Stereoscopic video quality assessment based on 3D convolutional neural networks

Jiachen Yang [a], Yinghao Zhu [a], Chaofan Ma [a], Wen Lu [b,*], Qinggang Meng [c]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[b] School of Electronic Engineering, Xidian University, Xian, China
[c] Department of Computer Science, Loughborough University, Loughborough, UK

A B S T R A C T

The research of stereoscopic video quality assessment (SVQA) plays an important role for promoting the development of stereoscopic video system. Existing SVQA metrics rely on hand-crafted features, which is inaccurate and time-consuming because of the diversity and complexity of stereoscopic video distortion. This paper introduces a 3D convolutional neural networks (CNN) based SVQA framework that can model not only local spatio-temporal information but also global temporal information with cubic difference video patches as input. First, instead of using hand-crafted features, we design a 3D CNN architecture to automatically and effectively capture local spatio-temporal features. Then we employ a quality score fusion strategy considering global temporal clues to obtain final video-level predicted score. Extensive experiments conducted on two public stereoscopic video quality datasets show that the proposed method correlates highly with human perception and outperforms state-of-the-art methods by a large margin. We also show that our 3D CNN features have more desirable property for SVQA than hand-crafted features in previous methods, and our 3D CNN features together with support vector regression (SVR) can further boost the performance. In addition, with no complex preprocessing and GPU acceleration, our proposed method is demonstrated computationally efficient and easy to use.

Keywords:
3D convolutional neural networks
Stereoscopic video quality assessment
Quality score fusion

1. Introduction

Nowadays, a large number of stereoscopic videos are created for various fields such as entertainment and education. Highly associated with the users' Quality of Experience (QoE), visual quality is a fundamental yet complex characteristic of a stereoscopic video that may suffer varying degrees of damage in the successive stages of stereoscopic video production, including processing, compression, transmission and display. Hence, the research of stereoscopic video quality assessment (SVQA) plays an important role in the development of stereoscopic video systems. In order to reach higher efficiency and feasibility, unattended and automatic objective SVQA methods instead of subjective methods are in great demand.

Depending on the amount of pristine video information available, the objective SVQA methods can fall into three types: full-reference (FR), reduced-reference (RR) and no-reference (NR). NR methods can assess the quality of tested stereoscopic videos without any information from reference content, while FR methods and NR methods require pristine video or its partial information. Unfortunately, considering that reference video is unavailable in most practical applications, only the NR methods have potential to satisfy the actual requirement. As a result, our work focuses on more appealing and challenging NR methods, and tries to propose a new general-purpose NR framework for SVQA.

Considering the development of NR metrics [1,2], we can find that most of them have similar frameworks, which can be generally divided into two steps: (1) extracts features that can reflect visual quality based on relatively perceptual models; (2) maps the obtained feature vectors to subjective quality scores by learning regression models. But in real-world scenarios, because of the diversity and complexity of video distortions, it is very difficult to identify what features are sensitive and robust to all sorts of distortions. Therefore, simply utilizing a group of artificially designed features to represent video quality results in inaccurate assessments and high computational costs.

Recently, it is a well-known fact that deep learning models, especially convolutional neural networks (CNN), have achieved great success in many challenge computer vision tasks, such as image classification [3,4], object detection [5,6], video classification [7,8] and video action recognition [9,10]. CNN is a biologically inspired architecture consisting of a stack of convolutional layers

* Corresponding author.
E-mail address: luwen@xidian.edu.cn (W. Lu).

and pooling layers, automatically extracting a hierarchy of powerful features from row data. Lately, CNN also demonstrated its superiority for image quality assessment (IQA) and stereoscopic image quality assessment (SIQA). Kang et al. [11] and Zhang et al. [12] proposed two CNN based NR metrics to realize effective quality assessment of 2D images and stereoscopic images, respectively.

In this work, we attempt to explore a CNN based stereoscopic video quality assessment metric. Apparently, a straightforward manner inspired by previous CNN based IQA/SIQA models is to regard stereoscopic video frames as stereoscopic pairs and employ CNN at the frame level sequentially. Nevertheless, such an image based method cannot yield superior performance due to neglect of the motion information contained in the adjacent video frames. To address this problem, we construct a 3D CNN architecture to learn spatio-temporal features for NR SVQA task, which is able to adequately encapsulate information related to stereoscopic video quality. The experimental results demonstrate that our method significantly outperforms the cutting-edge methods. In summary, our key contributions are as follows:

- We present a 3D CNN based framework for SVQA, which is able to model local spatio-temporal information but also global temporal information with cubic difference video patches as input. To the best of our knowledge, we are the pioneers to exploit the 3D CNN to evaluate the quality of stereoscopic video.
- After our 3D CNN architecture effectively capturing local spatio-temporal features, we design a quality score fusion strategy considering global temporal clues to pool patch-level quality scores into video-level quality score. Through a large number of extensive experiments, our proposed method achieves the best performance to date on both two challenging stereoscopic video quality databases and outperforms current best performing methods by a large margin.
- Our proposed framework takes cubic difference video patches as input and does not rely on any complex preprocessing such as optical flow and gradients, so it is computationally efficient and applicable in real applications compared with previous methods.

