This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

No-reference quality assessment of stereoscopic videos with inter-frame cross on a content-rich database

PLEASE CITE THE PUBLISHED VERSION

<https://doi.org/10.1109/TCSVT.2019.2948383>

PUBLISHER

IEEE

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Yang, Jiachen, Yang Zhao, Bin Jiang, Qinggang Meng, Wen Lu, and Xinbo Gao. 2019. "No-reference Quality Assessment of Stereoscopic Videos with Inter-frame Cross on a Content-rich Database". Loughborough University.

No-reference Quality Assessment of Stereoscopic Videos with Inter-frame Cross on a Content-rich Database

Jiachen Yang, *Member, IEEE*, Yang Zhao, Bin Jiang, Qinggang Meng, *Senior Member, IEEE*, Wen Lu, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*,

Abstract—With the wide application of stereoscopic video technology, the quality of stereoscopic video has attracted people's attention. Objective stereoscopic video quality assessment (SVQA) is highly challenging, but essential, particularly the no-reference (NR) SVQA method, where reference information is not needed and a large number of samples are required for training and testing sets. However, as far as we know, there are only a few samples in the established stereo video database, which is unsuitable for NR quality assessment and seriously hampers the development of NR-SVQA method. For these difficulties that we encountered, we carry out a comprehensive subjective evaluation of stereoscopic video quality in our newly established TJU-SVQA databases that contain various contents, mixed resolution coding and symmetrically/asymmetrically distorted stereoscopic videos. Furthermore, we propose a new inter-frame cross map to predict the objective quality scores. We compare and analyze the performance of several state-of-the-art 2D and 3D quality evaluation methods on our new databases. The experimental results on our established databases and a public database demonstrate that the proposed method can robustly predict the quality of stereoscopic videos.

Index Terms—Stereoscopic video quality assessment, inter-frame cross, TJU-SVQA database.

I. INTRODUCTION

WITH the rapid development of 3D video technology, the demand for 3D contents can be expected to become increasing in the near future. Stereoscopic videos have such promising application on account of their vivid stereo perception and excellent immersive experience [1]. Thus, the exploitation of subjective and objective visual quality models have become a hot research field [2]. However, there are some specific challenges to assessment of 3D video quality, in addition to the inevitable distortions in the transformation process which have already appeared in 3D image [3]. The main reasons are video contains a large amount of information

[4] and a more complex content than image, and the results of the prediction performance is difficult to achieve.

Research on VQA can be generally divided into two types: subjective and objective quality evaluation [5]. Subjective quality evaluation is to ask a sufficiently great number of subjects to assess the degree of distortion of image/video according to their viewing experience. And then averaging across these opinions produces a mean opinion score (MOS), which can represent the real viewing experience of the tested image/video [6]. In fact, such subjective assessment is the best indicator of perceived quality of image/video. At present, there are several general and freely available 3D Video Quality Assessment (3D-VQA) databases [7]–[9]. But these databases contain fewer videos and have a single video content. Especially with the arrival of the big data era and the extensive application of the 3D video display technologies, it is urgent to establish a SVQA database which involves a great number of samples and rich content.

Similar to image quality assessment (IQA), objective SVQA methods are generally classified into three types depending on requirements for reference videos. The pristine content of the full reference (FR) metrics is necessary for the comparison with the test content. And reduced-reference (RR) metrics for measuring quality require partial content of the pristine information, while NR evaluation metrics measure of quality only require processed/distorted information. Classical objective quality schemes, such as peak signal to noise ratio (PSNR) and structural similarity (SSIM) [10] are FR methods, but their results can not be consistent with subjective evaluation. To overcome this difficulty, new objective algorithms with stereo vision are being exploited. Hewage *et al.* [11] introduced a 3D video quality algorithm by assessing stereo pairs and analyzing the edges and contours of the depth map. Due to the characteristics of stereoscopic video, the quality is comprehensively evaluated from the perspectives of multiple domains. The experiment results demonstrate that this NR-SVQA metric is highly consistent with subjective quality evaluation [12].

SVQA has great challenges, for it has higher complexity than 2D-VQA. Due to the pupillary distance and slightly different visual angles of two views, there is a horizontal difference in the scene of two retinas and human eyes can recognize the depth of the scene. On the one hand, the additional dimension of 3D information (depth or disparity) will raise a number of important issues. On the other hand, the motion

This work was partially supported by National Natural Science Foundation of China (No. 61871283), Foundation of Pre-Research on Equipment of China (NO.61403120103) and Major Civil-Military Integration Project in Tianjin, China (NO.18ZXJMTG00170). (Corresponding author: Bin Jiang.)

J. Yang, Y. Zhao and B. Jiang are with School of Electrical and Information Engineering, Tianjin University, Tianjin, P.R. China (e-mail: yangjiachen@tju.edu.cn; yangzhao321@tju.edu.cn; jiangbin@tju.edu.cn).

Q. Meng is with the Department of Computer Science, Loughborough University, UK (email: q.meng@lboro.ac.uk).

W. Lu is with School of Electronic Engineering, Xidian University, Xi'an, P.R. China (e-mail: luwen@mail.xidian.edu.cn).

X. Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an, China (e-mail: xbgao@mail.xidian.edu.cn).

of videos will also lead to distortion in temporal domain. In this paper, a content-rich SVQA database is constructed and a NR-SVQA metric with inter-frame cross information is put forward. The main contributions of this paper are as follows:

1) A content-rich SVQA database is established with two parts: TJU-SVQA database Phase I and Phase II, and it contains symmetric and asymmetric compression videos created by varies content and mixed-resolution, respectively. And we analyze the subjective experiment results to prove the stability and rationality of the database.

2) We present a binocular characteristic inspired model, which leads to a markedly improved NR-SVQA model. This method uses a new infer-frame cross to effectively reflect the relationship between video distortion degree and the MOS, and increases efficiency.

3) According to the property of the extracted features, we adopt different methods for different views to obtain local quality score. We first combine dictionary learning with support vector regression (SVR), which is utilized in spatial, spatio-temporal and temporal domain. The experimental results show that our framework performs better than using dictionary learning or SVR alone.

The rest of the paper is organized as follows. The existing models of 2D/3D-VQA are reviewed in Section II. Section III describes the details of TJU-SVQA database, including scene choice and descriptions, subjective research project, and the analysis of subjective evaluation result. Then in Section IV, an objective SVQA model considering binocular characteristic is proposed. Section V reports the experimental results of the proposed metric on our TJU-SVQA database and a public database. The analysis shows the availability of the TJU-SVQA database and the effectiveness of the proposed algorithm. Finally, in Section VI, conclusion and ideas for the optimization of the algorithm are presented.

II. RELATED WORK

Existing video quality evaluation mainly includes 2D-VQA and 3D-VQA based on different information sources. Besides, according to the design concept of 3D quality assessment algorithm it can be divided into two categories: one of which is directly extended from 2D and the other 3D-VQA based on human visual system (HVS).

A. 2D Video Quality Assessment

In recent years, the theme of 2D quality evaluation has been extensively explored and studied [13]. Objective VQA methods are used to automatically predict the quality of the visual perception of the video sequence [14], while striving to be consistent with the subjective perception of the human eye. Some eminent FR-VQA methods include MOVIE [15], VQA based on SSIM [16], VADM [17], optical flow-based VQA [18]. Moreover, to imitate the processing of video signal by HVS, video sequences energy characteristics were described by statistical features with 3D discrete cosine transform [19]. Compared with RR-VQA, NR-VQA is much more challenging but has higher potential possibility of wide application. Saad *et al.* [20] proposed a NR-VQA (video BLIINDS) which

indicates that structure a spatio-temporal model in the DCT domain and exhibited statistical regularities. Mahttal *et al.* [21] developed a completely blind VQA VIIDEO (video intrinsic integrity and distortion evaluation oracle), which demands not the supererogatory distortional video information but the intrinsic statistical regularities of natural videos. Considering that different features have different sensitivity to high quality video and the inferior quality, noticeable distortion and blurring artifacts were applied comprehensively to guarantee accurate prediction [22]. In [23], intensity and texture of decoded sequence were combined as complementary information to measure duration and the stall number of distorted streaming video. Li *et al.* [24] extracted the natural scene statistical characteristics of video blocks with 3D shearlet transform and amplified disparity between primary features to effectively perceive quality with convolutional neural network.

B. 2D-to-3D Quality Assessment

The history of 3D-VQA is relatively short when being compared with that of 2D-VQA. First of all, there are few publicly stereoscopic video databases in the past few years, which greatly hinders the study of subjective evaluation algorithms. Recently, a variety of 3D video datasets have emerged according to different researching purposes and requirement [25]. A brief of most of the SVQA databases can be found in [7]. However, those existing databases have a small amount of samples, together with single video content and resolutions.

