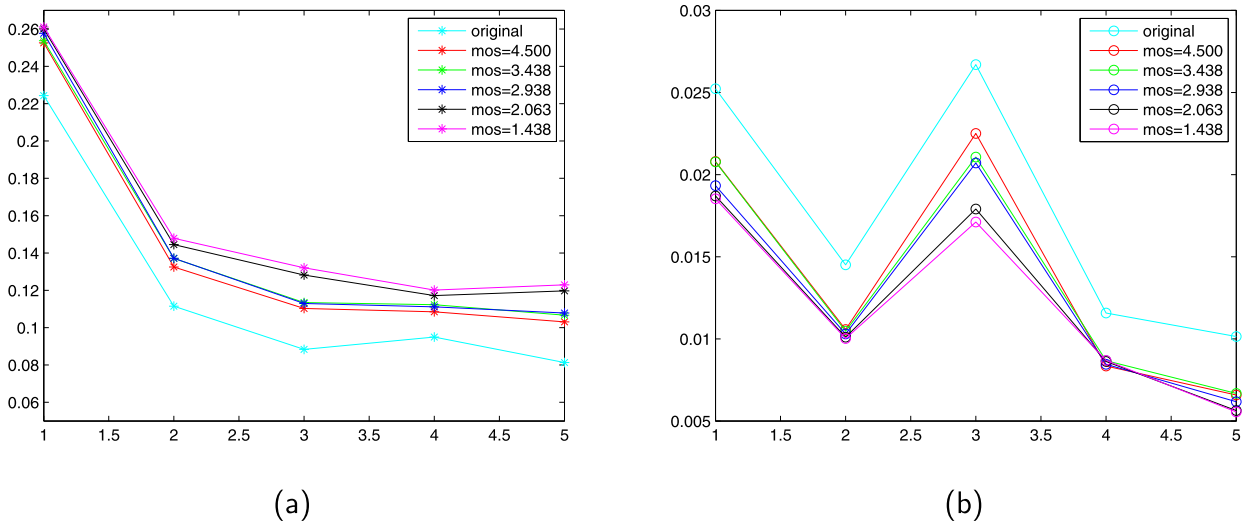

Fig. 7. Illustration of the effectiveness of selected temporal features. (a) Scatter plot of 5 temporal features shown in Eq. (15) vary with different MOS. (b) Scatter plot of 5 temporal features shown in Eq. (16) vary with different MOS.

Fig. 8. First frames of nine original stereo videos in QI-SVQA database. From left to right and top to bottom: Ballons, BookArrival, Champagne, Dog, Kendo, Outdoor, Pantomime, Poznan Street, Tsinghua Classroom.

where

$$\varphi[\xi] = \frac{\|\sum \xi_i\|}{\sum \|\xi_i\|} \tag{18}$$

$$\Lambda[\xi] = \frac{(\sum \|\xi_i\|)^2}{N \sum \|\xi_i\|^2} \tag{19}$$

Fig. 7 is the scatter plot of temporal features versus MOS (Mean Opinion Scores) for the video sequences. It can be observed that the parameter φ increases as the amount of perceived distortion in the video increases, while the parameter Λ decreases as the amount of perceived distortion in the video increases.

4. Experimental results and discussion

In this section, we describe the databases used for testing the proposed method, and conduct an extensive set of experiments to prove the discrimination effectiveness of the proposed method. We also discuss the evaluation results.

4.1. Stereo video database

To test the performances of the proposed NR-SVQA method, the publicly available stereo video database in [33] (we name it QI-SVQA) and the NAMA3DS1-COSPAD1 stereo video database [35] are used. The QI-SVQA database is in the format of uncompressed YUV 4:2:0, which has a total of 450 stereo video clips at 25 frames/s derived from nine reference videos. The majority of videos in the QI-SVQA database are asymmetrically distorted, only a small part are symmetrical distortion videos. Fig. 8 shows the left and right views of each reference video. The database contains videos distorted by two distortion types:

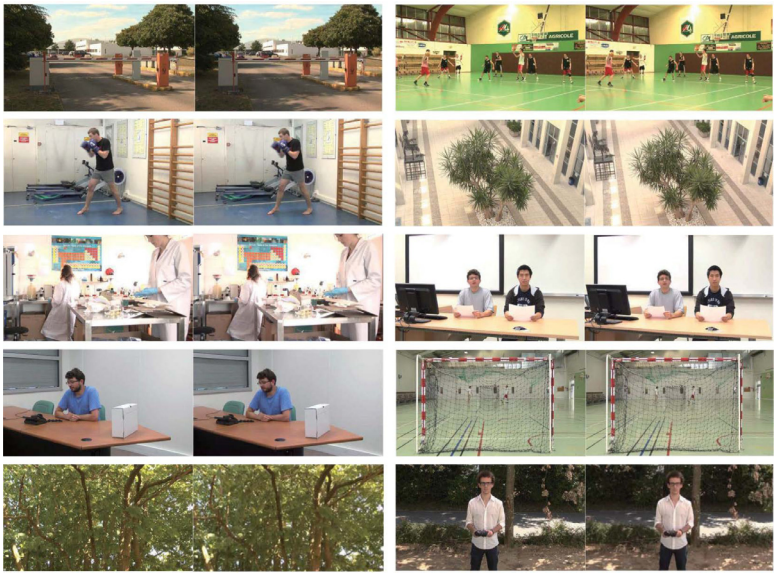

Fig. 9. First frames of ten original stereo videos in the NAMA3DS1-COSPAD1 stereo video database. From left to right and top to bottom: Barrier gate, Basket, Boxers, Hall, Lab, News report, Phone call, Soccer, Tree branches, Umbrella.

Gaussblur and H.264. The MOS scores are representative of the perceived quality of videos, and the value of which in the QI-SVQA database range from 1 (bad) to 5 (excellent). In order to reduce the time of video format conversion and improve the computational efficiency, Y channel videos of QI-SVQA database are directly used as test videos in the experiment.

The NAMA3DS1-COSPAD1 stereo video database consists of 10 original 1920 $\times$ 1080 3D full HD stereo videos of highly diverse spatial and temporal content at 25 frame/s (Fig. 9), and corresponding 100 symmetrically distorted stereo videos that span five distortion categories, including H.264/AVC, JPEG 2000, reduction of resolution, image sharpening, and down-sampling & sharpening. The MOS also range from 1 to 5.

4.2. Algorithms and performance measures

Four evaluation criteria are chosen in the performance evaluation: Pearson linear correlation coefficient (PLCC), Spearman rank correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC) and Root mean squared error (RMSE). A value approaching 1 for PLCC, SROCC and KROCC and a value approaching 0 for RMSE indicate the best correlation between the predicted scores after a five-parameter nonlinear regression and the human subjective scores of the database. At the feature extraction stage, in order to improve the real-time function, our proposed algorithm selected one frame out of every four frames, and then averaged and normalized their values. During the training and learning phase, we randomly split a database into independent training and testing sets: 80% of the database was used for training and the remaining 20% was used for testing. We repeated 80% training – 20% testing 1000 times, and the median performance evaluation indices across 1000 iterations were used as the final algorithm performance evaluation. The performance of the proposed approach is demonstrated with two experiments. Excellent results in each experiment testify that the proposed framework is highly correlated to the subjective scores.

In order to evaluate the performance of the proposed algorithm, we compared its performance on the two databases with two widely used 2D quality metrics: PSNR and SSIM [36]. Besides, to study the benchmark performance, six existing state-of-the-art 3D objective video quality metrics have been tested. For 2D quality metrics, PSNR and SSIM are calculated as average grade over grades from all frames and both views. For SVQA metrics BEVQM [10], BEVQM_mean indicates a temporal average pooling strategy of BEVQM algorithm, while BEVQM_minkowski indicates a minkowski summation pooling strategy of BEVQM algorithm.

4.3. Parameter optimization

To determine the five parameters in the proposed metric, we compare the performances of one parameter with different values. We firstly conduct an experiment to determine the μ and η in Eq. (3), which denote the weights adjusting the balance of summation and difference channels in the spatial quality; such that $\mu + \eta = 1$. The PLCC and SROCC of the spatial quality performed in QI-SVQA database are shown in Fig. 10. The rational numbers for μ range from 0 to 1 at an interval of 0.1. It can be concluded that among them $\mu = 0.4$ and $\eta = 0.6$ deliver the best performance, which is consistent with the