The rest of this paper is structured as follows: we first review the works related to our method in Section 2. Section 3 details proposed method. The experimental results and some analysis of key issues of the proposed method are shown in Section 4. Finally we conclude our work in Section 5.

2. Related work

2.1. Conventional Stereoscopic video quality assessment

SVQA is a significant but intractable task in computer vision, attracting more and more attention in recent years. As a result, various methods were proposed for this subject. Initially, researchers expected to accomplish the evaluation of stereoscopic video quality using 2D IQA metrics or 2D video quality assessment (VQA) metrics, proposed Peak Signal to Noise Ratio (PSNR) based method [13], Structural Similarity Index Metric (SSIM) based method [14] and VQM based method [15]. In this case, 2D metrics were performed on two views of stereoscopic video separately, and then averaged to integrate the final quality score. However, since the depth information and temporal information of stereoscopic video were not considered sufficiently, the aforementioned methods failed to obtain convincing results. As a consequence, some methods were proposed to measure the 3D video perceptual quality by applying both 2D and 3D information extracted from stereoscopic video. For example, Malekmohamadi et al. [16] proposed a RR method that encoded spatial neighboring information from gray level co-occurrence matrices for both color and depth sections. In [17], left-right views quality metric and depth perception metric were designed and pooled into SVQA score. More concretely, left-right views quality was assessed based on significant pixels and just noticeable distortion model, while depth perception quality evaluated by deploying three-dimensional wavelet transform. Recently, some approaches exploiting the human visual system (HVS) model began to emerge, showing more reliable performance. Taking the temporal characteristics of video and binocular perception in HVS into account, a RR SVQA method was proposed by Yu et al. [18]. In [19], Galkandage et al. presented a FR SVQA method built on a HVS model incorporating the phenomena of binocular suppression and recurrent excitation. Despite of these achievements, all of the above methods are FR or NR methods without practical value and few NR methods were presented. In our previous work [20], a NR SVQA metric was proposed for the first time, which jointly focused on the spatial information, the temporal information and the inter-frame spatiotemporal information employing local binary patterns statistical features and local flow statistical features. Overall, whether based on 2D metrics or HVS model, most of the existing SVQA methods rely on artificially designed features that can represent stereoscopic video quality, which is inflexible and time-consuming.

2.2. Neural network based visual content quality assessment

There were many early works applying neural networks to visual content quality assessment. In [21], Li et al. developed a NR IQA algorithm that deployed a general regression neural network (GRNN) with perceptual features including phase congruency, entropy and the image gradients as input. Chetouani et al. [22] used a neural network to combine multiple distortion-specific NR IQA measures. In [23], a machine learning method was presented for evaluating blocking artifacts in JPEG images and a SVR model is adopted to learn the underlying relations between features and perceived blocking artifacts. However, these methods require artificially designed features and only adopted shallow neural networks with only one or two hidden layers to learn the regression function.

2.3. Deep learning based visual content quality assessment

With remarkable success deep learning models have achieved in various computer vision tasks, a few deep learning based visual content quality assessment methods have shown remarkable performance. On the one hand, some researchers explored deep learning models to transform lower-level features into more abstract and higher-level representations for visual quality. For example, Tang et al. [24] constructed a semi-supervised rectifier neural network to blindly measure 2D image quality. A deep belief network (DBN) [25] of three layers was adopted to provide a high-level feature representations of LBIQ features [26], and the final quality was predicted with Gaussian Process regression. Ghadiyaram and Bovik [27] proposed natural-scene-statistics-based perceptual image features with a DBN to tackle the difficult problem of blindly image quality assessment on authentically distorted images. Hou et al. [28] proposed a NR IQA model to learn qualitative evaluations by using a four-layer discriminative deep model, which is pre-trained with DBN and discriminatively fine-tuned by back-propagation. Shao et al. [29] trained two separate 2D deep neural networks (DNN) from 2D monocular images and cyclopean images to evaluate the quality of stereoscopic image. In [30], a NR VQA approach based on 3D shearlet transform and 1D CNN was constructed, and high-level spatiotemporal features were produced by performing 1D CNN on the simple features directly extracted from videos.