According to algorithm design principles, the existing SVQA schemes could be classified as two categories. An earlier type of SVQA is extended from 2D image/video by averaging or weighting stereo pairs quality scores. In [25], Chen *et al.* predicted the content quality of stereo pair sequences on a weighted average with PSNR and multiscale SSIM (MS-SSIM). In [26] and [27], IQA approaches (PSNR, SSIM) and 2D-VQA approaches VQM [28] were implemented on the two views of stereo video respectively and then taken weighted average to obtain stereoscopic perceived quality. Experimental results demonstrated that VQM was more accurate than PSNR and SSIM. Due to the obvious phenomenon of binocular competition in asymmetric distortion, a new dynamic weight scheme was proposed to predict system bias and significantly improve the performance of SVQA [7].

C. 3D-VQA based on HVS

The other category of 3D-VQA metrics are to estimate the impact of compression distortion on visual perception from the perspective of stereo vision. In [29], a new binocular energy framework with suppression and recurrent excitation was presented in which multivariate regression method was applied to map the quality perception features to MOS. For improving the performance of the algorithm, Yu *et al.* [30] introduced a new frame extraction algorithm by judging the motion intensity to reduce the volume data processing. In [31], both views were represented by constructing a joint descriptor with the spatial-temporal structure metric and inter-view correlation was considered between 3D structure tensors of two views. In [32], regions of significance based on human perception

were extracted. In their framework, region of interest and just noticeable distortion (JND) were used to measure two views quality and depth information was estimated in 3D wavelet domain. In [33], Jin *et al.* presented a FR-SVQA based on 3D-DCT transform to stack the matched two views blocks in the stereo video sequence to form a 3D stack and then analyzed the visual similarity through 3D-DCT. Inspired by the Bayesian model, Shao *et al.* [34] decomposed the synthesized view into three terms and quantified the impact of color and depth distortions and their interactions to predict their synthesized videos quality. Recently, a NR-SVQA method, the depth perception quality metric (DPQM), has been developed with depth perception by Chen *et al.* [8]. The natural scene statistic was employed in fusion map of binocular summation and auto-regressive prediction based disparity entropy (ARDE) was applied in suppression map of binocular difference, which indicated key factors affecting texture and parallax.

III. SUBJECTIVE STEREO VIDEO QUALITY ASSESSMENT DATABASE CREATION

The new SVQA database is created and we call our new stereoscopic video database Tianjin University-Stereoscopic Videos Quality Assessment database (TJU-SVQA database). We first describe the choice of source stereo videos, and then conduct the subjective test on the built database. Finally we discuss the subjective experimental results.

A. Source Video Selection

To build new stereoscopic video database, 20 pristine stereoscopic video sequences were collected from pre-existing subjective 3D video quality researches [9], [35], [36]: Barrier gate, Basket, Boxers, Hall, Lab, News report, Phone call, Soccer, Tree branches, Umbrella, Street, Champagne, Pantomime, Dog, Sofa, Classroom, Bike, Car, Balloons, and Kendo. All these source video sequences are with 25 fps frame rate and in uncompressed .AVI format. All the 20 source videos are used in Phase I and Phase II. Sample frames for each source video are given in Fig. 1. The details of all the source videos are exhibited in Table I.

The selection of the original video plays an essential part in the quality evaluation of stereo video. The original videos should try to contain all the possible scenes in real life. For the universality and completeness of our video database, the selection of the source video is basically based on the video resolution, scene content, motion intensity and background complexity. The specific information of the source video is shown in Table I. Compared with the existing SVQA databases, the video resolution of TJU-SVQA databases are more abundant. The source videos contain eleven 1920×1080 videos, four 1280×960 videos, three 1280×720 videos, and two 1024×768 videos. From the point of scene content, there are 6 outdoor scene videos and 14 indoor scene videos. From the point of motion intensities, of the 20 original videos, there are 8 videos with large motion intensity, 8 videos with moderate motion intensities, and 4 videos with small motion intensities. From the point of the background complexity, there are 13 videos that have simple backgrounds and 7 Videos that

TABLE II: Degradation type in TJU-SVQA database Phase I.

Degradation type	Control parameter	File	Parameters	Number	Total
None	None	SYD_00	None	20	20
H.264	QP	SYD_01	QP30	20	80
		SYD_02	QP35	20	
		SYD_03	QP40	20	
		SYD_04	QP45	20	
JPEG2000	Bit-rate	SYD_05	2Mb/s	20	60
		SYD_06	4Mb/s	20	
		SYD_07	8Mb/s	20	

have complex backgrounds. From the point of shooting distance, there are 12 videos taken at close distance (the shooting distance is less than 5 meters), 5 videos taken at medium distance (the shooting distance is between 5 and 10 meters) and 3 videos taken at long distances (the shooting distance is more than 10 meters). Therefore, our video database contains rich video content, which makes the versatility of this video library better. The objective quality evaluation model based on our database has better universality in real life. In addition, existing stereo video databases have a small sample size, with most databases containing less than 10 original videos and about 100 distorted videos [9], [34], [37]. However, most of NR-SVQA algorithms need a great amount of samples to train the prediction models, so as to accurately obtain the test video quality scores. The TJU-SVQA database has larger sample capacity and is more suitable for NR-SVQA methods.

B. Distortion Simulation

The source sequences (SRC) from TJU-SVQA were impaired by two compress degradations. H.264 damages are introduced through H.264/AVC video coder [38], and JPEG2000 still image coder is impaired through Kakadu software, v7.0 [39]. TJU-SVQA database Phase I only includes symmetrically distorted videos. In Phase I, the two views are encoded separately based on the same parameters. For H.264 compression, it is compressed by four levels of quantization with $QP=\{30, 35, 40, 45\}$. For JPEG2000 compression, it was compressed by three levels of bitstream with $Bitrate=\{2Mb/s, 4Mb/s, 8Mb/s\}$. The details of Phase I are listed in Table II, eight symmetrically distorted conditions (SYD_00-SYD_07), including the unprocessed reference as SYD_00, and each condition has 20 stereo pairs. Therefore, there are totally 160 stereo videos in Phase I.

TJU-SVQA database Phase II only includes asymmetrically distorted videos. In Phase II, stereo pairs are encoded separately based on different parameters. The details of Phase II are listed in Table III, thirty-three different kinds of asymmetrically distorted conditions (ASD_00-ASD_32), including the unprocessed reference as ASD_00. Therefore, there are totally 660 stereo videos in Phase II.

Fig. 1: The 20 reference videos utilized in the subjective research. Only the left-views are displayed here. The stereoscopic videos contain both outdoor scenes and indoor scenes.

TABLE I: The detailed description of test videos in TJU-SVQAD Phase I and Phase II.

	Sequences	Resolution	Length(s)	Frame/s	Condition	Shooting conditions	Motion intensity	Background
1	Barrier gate	1920×1080	16	25	Outdoor	Medium	Medium	Clutter
2	Basket	1920×1080	16	25	Indoor	Long	Strong	Clutter
3	Boxers	1920×1080	16	25	Indoor	Close	Strong	Clutter
4	Hall	1920×1080	16	25	Indoor	Long	Weak	Clutter
5	Lab	1920×1080	16	25	Indoor	Close	Medium	Clutter
6	News report	1920×1080	16	25	Indoor	Close	Weak	Plain
7	Phone call	1920×1080	16	25	Indoor	Close	Medium	Plain
8	Soccer	1920×1080	16	25	Indoor	Long	Strong	Clutter
9	Tree branches	1920×1080	16	25	Outdoor	Medium	Strong	Clutter
10	Umbrella	1920×1080	13	25	Outdoor	Close	Medium	Clutter
11	Street	1920×1080	10	25	Outdoor	Medium	Weak	Clutter
12	Champagne	1280×960	16	25	Indoor	Close	Medium	Plain
13	Pantomime	1280×960	16	25	Indoor	Close	Strong	Plain
14	Dog	1280×960	12	25	Indoor	Close	Strong	Clutter
15	Sofa	1280×960	10	25	Indoor	Close	Medium	Plain
16	Classroom	1280×720	16	25	Indoor	Close	Medium	Plain
17	Bike	1280×720	10	25	Outdoor	Medium	Weak	Clutter
18	Car	1280×720	10	25	Outdoor	Medium	Weak	Plain
19	Balloons	1024×768	16	25	Indoor	Close	Strong	Clutter
20	Kendo	1024×768	16	25	Indoor	Close	Strong	Clutter

TABLE III: Degradation type in TJU-SVQA database Phase II.