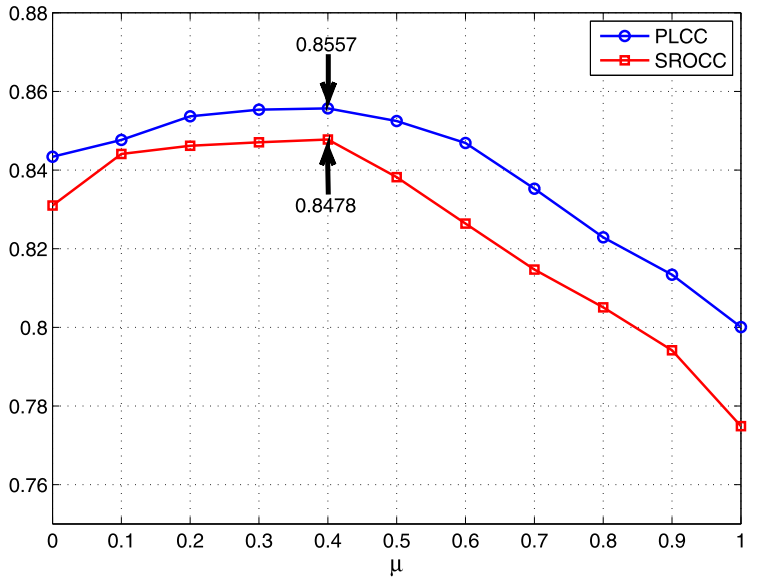

Fig. 10. PLCC and SROCC of the spatial quality on QI-SVQA database while changing μ value ($\mu + \eta = 1$). $\mu = 0.4$ delivers the best performance.

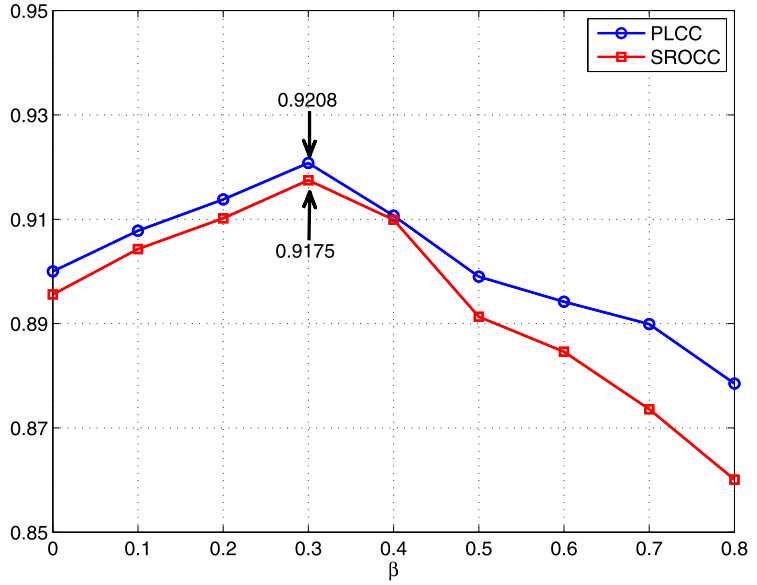

Fig. 11. PLCC and SROCC of the overall quality on QI-SVQA database while changing β, and γ ($\alpha = 0.2$ and $\alpha + \beta + \gamma = 1$). $\alpha = 0.2$, $\beta = 0.3$, and $\gamma = 0.5$ deliver the best performance.

theory that applying a larger gain to the difference channel would make more sense in efficiently encoding the respective images as perceived by two human eyes.

Then, we determined the three parameters in the combination process; such that $\alpha + \beta + \gamma = 1$. We fixed the value of parameter α, and compared the performance of the other two parameters as they were assigned different values. Our metric achieved the best performance with $\alpha = 0.2$, $\beta = 0.3$, and $\gamma = 0.5$. In Fig. 11, we have plotted PLCC and SROCC values of the overall quality on QI-SVQA database with the value of α fixed at $\alpha = 0.2$.

4.4. Overall prediction performance

Table 2 shows the overall performance of the proposed approach on QI-SVQA database and Table 3 shows the overall performance on NAMA3DS1-COSPAD1 database. The highest performed metrics are highlighted in boldface. As shown in Tables 2 and3, compared with other metrics, the proposed method delivers better correlation with the MOS values for the test videos. This provide strong support for the superiority of the proposed algorithm compared to other algorithms except

Table 2
Overall performance comparison on the QI-SVQA database. (80% of database was used for training).