On the other hand, only a few researchers tried to apply 2D CNN on IQA/SIQA task without using hand-crafted features, which took raw visual content as input and incorporated feature learning into the training process. In [11], Kang et al. pioneered a CNN based 2D IQA method to integrate feature extraction and regression into one optimization process. The experiment demonstrated that the proposed CNN can learn the local structures which are sensitive to human perception and representative for perceptual quality evaluation. Inspired by this work, Zhang et al. [12] designed two different CNNs with different inputs, namely one-column CNN with only the image patch from the difference image as input, and three-column CNN with the image patches from left-view image, right-view image, and difference image as the input.

2.4. Video analysis using convolutional neural networks

In last years, CNNs have made a series of significant breakthroughs on image recognition, a lot of powerful CNN architectures [3,31,32] were created for image feature learning. Driven by the success in 2D image processing tasks, CNN has also been utilized for video analysis. Karpathy et al. [33] presented several methods for extending the connectivity of CNN to capture the spatiotemporal information, which showed that CNN can generate strong improvement over hand-crafted features. Simonyan and Zisserman [34] proposed a two stream CNN network for video classification. One network analyzed the spatial information while the second analyzed multi-frame dense optical flow. Jain et al. [35] articulated human pose estimation in videos using a CNN architecture, which incorporated both color and motion features. However, only 2D convolution and 2D pooling operations were adopted in these approaches, which cannot naturally make use of the motion information in the network.

In this work, we extend 2D CNN to 3D CNN to remain and propagate temporal information across the network, which is well-suited for SVQA task. There have been some exploration of 3D CNN in numerous research topics. For instance, in [36], a novel 3D CNN architecture was proposed for human action recognition, implementing data representations from both spatial and temporal dimensions. Besides, Tran et al. [37] designed and trained 3D CNN models on large video datasets, which were demonstrated effective for several video analysis tasks including action recognition and scene recognition. In [38], a 3D CNN based discrimination model was developed to participate in the detection of cerebral microbleeds, sufficiently representing the spatial contextual information and hierarchically extracting high-level features. As evidenced by these previous works, CNN in 3D fashion is a promising solution to tackle video analysis problems.

3. Proposed method

As illustrated in Fig. 1, we construct a no-reference stereoscopic video quality assessment framework built on 3D CNN, which is able to adequately encapsulate local spatiotemporal information and global temporal information. Specifically, a 3D CNN architecture is devised to extract local spatiotemporal information while a quality score fusion strategy considering global temporal clues is adopted to obtain final video-level predicted scores. In this section, we describe the key components of the proposed framework, including data preprocessing, 3D CNN architecture and quality score fusion.

3.1. Data preprocessing

3.1.1. Difference video

Stereoscopic video consists of two 2D videos with disparity, which incorporates more visual information than ordinary 2D image and video. Therefore, to measure the perceived quality of 3D video, we should consider the non-intuitive interaction of several complicated visual factors, including video content quality and depth perception. In our previous work [39–41], the difference image calculated from left and right views has been demonstrated to retain stereoscopic perception information, which can be used to represent the quality of stereoscopic image. Additionally, Ma et al. [42] concluded that the difference image is more valuable than the left and right views in SIQA task. Similarly, we evaluate stereoscopic video quality by conducting our 3D CNN on the difference video rather than directly on the left and right views in this work. One reason is that the difference video incorporates video content together with depth perception information, which is suitable as raw data for further analysis of stereoscopic video quality. Another reason is that applying the difference video is applicable for stereo video analysis with massive data due to its low computational complexity. Suppose V_L and V_R denote left and right views of stereoscopic video and the value of the difference video D_L at position (x, y, z) is computed as:

$$D_L(x, y, z) = |V_L(x, y, z) - V_R(x, y, z)| \tag{1}$$

Fig. 2 shows individual frames sampled form two difference videos with different quality. Comparing Fig. 2 (a) and (b), it can be noticed that the difference video captures contour information coupled with depth information, and can represent the stereoscopic video quality.

3.1.2. Dataset augmentation

Note that training effective deep learning models requires a large number of labeled data. However, unlike visual recognition subjects, the amount of data in existing stereoscopic video quality datasets is limited. As a result, in order to make our deep learning model generalize better, we first need to tackle the basic problem of efficient data scarcity before applying CNN to our task.