Degradation type	Control parameter	File	Parameters	Number	Total
None	None	ASD_00	Ref-Ref	20	20
H.264	QP	ASD_01-ASD_04	Ref-QP30, Ref-QP35, Ref-QP35, Ref-QP35	20×4	400
		ASD_05-ASD_08	QP30-Ref, QP30-QP35, QP30-QP40, QP30-QP45	20×4	
		ASD_09-ASD_12	QP35-Ref, QP35-QP30, QP35-QP40, QP35-QP45	20×4	
		ASD_13-ASD_16	QP40-Ref, QP40-QP30, QP40-QP35, QP40-QP45	20×4	
		ASD_17-ASD_20	QP45-Ref, QP45-QP30, QP45-QP35, QP45-QP40	20×4	
JPEG2000	Bit-rate	ASD_21-ASD_23	Ref-BR2, Ref-BR4, Ref-BR8	20×3	240
		ASD_24-ASD_26	BR2-Ref, BR2-BR4, BR2-BR8	20×3	
		ASD_27-ASD_29	BR4-Ref, BR4-BR2, BR4-BR8	20×3	
		ASD_30-ASD_32	BR8-Ref, BR8-BR2, BR8-BR4	20×3	

C. Subjective Test

Our subjective experiments were conducted in accordance to ITU-R BT. 500. The subjective test was conducted in the Lab for Image and Video Processing at Tianjin University. The evaluation was performed by using a HD 3D TV with passive polarized glasses. A CHANGHONG 3D55C2000i monitor with NVIDIA GTX 1080 graphics was used for testing equipment. The viewing distance was set to be 6 times longer than the screen height, and that is about 4.7m. The details of the viewing conditions are given in Table IV.

In Phase I, nineteen subjects participated in our subjective study with ages between 18 and 38. In Phase II, twenty-nine subjects participated in our subjective study with ages between 18 and 40. Before the subject evaluation, a visual acuity test and a 3D vision test was performed to examine their capacity to view stereo content.

In the phase of subjective experiment, in order to distinguish the quality of original videos that have content scenes, we opt to the single stimulus procedure using a 5-grade numerical categorical scale (SSNCS) protocol. A general introduction was given at the beginning of the whole test, including the purpose, process and scoring criteria of the experiment. The participants were required to evaluate the video quality according to their overall 3D viewing experience. In the training phase: for both Phase I and Phase II, we first selected three categories of original videos, intermediate damaged videos and severely damaged videos from the stereoscopic video database as training sets, and then played these videos and explained the evaluation criterion to observers. The observers were asked to give high scores (close to 5) to the original stereo video sequences, the mediate scores corresponded to the moderately distorted sequences, and low scores to severely distorted 3D videos. Subsequently, several videos were played at random to test each observer, until ensuring that observers can fully understand the scoring criteria and establish their own scoring strategies.

In the formal subjective test phase, for the purpose of reducing the time required for the subjective evaluation to ensure comfort duration, we divided the video databases into several sessions. And each session contains all the stimuli. In Phase I, the 160 video sequences were divided into 4

TABLE IV: Viewing conditions in the subjective test.

Parameter	Value
Graphics	NVIDIA GTX 1080
Display version	CHANGHONG 3D55C2000i
Screen Resolution	1920×1080
Screen size	55inch(1.249×0.7815m)
Screen aspect ratio	16:9
Viewing Distance	4.7m

sets, and each set contains 5 pristine and 35 corresponding degraded sequences (40 sequences per test). While in Phase II, the 660 video sequences were divided into 10 sets, and each set contains 2 pristine and 64 corresponding degraded sequences (66 sequences per test). When the video was played, the order of stimuli was randomized. When the video ended, the participants were requested to speak out a score between 1 and 5, high score correspond to high quality. Meanwhile, the instructor recorded the quality scores with another computer. To prevent visual fatigue and discomfort of participants, each subject's subjective test time is not more than 2 hours per day. After all the subjective results were collected, the impact of individual misjudgment on video quality was excluded (within 95% confidence interval) and the MOS values for each of the videos were calculated.

In addition, to demonstrate the consistency between the prediction of the individual scores and MOS, two indicators generally used to evaluate the fitting degree of subjective and objective values are employed: Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SROCC). Among those parameters, the higher the values of PLCC and SROCC are, the better the algorithm performance is. In particular, values PLCC and SROCC approaching 1 indicate higher accuracy and stronger monotonicity. The data of PLCC and SROCC associated with each subject was obtained, and the average values of PLCC and SROCC are shown in Table V, where the high average indicator present a good consistency between individual score and MOS.

TABLE V: The average PLCC and SROCC between individual subjective quality score and MOS.

	TJU-SVQA Phase I	TJU-SVQA Phase II
PLCC	0.9072	0.8689
SROCC	0.8967	0.8669

D. Results and Analysis

The indicator of the rationality of subjective evaluation research is uniformly distributed MOS scores. A histogram of MOS values is given in Fig. 2. It can be observed that the distribution ratios of MOS both in Phase I and Phase II are evenly spaced in the range of 1 to 5, which indicates that the MOS distortions in the TJU-SVQA dataset have a wide range of visual perception quality, which can be considered representative. Furthermore, original videos with different

Fig. 2: Subjective study data distribution in the TJU-SVQA database (a) Histogram of MOS scores of Phase I (b) Histogram of MOS scores of Phase II.

content scenes usually bring different viewing experiences to the audience. And the same distortion type and distortion degree usually cause various damage to the original videos with different content scenes. Therefore, we display the MOS distribution of the original videos with different content scenes in Fig. 3, and the MOS distribution of the same distortion on the original videos with different content is shown in Fig. 4. Fig. 3 shows that the quality scores of most original video sequences are within the range of 4 to 5, except for Dog and Lab, whose scores are close to 4 (3.95 and 3.90, respectively). Accordingly, the video quality of TJU-SVQA original sequences could be considered competent in large-scale assessment activities. In addition, Fig. 4 also plots the MOS distribution of the same distortion on the original videos with different MOS. To some extent, the reduction of the compressed videos are consistent with the human eyes' quality perception under the same type of distortion.

IV. THE PROPOSED 3D-VQA ALGORITHM

According to human visual perception, we propose a comprehensive SVQA method based on joint contribution of multiple domain information, and a new inter-frame cross about spatio-temporal information is presented. The framework of the proposed algorithm is displayed in Fig. 5. In particular, the

Fig. 3: MOS value of the original sequences on Phase I.

Fig. 4: MOS value of the same distortion on Phase I.

binocular summation/difference and inter-frame cross maps are produced firstly, and then the quality perception features from the above sequences are extracted. Since optical flow is sensitive to distortion, optical flow features are extracted in temporal domain to measure the degree of distortion. Finally, the quality scores of each part are obtained through dictionary learning and SVR, and pooled into the global objective quality.

A. Spatial Domain Analysis

In terms of HVS peculiarity, binocular vision could operate in several different 'modes' [40]. According to the Li and Atick's theory [41] of efficient stereo coding, binocular signals are encoded to be irrelevant to binocular summation signal and difference signal, and gain control is performed on the summation and difference channels to optimize their sensitivities. The summation map of distorted stereo video contains both additive impairments and detail losses. Detail losses refers to the damage of valuable information due to distortion, while additive impairments mean redundant information which does not show up in the pristine signal [42], [43]. Moreover, due to the distance between human eyes, two views have horizontal parallax and depth perception is formed. As an alternative absolute disparity map, difference map covers the distinction between two views and can reflect the depth information of stereo vision [8]. Given the left view L and right view R , the binocular summation and difference maps could be computed frame-by-frame as follows:

$$I_{Summation} = \frac{I_{L_i}}{2} + \frac{I_{R_i}}{2} \quad (1)$$

$$I_{Difference} = |I_{L_i} - I_{R_i}| \quad (2)$$

Fig. 5: Framework of the proposed methodology.

where i is the number of frame of the stereo video sequences.