Type	Metrics	PLCC	SROCC	KROCC	RMSE
2D-IQA	PSNR	0.8496	0.8637	0.6832	0.5122
	SSIM	0.8185	0.8281	0.6418	0.5580
3D-VQA	PQM [18]	0.7852	0.8165	0.6365	0.6158
	PHVS-3D [17]	0.7082	0.7195	0.5353	0.7021
	SFD [24]	0.6483	0.6633	0.5021	0.7571
	3D-STS [11]	0.8311	0.8338	0.6553	0.5520
	Ref. [33]	0.8415	0.8379	0.6650	0.5372
	Proposed	**0.9208**	**0.9175**	**0.7730**	**0.3709**

Table 3
Overall performance comparison on the NAMA3DS1-COSPAD1 database. (80% of database was used for training).

Type	Metrics	PLCC	SROCC	KROCC	RMSE
2D-IQA	PSNR	0.6699	0.6470	0.4800	0.8433
	SSIM	0.7664	0.7492	0.5444	0.7296
3D-VQA	BEVQM_mean [10]	0.8918	–	–	–
	BEVQM_minkowski [10]	**0.9052**	–	–	–
	Proposed	0.8949	**0.8552**	**0.6913**	**0.4929**

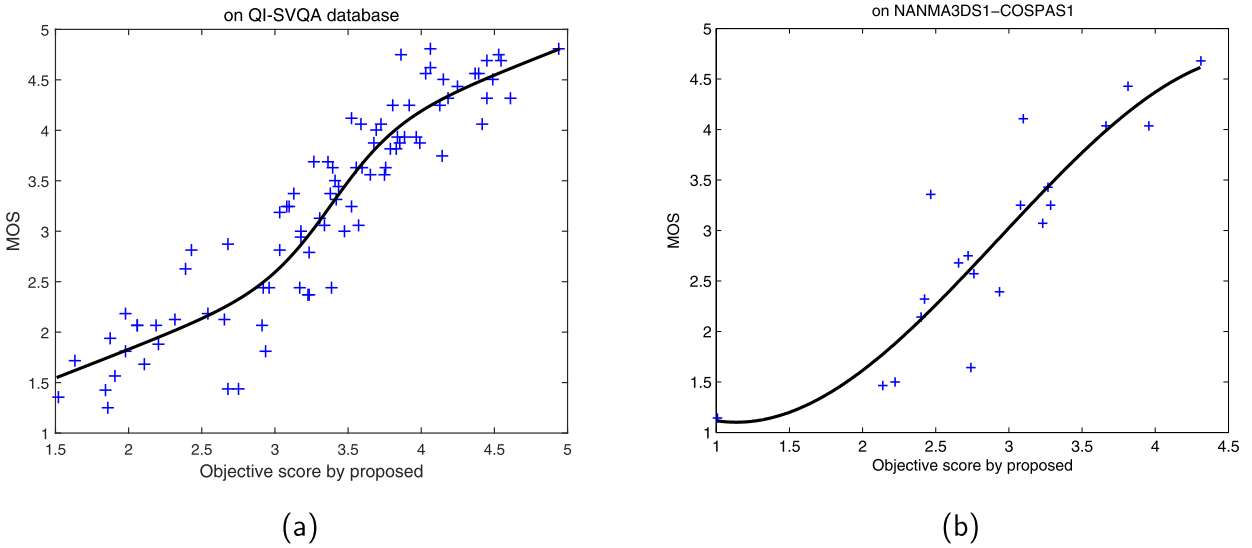

Fig. 12. Scatter plot of the MOS against mapped objective scores calculated by the proposed metric on the QI-SVQA and NAMA3DS1-COSPAD1 databases.

for BEVQM_minkowski. However BEVQM_minkowski is a FR-SVQA metric, while our method is a NR-SVQA metric. Furthermore, the temporal pooling strategy of our method is a temporal average, while that of BEVQM_minkowski is minkowski summation. In addition it can be seen from Table 2, BEVQM_minkowski is also superior to BEVQM_mean, which means that the temporal pooling strategy also plays an important role in video quality evaluation.