For many deep learning tasks, dataset augmentation is completed through creating new fake data with transformations like translating, rotating or scaling. Unfortunately, we cannot adopt these transformations for SVQA dataset augmentation because they may influence the quality of stereoscopic video. In this work, we propose an applicable dataset augmentation scheme for SVQA task. The raw video is split in both spatial and temporal dimensions, resulting in numerous low-resolution short video cubes. Specifically, the size of each cubic video patch is set as $10 \times 32 \times 32$, namely 10 frames with a resolution of 32×32. In this design, 32×32 rectangle boxes are cropped at the same position of 10 sequential frames, generating cubes with visual perception information. Then we are required to label these resulting cubes in order to obtain valid data for training. Based on the assumption that the quality degradation is homogeneous throughout the whole stereoscopic video, each video cube is annotated with a quality score consistent with the subjective score of the entire video, thus expanding the amount of efficient data, which satisfies the demand of CNN for data. The experiment in Section 4 verifies the correctness of our assumption.

Video analysis and processing are always time-consuming and costly. However, there is high redundancy in video data. Thereby, to make our obtained data more efficient, we need to reduce redundancy while increasing the amount of data. In this work, a subsampling strategy in all dimensions of video is adopted to reduce redundancy. Specifically, we slide a 32×32 box with a stride of 32 to crop the whole video in spacial dimension and select frames with a stride of 8 in temporal dimension. As a result, we obtain a cubic video patch set for each video as follows:

$$P_{cubic} = \left[SP^{(1)}, SP^{(2)}, \ldots, SP^{(i)}, \ldots, SP^{(I)}\right] \tag{2}$$

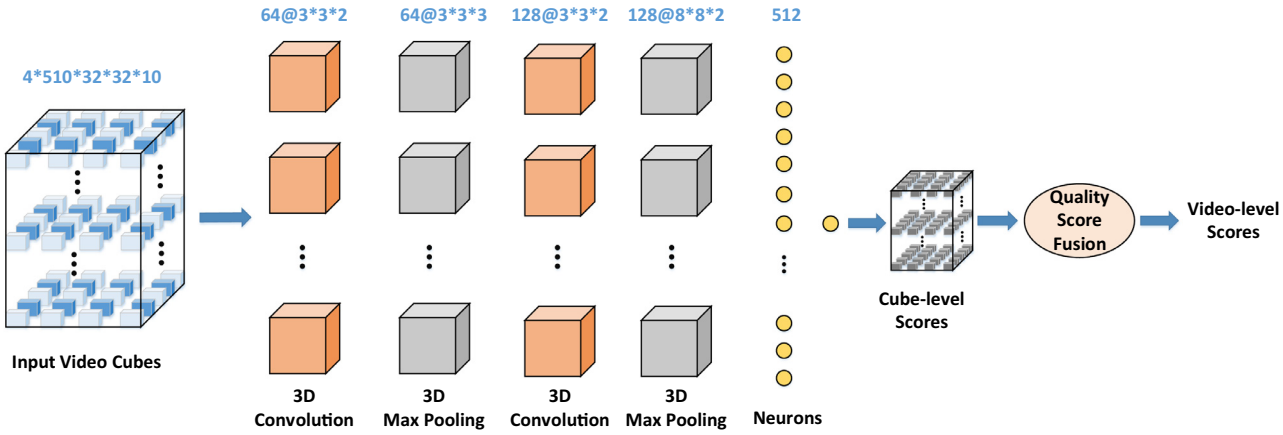


Fig. 1. The architecture of the 3D CNN based SVQA framework. A video is divided into numerous cubic patches and each cubic patch is fed to our 3D CNN to complete cube-level prediction. Finally, we employed a quality score fusion strategy to obtain final video-level perceptual quality score. Our 3D CNN architecture consists of two 3D convolution layers, two 3D pooling layers and two fully-connected layers. Detailed descriptions are given in the text.

(a)

(b)

Fig. 2. (a) 150*th* frame of a reference stereoscopic video (MOS = 4.75) and the corresponding difference frame. (b) 150*th* frame of a distorted stereoscopic video (MOS = 1.10) and the corresponding difference frame.

$$SP^{(i)} = \begin{pmatrix} C_{11}^{(i)} & C_{12}^{(i)} & \cdots & C_{1N}^{(i)} \\ C_{21}^{(i)} & C_{22}^{(i)} & \cdots & C_{2N}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ C_{M1}^{(i)} & C_{M2}^{(i)} & \cdots & C_{MN}^{(i)} \end{pmatrix} \tag{3}$$

where C is a $10 \times 32 \times 32$ cubic patch and $SP^{(i)}$ denotes the ith segment in temporal dimension. Ultimately, we build a training set consisting of 204,000 video cubes base on NAMA3DS1-COSPAD1 database [43], which is 2040 times larger than the original database. Our data preprocessing successfully lay the foundation for the training of our deep learning models.