The purpose of feature vector extraction is to extract the quality perception features that effectively indicate the distortion degree of stereoscopic video. The spatial features are firstly obtained in summation/difference maps. The human eye has selectivity of orientation and frequency in its perception of stimulus, and therefore the various scale and orientation sensitivities of the receptive fields can be simulated by the statistics of multi-scale and multi-directional filtering responses [44]. Since Log-Gabor filters can avoid direct current interference and bandwidth limitations, we utilize them for feature extraction and the kernel of the Log-Gabor filters used in this study can be generated by

$$L_{j,l}(\omega, \theta) = \exp\left(-\frac{\log^2(\omega/\omega_0)}{2\sigma_\gamma^2}\right) \cdot \exp\left(-\frac{(\theta - \theta_l)^2}{2\sigma_\theta^2}\right) \quad (3)$$

where ω_0 denotes the center frequency, $\theta_l = l\pi/L$, $l = \{0, 1, \dots, L-1\}$ is the orientation angle, and σ_θ and σ_γ control the angular and radial bandwidth of the filter respectively. In this paper, the Log-Gabor filter parameters are regreded as follow: $\sigma_\theta = 0.40$, $\sigma_\gamma = 0.65$ and $\omega_0 = 1/\min wavelength * mult^{j-1}$, where $\min wavelength = 3$ and $mult = 1.7$ (the scaling factor between successive filters). The local magnitude as spatial information at location on different scale with different orientation can be given by

$$m(j, l, k_1, k_2) = \sqrt{R_{j,l}(k_1, k_2)^2 + \text{Im}_{j,l}(k_1, k_2)^2} \quad (4)$$

where (k_1, k_2) represent coordinates in the spatial domain, $j = \{1, \dots, J\}$ and $l = \{1, \dots, L\}$ express the scale and orientation, which are respectively set to $J = 4$ and $L = 6$. $R_{j,l}(k_1, k_2)$ and $\text{Im}_{j,l}(k_1, k_2)$ represent real part and imaginary part which constitute Log-Gabor. In Oppenheim and Lim's study [45], the high degree of intelligibility image was synthesized from the Fourier transform magnitude of an image

and the phase of another image, which strongly resembles the one with the same phase. As the human eye is more sensitive to the phase information of the image [46], the phase information of each frame is also extracted by Log-Gabor filter as spatial information. In order to extract the local phase at location (k_1, k_2) on different scale j with different orientation l is given by

$$h(j, l, k_1, k_2) = \arg \tan \left(\frac{R_{j,l}(k_1, k_2)}{\text{Im}_{j,l}(k_1, k_2)} \right) \quad (5)$$

It is well known that the primary visual area (V1) is the first region in visual cortex to directly receive visual signals from the retina. The local semantic information formed in the V1 is strongly correlated with the perceived quality, and the decline of visual quality can be clearly reflected in local information changes. Therefore, after the computation of the magnitude and the phase information, an improved rotational invariance LBP ($LBP_{P,R}^{riu2}$) [47] is applied on the extracted and effectively studied local structure descriptors.

The basic patterns (the number of changes of adjacent codes is less than or equal to 2) of $LBP_{P,R}^{riu2}$ are uniform patterns with very little spatial variation. Due to the high proportion of basic patterns in the process of rotation invariant binarization coding, $LBP_{P,R}^{riu2}$ encodes nine basic patterns respectively and categorizes the rest patterns into one code as follows:

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(z_p - z_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P + 1 & \text{otherwise} \end{cases} \quad (6)$$

where uniform measure is defined as

$$U(LBP_{P,R}) = |s(z_{P-1} - z_c) - s(z_0 - z_c)| + \sum_{p=1}^{P-1} |s(z_p - z_c) - s(z_{p-1} - z_c)| \quad (7)$$

The superscript $riu2$ in Eq. 6 indicates that the rotation invariant uniform patterns have U at most 2. z_c represents

TABLE VI: Summary of features extracted in spatial domain.

Channel	Feature Type	Feature Dimension
Summation	Global Feature	$f_{SD,1}, \dots, f_{SD,3}$
	Magnitude	$f_{SD,4}, \dots, f_{SD,13}$
	Phase	$f_{SD,14}, \dots, f_{SD,23}$
Difference	Global Feature	$f_{SD,24}, \dots, f_{SD,26}$
	Magnitude	$f_{SD,27}, \dots, f_{SD,36}$
	Phase	$f_{SD,37}, \dots, f_{SD,46}$

the central magnitude/phase, and z_p is the magnitude/phase of the neighborhood. According to the above equation, there are a total of $P + 2$ patterns in $LBP_{P,R}^{riu2}$, which are coded from 0 to $P + 1$ respectively. We selected $P = 8$ points as the neighborhood and determined the range of the interest to be $R = 1$. After completing $LBP_{P,R}^{riu2}$ coding, the occurrence times of the various patterns was counted as:

$$H(k) = \sum_{i=1}^N \sum_{j=1}^M f(LBP_{P,R}^{riu2}(i, j), k), \quad k \in [0, K] \quad (8)$$

where the input of the $LBP_{P,R}^{riu2}$ operator is an $N \times M$ matrix and K is the maximum pattern of $LBP_{P,R}^{riu2}$,

$$f(x, y) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The statistical characteristics of the various patterns are regarded as the final magnitude/phase-based features. In addition, the magnitude, variance, and entropy of stereoscopic video are considered as auxiliary global features for the spatial domain. The spatial features are listed in Table VI.

B. Spatio-temporal Domain Analysis

Stereoscopic perception is based on binocular disparities. One scene with slightly different perspectives is sent to the lateral geniculate nucleus (LGN) through retinal ganglion cells, and some layers of the primary visual cortex (V1) receive information from the LGN. However, when the frames of two view at the same retinal location do not match, binocular rivalry, a perceptual effect in which perception alternates between different images presented to each eye, will occur. At the same time, binocular rivalry is not limited in V1, and continue to be in the higher and more abstract level of the visual pathway. When binocular rivalry occurs, two images are not seen overlapped, but perceived alternately at intervals [48]. Binocular rivalry can spread around the highly distorted and low correlated regions [49], and it can make visual processing unstable, unpredictable, and impair the ability of observers to direct attention to targets in the visual field [50].

Binocular rivalry and stereopsis are two core phenomena of binocular vision, which involve independent and parallel pathways in the early stages of visual processing. Moreover, binocular is a simple combination of the output of the pathways mediating stereopsis and binocular rivalry [51]. In the previous subsection, binocular summation and difference maps

TABLE VII: Summary of features extracted in spatio-temporal domain.

	Feature Type	Feature Dimension
Inter-frame cross	Global Feature	$f_{C,1}, \dots, f_{C,3}$
	Magnitude	$f_{C,4}, \dots, f_{C,13}$
	Phase	$f_{C,14}, \dots, f_{C,23}$

of spatial domain are established according to the stereopsis phenomenon. Here, a quality prediction model of spatio-temporal domain is proposed according to binocular rivalry phenomenon.

In the process of research, one underlying assumption is proposed that besides spatial and temporal information, human perception of stereoscopic video quality is also influenced by the inter-frame information of time transformation (as spatio-temporal information) [15]. The traditional analysis method of spatio-temporal information is exploring the relationship between adjacent frames respectively in the left and right video without consideration of the possible interactions between two views. Based on binocular rivalry, a new inter-frame difference is proposed by extracting frames from left and right sequences of stereo pairs alternately along the time axis, which we call inter-frame cross. Concrete implementation is to combine the first frame of the left video and the second frame of the right video to extract the spatio-temporal information of the stereoscopic video. Then the second frame of the right video is combined with the third frame of the left video. By this analogy, inter-frame cross formed. Moreover, stereoscopic video contains much more information than 2D video/image, so the computational efficiency is a major challenge for SVQA. As is shown in Fig. 6, in contrast to the previous inter-frame difference method, inter-frame cross will combine the left and right video with the same amount of data as left (right) video, which can reduce the computation time by half and obtain the same amount of information. The inter-frame cross (I_C) is given by

$$I_{C_i} = I_{R_{i+n}/L_{i+n}} - I_{L_i/R_i} \quad (10)$$

which means that the i^{th} frame of the left (right) view alternates with the $i + n^{th}$ frame of the right (left) view, and n is a constant.

The local features distribution of inter-frame cross are illustrated in Fig. 7. Similar to the spatial domain, the magnitude, variance, and entropy of stereoscopic video are considered as auxiliary global information for the spatio-temporal domain. The spatio-temporal features are listed in Table VII.

Furthermore, when light of the image for one eye leaks to the other eye, perceptual crosstalk, the significant factor that degrades the perceived quality of stereoscopic images/videos, might be visible. Crosstalk could occur with both spatial multiplexing and temporal multiplexing [52]. Based on the existing research in [53], we speculate that the proposed inter-frame cross model might be able to reflect perceptual crosstalk to some extent. In the future research, we will conduct further research on this.

Fig. 6: The forming process of inter-frame cross.