The QI-SVQA database has both symmetrically distorted videos and asymmetrically distorted videos, the excellent performance of the proposed method using this database means that our algorithm has strong flexibility and wide applicability. Particularly, in comparing the algorithms, it can be seen that the other six 3D video quality methods are all FR methods, while our method is a NR-SVQA method. Thus although our proposed algorithm does not refer to pristine stereo videos, its performance is still superior to most of the benchmarked FR-SVQA approaches. Accordingly, at present, our proposed NR-SVQA algorithm represents a pioneering approach with an outstanding performance that represents a significant advance in the field of stereo video quality assessment.

In addition, Fig. 12 shows the scatter plot of the MOS against mapped objective scores calculated by the proposed metric. It can be seen that our metric correlates well with subjective scores. Moreover, the complexity is as important as prediction accuracy in determining the feasibility of a SVQA scheme for practical deployment given the typically very large amount of data that would need to be processed for a stereo video. The proposed method is much simpler to implement than the other methods; since, unlike other schemes, we do not use a complex binocular vision model. Instead, we use a simple binocular

Table 4
Performance evaluation result of using features from each individual view in isolation from other features for quality prediction on the QI-SVQA database. (80% of database was used for training).

Database	Metrics	Spatial domain	Spatio-temporal domain	Temporal domain
QI-SVQA database	PLCC	0.8557	0.8620	0.8749
	SROCC	0.8478	0.8424	0.8754
	KROCC	0.6743	0.6778	0.7106
	RMSE	0.5105	0.4991	0.4756

summation and difference channels that have a low computational load to represent the binocular model. Furthermore, the proposed method benefits from the strategy of selecting one frame out of every four frames. These ensure efficiency and excellent prediction accuracy of the proposed scheme.

4.5. Each individual domain contribution to prediction performance

We now discuss the effectiveness of each individual domain feature. In order to examine and assess the contribution of each individual domain features to the overall prediction performance, we conducted an ablation study protocol whereby the feature of each individual domain in turn was used in isolation from the other features to predict quality, and the correlation between predicted and actual quality scores was computed in each case. Table 4 shows the performance evaluation result when using each individual domain feature in isolation from the other features for prediction of video quality. As can be seen from Table 4, the performance in temporal domain is better than that in the spatio-temporal domain, and the latter is slightly better than that in the spatial domain. That is, the contribution rate of temporal domain is larger than the spatio-temporal interaction and the spatial domain, which was also proved earlier in Section 4.3 of this paper ($\alpha = 0.2,\ \beta = 0.3$, and $\gamma = 0.5$).

However, some previous studies have shown that the spatial domain is more important than the temporal domain for video quality evaluation, and researchers have focused less on the distortion of temporal domain. Nevertheless, from the results of our experiments and some previous studies, it is concluded that the contribution rate of the temporal domain and the spatial domain are related to the selected features and the characteristics of the proposed algorithm. Moreover although, some previous studies have focused on the evaluation of spatial domain, few researchers have made outstanding contributions in the temporal domain. However in our algorithm, the evaluation of the temporal domain distortion has contributed to the excellent performance achieved and this indicates that the importance of the temporal domain distortion can not be ignored.

5. Conclusions

In this paper, we have introduced a novel stereo video quality assessment model by modelling the binocular perception effect in multi-views, including spatial domain, temporal domain and the spatial-temporal domain. Texture analysis features extracted by associating the curvelet transform and local binary pattern have been used in the analysis of distortion on the spatial and spatial-temporal domains. Besides, optical flow-based features have been explored for measuring temporal distortions and the significance of features to perceived quality was evaluated by applying the support vector regression. Finally, the results of multi-view feature learning have been pooled into a final score according to the contribution rate. We have conducted several experimental comparative studies of the proposed algorithm on the stereo video database in [33] and the NAMA3DS1-COSPAD1 stereo video database. Experimental results demonstrate that the proposed NR-SVQA algorithm is highly competitive with the state of-the-art methods and outperforms other methods whilst offering wider applicability, lower complexity and thus enhanced scalability beyond higher accuracy and simpler implementation due to being a no-reference method. Furthermore, the use of a simple binocular summation and difference channels and the strategy of selecting one frame of every four frames greatly improves the computational efficiency of the algorithm.

Future work could be undertaken on two aspects. One is to explore the dynamic contribution rate of each of the three domains (spatial, temporal, and spatio-temporal) video data sets of variabled content and spatial frequency and thus improve the performance of the proposed algorithm further. The other aspect is to further explore the effect of distortion on temporal characteristics.