3.2. 3D convolutional neural networks

Generally, typical CNN architecture stacks multiple convolution layers and pooling layers alternatively to process the input signal, and then implements the mapping between features and objective in the fully-connected layer. In 2D CNN, convolution operation and pooling operation are employed in spatial dimension merely, which is not suitable for the process of video streams with both spatial and temporal information. To this end, we extend 2D CNN to 3D CNN for SVQA task by virtue of carrying out 3D convolution and 3D pooling on the cubic video patches. Next we will describe 3D convolution operation, 3D pooling operation and our 3D CNN architecture.

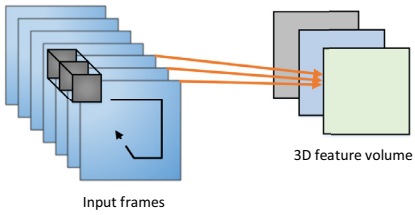
3.2.1. 3D convolution

Convolution employed at the convolutional layers in CNN is a special linear operation between input data and several kernel functions to produce feature maps. On this basis, 3D convolution is completed by convolving 3D convolution kernel with the cubic patches composed of multiple adjacent frames to ensure that the temporal information is preserved and abstracted across the network, which is formally expressed in an element-wise form:

$$v_{ki}^{l}(x,y,z) = \sum_{r}\sum_{p}\sum_{q} h_{k}^{l-1}(x-p, y-q, z-r) W_{i}^{l}(p,q,r) \tag{4}$$

where k denotes the index of the feature map in the $(l-1)$ layer connected to the current convolution kernel, h_k^{l-1} represents kth 3D feature map in $(l-1)$th layer, W_i^l is ith 3D convolution kernel in lth layer that convolves over the h_k^{l-1}. After convolution completed, an additive bias term and an nonlinear activation function are performed to get the final feature map. Formally, the ith

(a) The workflow of 2D convolution



(b) The workflow of 3D convolution

Fig. 3. Comparison of (a) 2D and (b) 3D convolutions. 2D convolution slides the 2D convolution kernels on spatial dimension whereas 3D convolution slides the 3D convolution kernels on both spatial dimension and temporal dimension. In 3D convolution, each location of the 3D feature maps are connected with several adjacent input frames, which preserves temporal information of input frames.

feature map in the lth layer is given as

$$h_i^l = f\left(\sum_k v_{ki}^l + b_i^l\right) \tag{5}$$

where b_i^l is the additive bias term, $f(\cdot)$ is the nonlinear activation function, such as sigmoid function, hyperbolic tangent function and rectified linear function. The comparison of 2D convolution and 3D convolution is shown in Fig. 3.

3.2.2. 3D pooling

In addition to the convolution layer, the pooling layer is also a main component of typical CNN, which sub-sampling the feature map transmitted from the convolution layer based on the local correlation principle. The pooling operation outputs the summary statistic of adjacent units at a certain location of feature map, thereby reduce the amount of data while retaining valuable information. For example, the max pooling operation [44] reports the maximum value within a rectangular neighborhood of feature map. Similarly, we apply 3D pooling to produce invariance to translation over both spatial and temporal dimensions of the cubic video patches. Taking max pooling as an example, 3D pooling operation is formulated as

$$u_i^l(x,y,z) = \max_{m,n,j} h_k^l(x+m, y+n, z+j) \tag{6}$$

3.2.3. 3D CNN architecture

Based on the 3D convolution and 3D pooling elaborated above, we construct a 3D CNN architecture to automatically and effectively capture local spatiotemporal features for SVQA task. In theory, the deeper model has greater capacity, which means it can accomplish more complex tasks but requires more data at the same time. Although dataset augmentation scheme is applied, available valid data for SVQA task is so scarce that the complex model is easy to fall into over-fitting. Hence, the proposed network has a simple yet effective architecture with a total of six layers, including two 3D convolution layers C1, C2, two 3D pooling layers S1, S2 and two fully-connected layers FC1, FC2. After data preprocessing, we consider numerous $10 \times 32 \times 32$ (10 in the temporal dimension and 32×32 in the spatial dimension) cubic video patches as inputs to the 3D CNN model. For each convolution kernel, we fix the spatial receptive field to 3×3 according to the findings in 2D CNN [31] that small receptive fields of 3×3 convolution kernels yield best results. Then we vary and search the temporal depth of the 3D convolution kernels according to our experiments. As a result, the two 3D convolutional layers have filters with a kernel size of $2 \times 3 \times 3$. With C1 and C2 layers, multiple 3D feature maps are hierarchically generated to represent the stereoscopic video. After each convolution layer, a 3D max-pooling layer performs sub-sampling on 3D feature maps, which reduces the resolution of feature maps in spatial and temporal dimensions simultaneously. Specifically, the kernel sizes of two 3D max-pooling layers are $3 \times 3 \times 3$ and $2 \times 8 \times 8$, respectively. Finally, the proposed network ends with two fully-connected layers: FC1 contains 512 neurons to flatten 3D feature maps into a 512-D feature vector and FC2 only contains 1 neuron to predict a cube-level score corresponding to the quality of input cubic video patches. In summary, our network settings are shown in Table 1.