Fig. 7: Illustration of the validity of the spatio-temporal features. The x-axis represents each pattern of $LBP_{P,R}^{riu2}$ operator, and the y-axis represents the frequency of each pattern. (a) Scatter plot of phase features of inter-frame cross vary with different MOS. (b) Scatter plot of magnitude features of inter-frame cross vary with different MOS.

C. Temporal Domain Analysis

Because the structures of stereoscopic video frames change along the timeline, stereoscopic video contains not only the structure and depth information in the spatial domain but also temporal information. Therefore, the objective scores are inconsistent with MOS when the 3D-IQA methods are directly applied to predict stereoscopic video quality with merely referring to the spatial distortion. Stereoscopic video quality can be represented by quality perception features involving spatial domain and temporal domain.

An significant part of SVQA is the analysis of the temporal information distortion and most of the previous approaches utilized the motion magnitude as the temporal feature to measure the temporal distortion directly. In order to relate the movement of brightness patterns to the movement of the frames, the optical flow metrics are applied to acquire the motion information between tow frames which are taken at times t and $t + \Delta t$ at every voxel position. In our studies, the Horn-Schunck algorithm [54] is proposed to gain the motion vectors in the temporal domain and its constraint equation is given by

$$I_x v_x + I_y v_y + I_t = 0 \quad (11)$$

TABLE VIII: Optical flow derivative.

Optic flow component	Feature
$v = (v_x^2 + v_y^2)^{\frac{1}{2}}$	Velocity
$div(\mathbf{v}) = I_x v_x + I_y v_y$	Scaling
$rot(\mathbf{v}) = I_x v_y - I_y v_x$	Rotation
$shA(\mathbf{v}) = I_x v_x - I_y v_y$	Shearing
$shB(\mathbf{v}) = I_x v_y + I_y v_x$	

where $I(x, y, t)$ represents the intensity at the point (x, y) of frame at time t . Optical flow is smooth in undistorted natural images but loses its smooth properties when distortion occurs. In particular, the degree of distortion will affect the magnitude and orientation of the optical flow.

The statistical regularities of natural stereo videos are correlated with video quality not only in the spatial domain but also in temporal domain. Based on the premise that local optical flow statistics are affected by distortions and the deviation from pristine flow statistics, the temporal statistical characteristics [9] based on optical flow are introduced to analyze the degree of temporal distortion. To describe the visual impact of distortion on video, the optical flow is divided into several components as sample ξ_i shown in Table VIII. According to two assumptions in [55], statistical methods of the optical flow component samples are as follows:

$$\varphi[\xi] = \frac{\|\sum \xi_i\|}{\sum \|\xi_i\|} \quad (12)$$

$$\Lambda[\xi] = \frac{(\sum \|\xi_i\|)^2}{N \sum \|\xi_i\|^2} \quad (13)$$

where ξ_i is N dimensions. The optical flow features $\varphi[\xi]$ and $\Lambda[\xi]$ estimate the probabilities of having characteristic orientation and characteristic speed of the above components respectively [55]. The temporal features of stereo video are obtained by averaging the statistics of each frame and listed in Table IX.

TABLE IX: Summary of features extracted in temporal domain.

	Feature Type	Feature Dimension
Optical flow	Direction: $\varphi[\xi]$	$f_{T,1}, \dots, f_{T,5}$
	Speed: $\Lambda[\xi]$	$f_{T,6}, \dots, f_{T,10}$

D. The Overall Stereo Video Quality Evaluation

Similar to the simple cells and complex cells encoding stereoscopic information into non-redundant binocular retinal information, dictionary learning is utilized for basic quality feature vectors to eliminate redundant information from the underlying features. Instead of just using dictionary learning or SVR alone to obtain the quality score, here different schemes are used in different domains to estimate the stereo video quality. An exclusive overcomplete dictionary matrix D_{SD} for binocular summation/different channel is constructed by concatenating all the feature vectors to sparsely represent spatial domain, i.e., $D_{SD} = [f_{SD,1}, f_{SD,2}, \dots, f_{SD,n}] \in R^{m \times n}$, where $f_{SD,i}$ means the feature vectors of the i -th sample and m is the dimension of features. Similar to spatial domain, an exclusive dictionary matrix D_C of spatial-temporal domain is formed by spatio-temporal features ($f_{C,i}$), i.e., $D_C = [f_{C,1}, f_{C,2}, \dots, f_{C,n}] \in R^{r \times n}$, where r is the dimension of spatio-temporal features. Then the dictionary learning is straightforwardly used in features and the corresponding MOS of the training set as follows:

$$\begin{bmatrix} D_{SD} \\ Q \end{bmatrix} = \begin{bmatrix} f_{SD,1} & f_{SD,2} & \dots & f_{SD,n} \\ q_1 & q_2 & \dots & q_n \end{bmatrix} \quad (14)$$

where q_i denotes the human subjective score associated with $f_{SD,i}$. The similar dictionary learning model of spatio-temporal view can be obtained. In order to find the most approximate and sparsest representation of the feature of a test video, the l_2 -minimization based approach which is less complicated than l_1 -minimization has been used as

$$\alpha_{SD} = \arg \min_{\alpha_{SD}} \|f_{SD} - D_{SD}\alpha_{SD}\|_2 + \lambda \|\alpha_{SD}\|_1 \quad (15)$$

where λ is a regularization parameter to balance the fidelity term and the sparse regularization term. According to above steps, the optimal sparse solution vector α_{SD} of f_{SD} is obtained. On the basis of the assumption that the videos with the approximate quality scores have similar distributions of their features, the predicted result of the test video can be quantified with training video, i.e.,

$$Q_{SD} = \frac{\sum_{i=1}^n \alpha_{SD,i} q_i}{\sum_{i=1}^n \alpha_{SD,i}} \quad (16)$$

$$Q_C = \frac{\sum_{i=1}^n \alpha_{C,i} q_i}{\sum_{i=1}^n \alpha_{C,i}} \quad (17)$$

where Q_{SD} denotes the spatial quality and Q_C denotes the spatio-temporal quality.

Simultaneously, an individual SVR is used to obtain temporal quality (Q_T) scores. Since the temporal statistical feature vector dimensions extracted in our framework are small, it is

hard to accurately describe the temporal features of the test set by using dictionary learning. Therefore, SVR is applied to obtain the temporal quality scores. Specifically, the temporal feature vector of the training set randomly generated by the database and the corresponding MOS value is used to train SVR, and the trained model is utilized for predict the quality of the local part. Finally, the global stereoscopic video quality is predicted by integrating the above three parts, so the overall stereo video quality (Q_o) is computed as:

$$Q_o = \mu \cdot Q_{SD} + \eta \cdot Q_C + \gamma \cdot Q_T \quad (18)$$

where the constraint is $\mu + \eta + \gamma = 1$, and μ , η and γ are constant. The determination of weight parameters is mentioned in section V.

V. EXPERIMENTS

In addition to the TJU-SVQA database Phase I and Phase II, the presented NR-SVQA method is also executed on the NAMA3DS1-COSPAD1 stereo video database [9] to verify the accuracy and robustness of the algorithm. The NAMA3DS1-COSPAD1 database contains 100 symmetrically distorted stereo videos with 10 original 1920×1080 3D full HD stereo videos which span five distortion categories including: H.264/AVC, JPEG2K, reduction of resolution, sharpening and downsample with sharpening.

A. Algorithms and Performance

On the premise that the performance of the algorithm is not affected, to reduce the consumption of time, the algorithm uses the strategy of extracting one frame every four frames for processing. Thus, the constant in the inter-frame cross is defined as $n = 4$. During the training of the regression model, the database is randomly divided into non-interacting training and testing sets: 80% of the database is applied to train and the rest 20% to test. Consequently, two dictionaries of spatial domain and spatio-temporal domain are formed directly by using feature vectors with corresponding MOS for the training set and sparse representation for the testing set. A SVR is trained with temporal domain features in training set to predict the effect of distortion on the temporal domain in the test set. We executed 1000 repetitions of the train-test process, and the medians of two performance indicators PLCC and SROCC are applied to measure the ultimate algorithm accuracy. The capability of the employed algorithm is verified with several experiments and the outstanding results show that the predicted scores are highly consistent with the MOS.

To demonstrate the effectiveness of the presented algorithm, we compared it with five existing metrics on the TJU-SVQA database Phase I and Phase II, including two widely used 2D-IQA models (PSNR and SSIM), one 3D-IQA model (Cyclopean [55]) and two state-of-the-art 3D-VQA models (VQM [12] and DPQM [8]). For 2D models, PSNR and SSIM are estimated as the average of all frame scores for each view and two views' scores are averaged to obtain the final scores. For 3D-IQA model, all frame features are averaged to obtain the final quality result. Moreover, in NAMA3DS1-COSPAD1 database, four quality evaluation methods are also employed to verify the robustness of the presented algorithm.