Acknowledgments

This research is partially supported by the Natural Science Foundation of China (No. 61471260), and Natural Science Foundation of Tianjin: 16JCYBJC16000.

References

[1] S. AlZubi, N. Islam, M. Abbod, Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation, J. Biomed. Imaging 4 (2011).
[2] R.T. Born, D.C. Bradley, Structure and function of visual area MT, Annu. Rev. Neurosci. 28 (2005) 157–189.

[3] S. Brahnam, L.C. Jain, L. Nanni, A. Lumini, Local Binary Patterns: New Variants and Applications, Springer, Berlin Heidelberg., 2014.
[4] E.J. Candès, What is... a curvelet? Not. Am. Math. Soc. 50 (11) (2003) 1402–1403.
[5] E. Candès, L. Demanet, D. Donoho, L. Ying, Fast discrete curvelet transforms, Multiscal. Model. Simul. 5 (3) (2006) 861–C899.
[6] M.J. Chen, C.C. Su, D.K. Kwon, L.K. Cormack, A.C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, Signal Process. Image Commun. 28 (9) (2013) 1143–1155.
[7] C. Ding, J. Choi, D. Tao, L.S. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 518–531.
[8] C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition, ACM Trans. Intell. Syst. Technol. (TIST) 7 (3) (2016) 37.
[9] D.W. Dong, J.J. Atick, Statistics of natural time-varying images, Netw. Comput. Neural Syst. 6 (3) (1995) 345–358.
[10] C. Galkandage, J. Calic, S. Dogan, J.Y. Guillemaut, Stereoscopic video quality assessment using binocular energy, IEEE J. Sel. Top. Signal Process. 11 (1) (2016) 102–112.
[11] J. Han, T. Jiang, S. Ma, Stereoscopic video quality assessment model based on spatial-temporal structural information, in: Proceedings of the IEEE Conference on Visual Communications and Image Processing (VCIP), 2012, pp. 1–6.
[12] S. Henriksen, J.C. Read, Visual perception: a novel difference channel in binocular vision, Curr. Biol. 26 (12) (2016) R500–R503.
[13] C.T. Hewage, S.T. Worrall, S. Dogan, A.M. Kondoz, Prediction of stereoscopic video quality using objective quality models of 2-d video, Electron. Lett. 44 (16) (2008) 963–965.
[14] C.T. Hewage, S.T. Worrall, S. Dogan, S. Villette, A.M. Kondoz, Quality evaluation of color plus depth map-based stereoscopic video, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 304–318.
[15] B.K. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1–3) (1981) 185–203.
[16] L. Jin, A. Boev, K. Egiazarian, A. Gotchev, Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3d video, EURASIP J. Image Video Process. 2014 1 (2014) 1–18.
[17] L. Jin, A. Boev, A. Gotchev, K. Egiazarian, 3d-DCT based perceptual quality assessment of stereo video, in: Proceedings of the Eighteenth IEEE International Conference on Image Processing, 2011, pp. 2521–2524.
[18] P. Joveluro, H. Malekmohamadi, W.C. Fernando, A.M. Kondoz, Perceptual video quality metric for 3d video quality assessment, in: Proceedings of the IEEE 3DTV Conference on the True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010, pp. 1–4.
[19] J. Kannala, E. Rahtu, BSIF: binarized statistical image features, in: Proceedings of the Twenty-first IEEE International Conference on Pattern Recognition (ICPR), 2012, pp. 1363–1366.
[20] A. Khan, S.A. Malik, A. Ali, R. Chamlawi, M. Hussain, M.T. Mahmood, I. Usman, Intelligent reversible watermarking and authentication: hiding depth map information for 3d cameras, Inf. Sci. 216 (2012) 155–175.
[21] Q. Li, W. Lin, Y. Fang, No-reference quality assessment for multiply-distorted images in gradient domain, IEEE Signal Process. Lett. 23 (4) (2016) 541–545.
[22] Y. Li, S. Wang, Q. Tian, X. Ding, A survey of recent advances in visual feature detection, Neurocomputing. 149 (2015) 736–751.
[23] L. Liu, P. Fieguth, Y. Guo, X. Wang, M. Pietikäinen, Local binary features for texture classification: taxonomy and experimental study, Pattern Recognit. 62 (2017) 135–160.
[24] F. Lu, H. Wang, X. Ji, G. Er, Quality assessment of 3d asymmetric view coding using spatial frequency dominance model, in: Proceedings of the 3DTV Conference on the True Vision-Capture, Transmission and Display of 3D Video, IEEE, 2009, pp. 1–4.