Table 1
Configurations of the proposed 3D CNN architecture.

Layer	Kernel sze	Stride	Output size	Feature maps
Input	–	1	$10 \times 32 \times 32$	1
C1	$2 \times 3 \times 3$	1	$9 \times 30 \times 30$	64
S1	$3 \times 3 \times 3$	1	$3 \times 10 \times 10$	64
C2	$2 \times 3 \times 3$	1	$2 \times 8 \times 8$	128
S2	$2 \times 8 \times 8$	1	$1 \times 1 \times 1$	128

We train the model using SGD optimizer with a minibatch size of 128 and apply a Nesterov momentum of 0.9. The learning rate is initialized to 0.001. The final network has 215361 parameters totally and all the trainable parameters in this model are initialized randomly and trained by the online error back-propagation algorithm as described in [45]. The rectifier linear unit(ReLU) [46] is utilized for the non-linear activation function in the C and FC layers.

In order to avoid over-fitting, we use dropout strategy [47] in the fully-connected layers to drop the input units with a fraction of 0.5, and adopt a objective function consisting of the original cost function and a regularization term as follows:

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 + \lambda||\theta||_F^2 \tag{7}$$

where y_i and $f(x_i)$ denote ground-truth quality score and predicted score, respectively. λ is the regularization parameter. Furthermore, batch normalization is used between each convolution and following activation to accelerate network training.

3.3. Quality score fusion

After the train-test process of our 3D CNN model, we can acquire the predicted score of each input cubic patch split from testing stereoscopic video. In order to obtain video-level quality score effectively, we employ a quality score fusion strategy considering global temporal information. First, average pooling is utilized to integrate the cube-level scores in the spatial dimension. Thereby, each video gets a score set $\{S_1, S_2, \ldots, S_i, S_I\}$ and S_i represents the quality of the ith segment in each stereoscopic video. To model global temporal information, we compute the weight of each segment based on the motion intensity. For efficiency, a simple way to acquire motion intensity is defined as:

$$I = \sum_{x,y,t}[V(x,y,t) - V(x,y,t-1)]^2 \tag{8}$$

Supposed I_i denotes the motion intensity of the ith segment of stereoscopic video in temporal dimension, the corresponding

(a) Barrier gate (b) Basket (c) Boxers (d) Hall (e) Lab

(f) News report (g) Phone call (h) Soccer (i) Tree branches (j) Umbrella

Fig. 4. The 100th frames of ten reference stereo videos in the NAMA3DS1-COSPAD1 stereo video database (only right views are shown).

weight of the ith segment is formulated as:

$$w_i = \frac{I_i}{\sum_i^N I_i} \tag{9}$$

Finally, we aggregate the video-level predicted score as follows:

$$F = \sum_i w_i S_i \tag{10}$$

where S_i is the quality score of the ith segment averaged from cubic patches in spatial dimension. Note that we adopt a score fusion method in temporal dimension based on motion intensity rather than using the simple average fusion, which incorporates global temporal information and models the affect of motion intensity on stereoscopic video quality.

4. Result and discussion

In this section, we first introduce the stereoscopic video databases and evaluation metrics used in the experiment. Then the effectiveness of the proposed method is validated on these databases. Furthermore, we treat our model as a feature extractor and reveal that our 3D CNN features have more desirable property for SVQA than hand-crafted features. Next we verify the correctness of the hypothesis that the quality degradation is homogeneous throughout the stereoscopic video. Finally, our proposed method is demonstrated computationally efficient compared with previous methods.

4.1. Stereoscopic video database

In our work, we utilize the NAMA3DS1-COSPAD1 stereoscopic video quality database [43] and the stereoscopic video quality database in [48] (we name it QI-SVQA for simplicity) to evaluate the performance of our proposed method.

4.1.1. NAMA3DS1-COSPAD1 stereoscopic video quality database

The NAMA3DS1-COSPAD1 database consists of 10 original stereoscopic videos with a resolution of 1080×1920 at 25 fps and 100 symmetrically distorted stereoscopic videos derived from the original videos. There are five types of distortion considered in this database, including H.264/AVC, JPEG 2000, reduction of resolution, image sharpening and downsampling&sharpening. The mean opinion scores (MOS) ranging from 1 to 5 is adopted to indicate the subjective quality of stereoscopic video, and higher MOS means better subjective quality. The first frames of ten original videos are summarized in Fig. 4.