TABLE X: The performance of three methods to get the final score on TJU-SVQA database Phase I and NAMA3DS1-COSPAD1 database.

Method	TJU-SVQA Phase I		NAMA3DS1-COSPAD1	
	PLCC	SROCC	PLCC	SROCC
DL	0.8508	0.8331	0.9204	0.8966
SVR	0.8635	0.8484	0.9166	0.8886
SVR+DL	0.8724	0.8602	0.9276	0.9028

TABLE XI: The performance of the predicted quality of individual domains on TJU-SVQA database Phase I and Phase II

Domain	TJU-SVQA Phase I		TJU-SVQA Phase II	
	PLCC	SROCC	PLCC	SROCC
Spatial	0.8584	0.8415	0.8955	0.8893
Spatio-temporal	0.8326	0.8251	0.8863	0.8794
Temporal	0.6261	0.5922	0.7830	0.7757
Overall	0.8724	0.8602	0.9055	0.8984

B. Parameter Optimization of Local Quality

To explore the characteristics of different visual features and optimize the performance of the algorithm simultaneously, we used the quantitative angle to analyze different prediction methods. In the process of predicting the final quality score, three schemes are considered, including SVR, dictionary learning (DL) and the combination of the two schemes (SVR+DL). Three regression schemes are explored respectively in TJU-SVQA database Phase I and NAMA3DS1-COSPAD1 database. As shown in Table X, we list the experimental results based on different predictors choices. From Table X, the presented metric (SVR+DL) achieves better performance on two evaluation criteria.

We further investigated the quality perception ability of each domain and explored the sensitivity of each domain to distortion. The quantitative results of the fitting degree between predicted scores of each domain and MOS are shown in Table XI. It can be observed that the quality of the spatial and spatio-temporal domains contributes similarly to the prediction and is more important than that of the temporal domain. Even so, temporal domain information is also valuable for quality evaluation and the local quality score of each domain does not exceed the composite quality score of three domains, which verifies the necessity of comprehensively considering the quality degradation of the three domains for visual quality perception.

The ability of global features, magnitude and phase to predict the quality was explored by taking the features of spatio-temporal domain as an example. The experimental results are displayed in Table XII. The results show that the prediction performance of composite features is better than that of separate type feature, which proves that each type of feature contributes to quality prediction, and the magnitude is the most sensitive to stereo video quality. A separate type of feature cannot comprehensively describe the quality of

TABLE XII: The performance of the predicted quality of each type of features in spatio-temporal domain TJU-SVQA database Phase I and Phase II

Feature	TJU-SVQA Phase I		TJU-SVQA Phase II	
	PLCC	SROCC	PLCC	SROCC
Global features	0.5082	0.4177	0.4571	0.4105
Magnitude	0.7206	0.6923	0.7091	0.6846
Phase	0.5996	0.5468	0.6735	0.6511
Overall	0.8326	0.8251	0.8863	0.8794

TABLE XIII: Performance comparison of inter-frame cross model with previous model in spatio-temporal domain on TJU-SVQA database Phase I and Phase II

Model	TJU-SVQA Phase I		TJU-SVQA Phase II	
	PLCC	SROCC	PLCC	SROCC
Previous model	0.7984	0.7638	0.8670	0.8472
Inter-frame cross model	0.8326	0.8251	0.8863	0.8794

stereo video. In order to ensure the prediction accuracy of the model, it is necessary to combine various quality-sensitive features that will not interfere with each other. In addition, the performance of inter-frame cross model is compared with that of the previous spatio-temporal domain model. The previous model here is to extract the above features from the inter-frame differences of left and right views, and the stereo video quality score is obtained by weighted average of the scores of two views. The results are shown in Table XIII. Since inter-frame cross model is proposed based on binocular rivalry phenomenon, it can more accurately reflect the quality perception process of human eyes. However, the previous model is not closely related to binocular perception, and its performance does not exceed that of the proposed model.

For purpose of determining the three parameters in the combination process, we assigned various values to one parameter and compared their performance. The experiments to determine the μ , η and γ in eq. 15 are conducted, and μ , η and γ denote the weights adjustment to coordinate the balance of three domains. The adjustment of these three parameters ranges from 0 to 1 with an interval of 0.1. We first adjusted parameters μ and η to the optimal weight ratio of spatial and temporal-spatial domains, and results are shown in Fig. 8 (a). When the assignment values of two parts are equal, the prediction accuracy reaches the best. Under the constraint of $\mu = \eta$ and $\mu + \eta + \gamma = 1$, different weight values are allocated to γ to achieve the best performance of the algorithm. The result is shown in Fig. 8 (b). Therefore, the proposed metric achieves the best performance with $\mu = 0.4$, $\eta = 0.4$ and $\gamma = 0.2$. Such weight scheme also matches contribution to the prediction performance of each domain provided in Table XI.

C. Prediction Performance Evaluation

The experimental results on TJU-SVQA database Phase I, Phase II and NAMA3DS1-COSPAD1 database are shown

TABLE XIV: Performance comparison of quality prediction models on TJU-SVQA database Phase I and Phase II.

Database	Method	H.264		JPEG2000		ALL	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
TJU-SVQA Database Phase I	PSNR (average)	0.5273	0.5004	0.5951	0.5740	0.4510	0.4416
	SSIM (average)	0.7460	0.7159	0.7001	0.6908	0.6882	0.6681
	Cyclopean (average)	0.6183	0.6071	0.6547	0.6770	0.6169	0.6052
	VQM	0.6796	0.6500	0.7886	0.7343	0.7563	0.7330
	DPQM	0.8185	0.7538	0.8625	0.8252	0.8525	0.8424
	Proposed	0.8803	0.8509	0.8081	0.7629	0.8724	0.8602
TJU-SVQA Database Phase II	PSNR (average)	0.5734	0.5794	0.5392	0.6172	0.4664	0.4667
	SSIM (average)	0.6611	0.6338	0.7117	0.6910	0.4501	0.4416
	Cyclopean (average)	0.5270	0.4700	0.6459	0.6216	0.5009	0.4648
	VQM	0.8496	0.8382	0.8947	0.8900	0.8020	0.7893
	DPQM	0.7803	0.7668	0.7466	0.7637	0.7613	0.7471
	Proposed	0.9190	0.9080	0.8728	0.8795	0.9055	0.8984

Fig. 8: PLCC and SROCC of the proposed method with different weight schemes on the TJU-SVQA database Phase I (a) PLCC and SROCC in regulating the relationship between μ and η ($\mu + \eta = 1$). (b) PLCC and SROCC in regulating the relationship between μ , η and γ ($\mu + \eta + \gamma = 1$).

in Table XIV and Table XV, respectively. The method with outstanding performance is highlighted in boldface. Overall, compared with other metrics, the proposed method has better performance in both evaluation criteria on individual databases, which provides strong support for its effectiveness. To more intuitively express the consistency between the objective results of each algorithms and MOS values, we displayed the scatter plots of the predicted results of the above algorithms on TJU-SVQA database Phase I in Fig. 9. To make the comparison more explicit, for the FR methods, we randomly selected 20% of the samples in the database for demonstration, which is consistent with the test set sample number of the NR methods. From Fig. 9, the fitting degree between the predicted score of the proposed method and MOS is impressive. Since the FR methods (PSNR, SSIM, Cyclopean) we compared are extended from 2D-IQA and 3D-IQA to 3D-VQA, binocular vision characteristics and temporal information are ignored. Therefore the performance is not satisfactory.

Since the prediction quality of symmetric distortion is relatively less difficult than that of asymmetric distortion, and video content in NAMA3DS1-COSPAD1 database is

TABLE XV: Overall performance comparison on the NAMA3DS1-COSPAD1 database.

Method	PLCC	SROCC	P_{GData}
PSNR (average)	0.6699	0.6470	-
SSIM (average)	0.7664	0.7492	-
VQM [9]	0.8949	0.8552	0.2863
DPQM [8]	0.9086	0.9239	0.3056
Proposed	0.9276	0.9028	0.3774

relatively less, the overall performance of the algorithms in NAMA3DS1-COSPAD1 database is better than performance in TJU-SVQA databases. Moreover, DPQM also has an impressive performance on the symmetric distortion databases, and particularly its SROCC indicator is superior to the others on the NAMA3DS1-COSPAD1 database. To further explore the performance of the proposed method and DPQM, we applied a new performance measure P_{GData} proposed in [56]. The higher the index is, the better the prediction performance of the algorithm is achieved. We trained the prediction model corresponding to each algorithm on the TJU-SVQA Database Phase I, and verified the performance of each model on the NAMA3DS1-COSPAD1 database. It can be observed that the proposed method performs relatively better. This indicator is required to use the training and validation set, so the FR methods are not implemented.