[25] H. Malekmohamadi, A. Fernando, A. Kondoz, A new reduced reference metric for color plus depth 3d video, J. Vis. Commun. Image Represent. 25 (3) (2014) 534–541.
[26] K. Manasa, S.S. Channappayya, An optical flow-based full reference video quality assessment algorithm, IEEE Trans. Image Process. 25 (6) (2016) 2480–2492.
[27] K.A. May, Z. Li, Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns, Curr. Biol. 26 (12) (2016) 1571–1576.
[28] A. Mittal, M.A. Saad, A.C. Bovik, A completely blind video integrity oracle, IEEE Trans. Image Process. 25 (1) (2016) 289–300.
[29] S. Nagaraja, C.J. Prabhakar, P.P. Kumar, Complete local binary pattern for representation of facial expression based on curvelet transform, in: Proceedings of the International Conference on Multimedia Processing, Communication and Information Technology (MPCIT), 2013, pp. 48–56.
[30] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[31] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: Proceedings of the International Conference on Image and Signal Processing, Springer, Berlin, Heidelberg, 2008, pp. 236–243.
[32] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[33] F. Qi, D. Zhao, X. Fan, T. Jiang, Stereoscopic video quality assessment based on visual attention and just-noticeable difference models, Signal Image Video Process. 10 (4) (2016) 737–744.
[34] M.A. Saad, A.C. Bovik, C. Charrier, Blind prediction of natural video quality, IEEE Trans. Image Process. 23 (3) (2014) 1352–1365.
[35] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, et al., NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences, in: Proceedings of the Fourth IEEE International Workshop on Quality of Multimedia Experience (QoMEX), 2012.
[36] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[37] J. Wang, S. Wang, Z. Wang, Asymmetrically compressed stereoscopic 3d videos: quality assessment and rate-distortion performance evaluation, IEEE Trans. Image Process. 26 (3) (2017) 1330–1343.
[38] J. Wu, W. Lin, G. Shi, L. Li, Y. Fang, Orientation selectivity based visual pattern for reduced-reference image quality assessment, Inf. Sci. 351 (2016) 18–29.
[39] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert, R. Scopigno, Browsing and exploration of video sequences: a new scheme for key frame extraction and 3d visualization using entropy based jensen divergence, Inf. Sci. 278 (2014a) 736–756.
[40] C. Xu, D. Tao, C. Xu, Large-margin multi-view information bottleneck, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014b) 1559–1572.
[41] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015a) 2531–2544.
[42] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Trans. Image Process. 24 (12) (2015b) 5812–5825.
[43] S.L.P. Yasakethu, C.T. Hewage, W.A.C. Fernando, A.M. Kondoz, Quality analysis for 3d video using 2d video quality models, IEEE Trans. Consum. Electron. 54 (4) (2008).
[44] J. Yang, Y. Liu, Z. Gao, R. Chu, Z. Song, A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior, J. Vis. Commun. Image Represent. 31 (2015a) 138–145.
[45] J. Yang, Y. Lin, Z. Gao, Z. Lv, W. Wei, H. Song, Quality index for stereoscopic images by separately evaluating adding and subtracting, PloS ONE 10 (12) (2015b) E0145800.
[46] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, Inf. Sci. 385 (2017) 338–352.
[47] J. Yang, Y. Wang, B. Li, W. Lu, Q. Meng, Z. Lv, D. Zhao, Z. Gao, Quality assessment metric of stereo images considering cyclopean integration and visual saliency, Inf. Sci. 373 (2016) 251–268.
[48] M. Yu, K. Zheng, G. Jiang, F. Shao, Z. Peng, Binocular perception based reduced-reference stereo video quality assessment method, J. Vis. Commun. Image Represent. 38 (2016) 246–255.
[49] H. Zhu, M. Yu, Y. Song, G. Jiang, A stereo video quality assessment method for compression distortion, in: Proceedings of the IEEE International Conference on Computational Science and Computational Intelligence (CSCI), 2015, pp. 481–485.