4.1.2. QI-SVQA database

The QI-SVQA database is in the format of uncompressed YUV 4:2:0, which has 9 original videos and 450 symmetric or asymmetric distorted videos that are divided into two distortion type: Gaussblur and H.264. The frame rate of all videos in this database is 25 fps while the resolution and the number of frames are diverse. Similarly, the MOS accessible in the database also varies from 1 (bad) to 5 (excellent).

Table 2
Overall performance comparison on NAMA3DS1-COSPAD1.

NAMA3DS1-COSPAD1 database				
Algorithm	PLCC	SROCC	KROCC	RMSE
PSNR	0.6699	0.6470	0.4800	0.8433
SSIM	0.7664	0.7492	0.5444	0.7296
PQM [49]	0.6340	0.6006	0.4391	0.8784
PHVS-3D [50]	0.5480	0.5146	0.3572	0.9501
SFD [51]	0.5965	0.5896	0.4025	0.9117
3D-STS [52]	0.6417	0.6214	0.4544	0.9067
Feng [48]	0.6503	0.6229	0.4575	0.8629
Yang et al [20]	0.8949	0.8552	0.6913	0.4929
MNSVQM [53]	0.8545	0.8394	0.6439	0.4538
BSVQE [54]	0.9239	0.9086	0.7622	0.3754
Image-based method 1	0.9145	0.9082	0.7107	0.3908
Image-based method 2	0.9012	0.8985	0.7225	0.4056
3D CNN	0.9316	0.9046	0.7533	0.4161
3D CNN+SVR	**0.9478**	**0.9231**	**0.7883**	**0.3514**

4.2. Overall performance evaluation

Four commonly used measures are applied to quantitatively evaluate the effectiveness of the proposed method: Pearson linear correlation coefficient (PLCC), Spearman rank correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC) and Root mean squared error (RMSE). The PLCC assesses the linearity of an IQA index, while the SROCC measures its monotonicity. The objective scores are passed through the five-parametric nonlinear regression before computing PLCC and SROCC for mapping to DMOS or MOS space. When the objective score and the subjective score are exactly matched, PLCC = SROCC = KROCC= 1, RMSE = 0.

In this subsection, we evaluate our 3D CNN based SVQA method on abovementioned NAMA3DS1-COSPAD1 and QI-SVQA database, and compare its effectiveness with the best performing methods. In order to guarantee the reliability of the test results, we conduct the train-test process 100 times and adopt median value as the final performance evaluation results. Each time we picked 60% of the dataset as training set, 20% as the validation set and the remaining 20% as test set. The overall performance of the proposed method on two databases and comparison are reported in Tables 2 and 3, respectively. The best performance across all the methods are highlighted in boldface.

Notably, our proposed method yields competitive results, which is significantly outperformed than all previous SVQA methods including FR and NR models. Compared with current best performing

Table 3
Overall performance comparison on QI-SVQA database.

QI-SVQA database				
Algorithm	PLCC	SROCC	KROCC	RMSE
PSNR	0.8496	0.8637	0.6832	0.5122
SSIM	0.8185	0.8281	0.6418	0.5580
PQM [49]	0.7852	0.8165	0.6365	0.6158
PHVS-3D [50]	0.7082	0.7195	0.5353	0.7021
SFD [51]	0.6483	0.6633	0.5021	0.7571
3D-STS [52]	0.8311	0.8338	0.6553	0.5520
Feng [48]	0.8415	0.8379	0.6650	0.5372
Yang et al [20]	0.9208	0.9175	0.7730	0.3709
MNSVQM [53]	0.8823	0.8573	0.7039	0.4073
BSVQE [54]	0.9394	0.9387	0.7963	0.3543
Image-based method 1	0.9166	0.9051	0.7259	0.3891
Image-based method 2	0.8978	0.8856	0.7012	0.3908
3D CNN	0.9318	0.9284	0.7848	0.3586
3D CNN+SVR	**0.9503**	**0.9426**	**0.8038**	**0.3333**

method, we observe absolute performance gains of around 0.02 for PLCC, SROCC, KROCC and RMSE on NAMA3DS1-COSPAD1 database. And although QI-SVQA database has both symmetrically and asymmetrically distorted stereoscopic videos, we also acquire absolute performance gains of around 0.01 for PLCC, SROCC, KROCC and RMSE. These desirable experimental results demonstrate that our 3D CNN based architecture is an efficient and robust solution for evaluating the quality of whether symmetrically or asymmetrically distorted stereoscopic videos, as it has capability to capture local spatiotemporal information and global temporal information automatically in a data driven way.