To evaluate the performance difference between any two algorithms, the significance experiment is conducted on the SROCC statistical sequence of any two methods. T-test [57] operates on 1000 SROCC of two random methods on TJU-SVQA databases, and results are displayed in Fig. 10. Thereinto, 1 manifests that the horizontal algorithm performance is better than the columnar method performance, while -1 manifests that the columnar method is superior. In addition, 0 manifests that the two algorithms are statistically correlated and their performance is similar. Since only the NR method

Fig. 9: A scatter diagram of MOS mapping with objective quality scores of six methods on TJU-SVQA database Phase I. (a) PSNR, (b) SSIM, (c) Cyclopean, (d) VQM, (e) DPQM and (f) the proposed metric.

	Proposed	DPQM	VQM
Proposed	0	1	1
DPQM	-1	0	1
VQM	-1	-1	0

(a)

	Proposed	DPQM	VQM
Proposed	0	1	1
DPQM	-1	0	-1
VQM	-1	1	0

(b)

Fig. 10: The result of the significance test for each NR-SVQA performance index SROCC. (a) T-test results on TJU-SVQA database Phase I. (b) T-test results on TJU-SVQA database Phase II

needs to execute training model and multiple iterations, the significance test is just carried out for the three NR methods. The experimental results are consistent with Table XIV. The performance of each algorithm is divergent, and the proposed algorithm performs better.

Compared with symmetrical distortion, asymmetrical distortion is more realistic and more difficult to predict. Since many methods can not accurately reflect the binocular perception of asymmetric distortion, it is difficult to match the predicted results with the subjective scores, and their performance would decline on the TJU-SVQA Database Phase II. However, as shown in Table XIV, the proposed algorithm and VQM perform better on the TJU-SVQA Database Phase II. This might be due to the fact that the number of stereoscopic video samples in the Phase II is significantly larger than that of the Phase I, which makes the training sufficient and finally achieves better results in the prediction stage. To verify the conjecture, stereoscopic video samples with an equal amount of Phase I are randomly selected on Phase II, during the training and testing phase. As shown in Table XVI, experimental results show that with the the same sample size, the proposed algorithm performs better on Phase I database, which

TABLE XVI: The performance of the same number of content on TJU-SVQA database Phase I and Phase II.

Database	PLCC	SROCC
TJU-SVQA Phase I	0.8724	0.8602
TJU-SVQA Phase II (part)	0.7080	0.6748

demonstrates that the sample size has significant influence on the experimental results.

Furthermore, to verify the predictive capacity of the proposed algorithm in individual distortion types, we divided Phase I and Phase II into two parts respectively according to the type of distortions (H.264 and JPEG2000). The training and testing process is the same as the experiment based on the whole database, and the results are shown in table XIV. Obviously, the proposed algorithm has a stable performance on H.264 subsets, while DPQM and VQM have remarkable performance on JPEG2000 subsets.

D. Cross-database Experiment

To illustrate the generalization ability of the presented method, we executed cross-database experiment on the above three databases. Since the FR method does not require training, cross-database validation are conducted on only three NR methods. In this paper, TJU-I/NAMTE means that the NR method is trained on the TJU-SVQA database Phase I and tested on the NAMA3DS1-COSPAD1 database. A total of six groups of experiments are presented: 1) TJU-I/TJU-II, 2) TJU-II/TJU-I, 3) TJU-I/NAMTE, 4) NAMTE/TJU-I, 5) TJU-II/NAMTE, 6) NAMTE/TJU-II. The results are displayed in Table XVII.

We observed that although the proposed approach performed relatively well, these algorithms show significant accuracy degradation in cross-database experiments compared to the individual database experiments, especially when setting TJU-SVQA database Phase II as test set. Since TJU-SVQA database Phase II has a larger sample size and a richer variety of distortion types than the other two databases, the prediction model cannot be adequately trained and the veracity decreases greatly. Meanwhile, those algorithms performance is relatively high when TJU-SVQA database Phase II is set as the training set, especially for VQM.

VI. CONCLUSION

In this paper, we built two stereo video quality assessment databases, namely TJU-SVQA Phase I and TJU-SVQA Phase II, which are all freely available, making it practical for cross-database experiments and the comparison between different studies. TJU-SVQA databases contain various kinds of content, symmetrically and asymmetrically distorted videos, and have a large number of video samples, which makes it more in line with actual demand and more versatile. In addition, we implemented subjective quality assessment experiments on the two databases. Results show that the presented evaluation is well balanced on the sequences variability. Then we proposed

TABLE XVII: Performance comparison of quality prediction models in cross database experiments on TJU-SVQA databases and NAMA3DS1-COSPAD1 database.

	TJU-I/TJU-II		TJU-II/TJU-I		TJU-I/NAMTE		NAMTE/TJU-I		TJU-II/NAMTE		NAMTE/TJU-II	
Method	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
VQM	0.4128	0.3840	0.8378	0.7962	0.6656	0.6472	0.6058	0.6100	0.8244	0.8140	0.4606	0.1757
DPQM	0.6139	0.6076	0.7889	0.7924	0.6214	0.5535	0.6833	0.7067	0.7570	0.6626	0.5474	0.5175
Proposed	0.5601	0.5498	0.8376	0.8381	0.7304	0.7330	0.7069	0.6866	0.7860	0.7874	0.5340	0.5245

a comprehensive NR-SVQA model to verify the TJU-SVQA databases' validity with inter-frame cross map, leading to an improvement of the quality prediction performance of stereoscopic videos. In the future, three aspects could be taken into consideration: first to explore the synthesis of the final quality score by combining the quality scores of each domain with the dynamic weight, second to explore the novel good property model with inter-frame in TJU-SVQA database, third to further study on the effect of video distortion on temporal information.

REFERENCES

- [1] S. Delis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Automatic detection of 3D quality defects in stereoscopic videos using binocular disparity," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 5, pp. 977-991, 2017.
- [2] T. Kim, J. Kang, S. Lee, and A. C. Bovik, "Multimodal interactive continuous scoring of subjective 3D video quality of experience," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 387-402, 2014.
- [3] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable Vision Assistance System Based on Binocular Sensors for Visually Impaired Users," *IEEE Internet of Things Journal*, to be published.
- [4] Z. Zhang, C. Zhou, R. Wang, Y. Wang, and W. Gao, "A compact representation for compressing converted stereo videos," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2343-2355, 2014.
- [5] Q. Huynh Thu, M. N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1-14, 2011.
- [6] K. Wang, M. Barkowsky, K. Brunnstrom, M. Sjostrom, R. Cousseau, and P. L. Callet, "Perceived 3D TV transmission quality assessment: multi-laboratory results using absolute category rating on quality of experience scale," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 544-557, 2012.
- [7] J. Wang, S. Wang, and Z. Wang, "Asymmetrically compressed stereoscopic 3D videos: quality assessment and rate-distortion performance evaluation," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1330-1343, 2017.
- [8] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: from depth perception to overall experience," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 721-734, 2018.
- [9] M. Urvoay et al., "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Proc. IEEE 4th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Yarra Valley, VIC, Australia, Jul. 2012, pp. 109-114.
- [10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.
- [11] C. T. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304-318, 2009.
- [12] J. Yang, H. Wang, W. Lu, B. Li, A. Badii, and Q. Meng, "A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain," *Information Sciences*, vol. 414, pp. 133-146, 2017.
- [13] B. Jiang, J. Yang, Q. Meng, B. Li and W. Lu, "A Deep Evaluator for Image Retargeting Quality by Geometrical and Contextual Interaction," *IEEE Transactions on Cybernetics*, to be published.
- [14] A. Aldahdooh, E. Masala, G. V. Wallendaal, and M. Barkowsky, "Comparing temporal behavior of fast objective video quality measures on a large-scale database," *Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016, pp. 1-5.
- [15] K. Seshadrinathan, and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335-350, 2010.
- [16] Z. Wang, L. Lu, Ligang and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing Image Communication*, vol. 19, no. 2, pp. 121-132, 2004.
- [17] S. Li, L. Ma, and K. N. Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1100-1112, 2012.
- [18] K. Manasa, and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480-2492, 2016.
- [19] L. He, W. Lu, C. Jia, and L. Hao, "Video quality assessment by compact representation of energy in 3D-DCT domain," *Neurocomputing*, vol. 269, pp. 108-116, 2017.
- [20] M. A. Saad, A.C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352-1365, 2014.
- [21] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *A Completely Blind Video Integrity Oracle*, *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289-300, 2015.
- [22] F. Zhang, and D. R. Bull, David, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017-1028, 2016.
- [23] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "Toward a blind quality metric for temporally distorted streaming video," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 367-378, 2018.
- [24] Y. Li, L. M. Po, C. H. Cheung, X. Xu, L. Feng, F. Yuan, and K. W. Cheung, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1044-1057, 2016.
- [25] M. J. Chen, D. K. Kwon, and A. C. Bovik, "Study of Subject Agreement on Stereoscopic Video Quality," in *Proc. IEEE Southwest Symposium on Image Analysis and Interpretation. (SSIAI)*, Santa Fe, NM, USA, April. 2012, pp. 173-176.
- [26] C. T. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *Electronics letters*, vol. 44, no. 16, pp. 963-965, 2008.
- [27] S. L. P. Yasakethu, C. T. Hewage, W. A. C. Fernando, and A. M. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1969-1976, 2008.
- [28] M. H. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312-322, 2004.
- [29] C. Galkandage, J. Calic, S. Dogan, and J. Y. Guillemaut, "Stereoscopic video quality assessment using binocular energy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 102-112, 2016.
- [30] M. Yu, K. Zheng, G. Jiang, F. Shao, Z. Peng, "Binocular perception based reduced-reference stereo video quality assessment method," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 246-255, 2016.
- [31] J. Han, T. Jiang, and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," in *Proc. IEEE Visual Communications and Image Processing*, San Diego, CA, USA, Nov. 2013, pp. 1-6.

- [32] H. Zhu, M. Yu, Y. Song, and G. Jiang, "A stereo video quality assessment method for compression distortion," in *Proc. IEEE Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, Dec. 2015, pp. 481-485.
- [33] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," in *Proc 18th IEEE International Conference on Image Processing*, Brussels, Belgium, Dec. 2011, pp. 2521-2524.
- [34] F. Shao, Q. Yuan, W. Lin, and G. Jiang, "No-reference view synthesis quality prediction for 3D videos based on color-depth interactions," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 659-674, 2017.
- [35] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal, Image and Video Processing*, vol. 10, no. 4, pp. 737-744, 2016.
- [36] G. A. Lutz D. S. A. Francesca, and E. A. Touradj, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *The International Society for Optical Engineering*, vol. 7526, no. 2, pp. 158-160, 2010.
- [37] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427-1441, 2010.
- [38] <http://iphome.hhi.de/suehring/tml/download/>. Accessed 21 July 2015.
- [39] http://www.kakadusoftware.com/index.php?option=com_content&task=view&id=26&Itemid=22. Accessed 21 July 2015.
- [40] L. Jin, A. Boev, K. Egiazarian, and A. Gotchev, "Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3D video," *EURASIP Journal on Image and Video Processing*, vol. 26, no. 6, pp. 1044-1057, 2016.
- [41] Z. Lit, and J. J. Atick, "Efficient stereo coding in the multiscale representation," *Network: Computation in Neural Systems*, vol. 5, no. 2, pp. 157-174, 1994.
- [42] J. Yang, K. Sim, X. Gao, W. Lu, Q. Meng and B. Li, "A Blind Stereoscopic Image Quality Evaluator With Segmented Stacked Autoencoders Considering the Whole Visual Perception Route," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1314-1328, 2019.
- [43] J. Yang, C. Ji, B. Jiang, W. Lu and Q. Meng, "No Reference Quality Assessment of Stereo Video Based on Saliency and Sparsity," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 341-353, 2018.
- [44] W. Zhou, W. Qiu, and M. W. Wu, "Utilizing Dictionary Learning and Machine Learning for Blind Quality Assessment of 3-D Images," *IEEE Transactions on Broadcasting*, vol. 63, no. 2, pp. 404-415, 2017.
- [45] A. V. Oppenheim, and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529-541, 1981.
- [46] J. Bai, X. Dong, S. He, and M. Bao, "Monocular deprivation of Fourier phase information boosts the deprived eye's dominance during interocular competition but not interocular phase combination," *Neuroscience*, vol. 352, pp. 122-130, 2017.
- [47] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.
- [48] F. Tong, K. Nakayama, J. T. Vaughan, and N. Kanwisher, "Binocular rivalry and visual awareness in human extrastriate cortex," *Neuron*, vol. 21, no. 4, pp. 753-759, 1998.
- [49] K. Lee, and S. Lee, "3D perception based quality pooling: stereopsis, binocular rivalry, and binocular suppression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 533-545, 2015.
- [50] R. Patterson, "Review Paper: Human factors of stereo displays: An update," *Journal of the Society for Information Display*, vol. 17, no. 12, pp. 987-996, 2012.
- [51] J. M. Wolfe, "Stereopsis and binocular rivalry," *Psychological Review*, vol. 93, no. 3, pp. 269-282, 1986.
- [52] L. Wang, C. Teunissen, Y. Tu, L. Chen, P. Zhang, and T. Zhang, "Crosstalk evaluation in stereoscopic displays," *Journal of Display Technology*, vol. 7, no. 4, pp. 208-214, 2011.
- [53] J. Kim, and T. Kim, S. Lee, and A. C. Bovik, "Quality assessment of perceptual crosstalk on two-view auto-stereoscopic displays," *IEEE Transaction Image Processing*, vol. 26, no. 10, pp. 4885-4899, 2017.
- [54] B. K. P. Horn, and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185-203, 1981.
- [55] S. Fazekas, and D. Chetverikov, "Normal versus complete flow in dynamic texture recognition: a comparative study," *Texture International Workshop on Texture Analysis and Synthesis*, pp. 37-42, 2005.
- [56] A. Aldahdooh, E. Masala, O. Janssens, G. V. Wallendael, M. Barkowsky, and P. L. Callet, "Improved Performance Measures for Video Quality Assessment Algorithms Using Training and Validation Sets" *IEEE Transactions on Multimedia*, to be published.
- [57] G. D. Ruxton, "The Unequal Variance T-Test is an Underused Alternative to Student's T-Test and the Mann-Whitney U Test," *Behavioral Ecology*, vol. 17, no. 4, pp. 688-690, 2006.

Jiachen Yang received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, China, in 2005 and 2009, respectively. He is currently a professor at Tianjin University. His research interests include stereo vision research, pattern recognition and image quality evaluation.

Yang Zhao received a B.S. degree in Information Engineering from Zhengzhou University in 2016 and received the M.S. degree in communication and information engineering from Tianjin University, China, in 2019. Currently, she is pursuing the Ph.D. degree with School of electrical and information engineering, Tianjin University, Tianjin, China. Her research interests include stereoscopic video and multimedia quality evaluation.

Bin Jiang received the B.S. and M.S. degree in communication and information engineering from Tianjin University, China, in 2013 and 2016. He is currently pursuing the Ph.D. degree at the school of electrical and information engineering, Tianjin University, China. His research interests include multimedia quality assessment, virtual reality, pattern recognition and stereo vision research.

Qinggang Meng (M'06-SM'18) received the BSc and MSc degrees from the School of Electronic Information Engineering, Tianjin University, China, and the Ph.D. degree in computer science from Aberystwyth University, U.K. He is currently a Professor in Robotics and AI with the Department of Computer Science, Loughborough University, UK. He is a fellow of the Higher Education Academy, UK. His research interests include biologically inspired learning algorithms and developmental robotics, service robotics, agricultural robotics, robot learning and adaptation, multi-UAV cooperation, human motion analysis and activity recognition, pattern recognition, artificial intelligence, computer vision, and embedded intelligence.

Wen Lu received the B.Sc., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xian, China, in 2002, 2006, and 2009, respectively. Since 2009, he joined the School of Electronic Engineering, Xidian University. From 2010 to 2012, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Stanford University, Stanford, CA, USA. He is currently a Professor with the School of Electronic Engineering, Xidian University. He has authored or co-authored more than 2 books and around 50

technical articles in refereed journals and proceedings including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, Information Science, and Neurocomputing. His current research interests include multimedia analysis, computer vision, pattern recognition, and deep learning. Dr. Lu is on the Editorial Boards and serves as a reviewer for many journals, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON MULTIMEDIA.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xian, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian

University. He is currently a Cheung Kong Professor of the Ministry of the Peoples Republic of China, a Professor of pattern recognition and intelligent system with Xidian University, and the Dean of the Graduate School of Xidian University. He has authored or co-authored more than six books and around 200 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. Dr. Gao is a Fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics. He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).