In addition, we also compare our method with image-based methods and the corresponding experimental results are shown in Tables 2 and 3. There are two image-based methods we have tried. The first one only takes the difference image of each frame as the input of 2D CNN while the second one takes left and right views of each frame as input. In the second method, we set up two 2D CNN branches to encode the left view input and the right view input as representation vectors, then we concatenate these vectors into final representation of each frame. These results demonstrate the superiority of the proposed 3D CNN based method compared to the straightforward method based on 2D CNN.

We also present scatter plots of predicted quality scores by proposed method against the corresponding MOS values on two databases in Fig. 5 (a)– (b). These scatter plots illustrate that the objective scores obtained by our method have a good linear correlation with the subjective score.

4.3. 3D CNN video descriptor

In this subsection, we use our 3D CNN model as a feature extractor. In such a use case, the last fully-connected layer FC2 is removed and 512-D activations of the penultimate layer FC1 is extracted as features to represent each input cubic video patches. Then we aggregate these cube-level features to a video-level representation using L1 - norm. As a result, a 512-D feature vector is learned for each stereoscopic video by our 3D CNN model. For the sake of verifying the discrimination capability of our 3D CNN video descriptor, we embed these high dimensional feature vectors to 2D space utilizing the t-SNE toolbox [55]. Fig. 6 qualitatively visualizes and compares the capability of our 3D CNN features and the handcrafted features extracted from the very recent state-of-the-art NR SVQA method [20]. Yang et al. [20] proposed a SVQA method by modeling the binocular perception effect in multi-views, including spatial domain, temporal domain and the spatial-temporal domain, which extracts texture analysis features by associating the curvelet transform and local binary pattern have been used in the analysis

Table 4
The contribution analysis of each component on NAMA3DS1-COSPAD1 database.

3D CNN	Score Fusion	SVR	NAMA3DS1-COSPAD1			
			PLCC	SROCC	KROCC	RMSE
√	×	×	0.9216	0.9046	0.7533	0.4161
√	√	×	0.9374	0.9144	0.7592	0.4120
√	×	√	0.9410	0.9182	0.7650	0.3910
√	√	√	**0.9478**	**0.9231**	**0.7883**	**0.3614**

Table 5
The contribution analysis of each component on QI-SVQA database.

3D CNN	Score Fusion	SVR	QI-SVQA			
			PLCC	SROCC	KROCC	RMSE
√	×	×	0.9318	0.9284	0.7848	0.3586
√	√	×	0.9362	0.9335	0.7892	0.3521
√	×	√	0.9465	0.9395	0.7954	0.3485
√	√	√	**0.9503**	**0.9426**	**0.8038**	**0.3333**

of distortion on the spatial and spatial-temporal domain. As illustrated in Fig. 6, our 3D CNN features are semantically separable compared with artificially designed features, which indicates that our learned features are more effective than artificially designed features used in previous methods for SVQA.

Additionally, the 3D CNN representations are passed to a support vector regressor (SVR) instead of fully connected Multi-Layer Perceptron (MLP) for quality prediction. As shown in Fig. 5 (c)–(d), the predicted scores of 3D CNN feature extractor combined with SVR have better correlation with the MOS of testing set than using 3D CNN only. First, the learning algorithm of MLP is based on the Empirical Risk Minimization, which attempts to minimize the errors by the back-propagation algorithm in the training set. So MLP often converges on local minima rather than global minima. Furthermore, SVR aims to minimize the generalization errors on the unseen data with a fixed distribution for the training set, by using the Structural Risk Minimization principle. Therefore, the generalization ability of MLP is lower than that of SVR. These results reconfirm the fact that our 3D CNN feature extractor is superior to the artificially designed feature extractors.

4.4. The contribution analysis of each component

Our framework is made up of three components including 3D CNN, quality score fusion and SVR, where 3D CNN is the major part and the other two are the auxiliary parts. We now investigate the significance of each component and analyze their contribution via a comparison experiment. In our experiment, different combinations of these components are conducted on SVQA datasets for discussion. Tables 4 and 5 show the performance evaluation results when using each combination for predicting video quality. The symbol √ and × in the first three columns of Tables 4 and 5 indicate whether the corresponding component is used or not. Specifically, the symbol × in the column "Score Fusion Strategy" means adopting average fusion strategy instead of our proposed fusion strategy, and the symbol × in the column "SVR" means using MlP instead of SVR to complete regression. For example, the methods described in the first row of Tables 4 and 5 obtain patch-level score by 3D CNN representations with MLP and adopt average fusion strategy to pool the scores of cubic video patches into video-level score. It is clearly observed that the combination adopting all three components yields best performance. Meanwhile, our quality score fusion strategy considering global temporal clues are verified complementary to 3D CNN, and 3D CNN features in combination with SVR can further improve the performance.

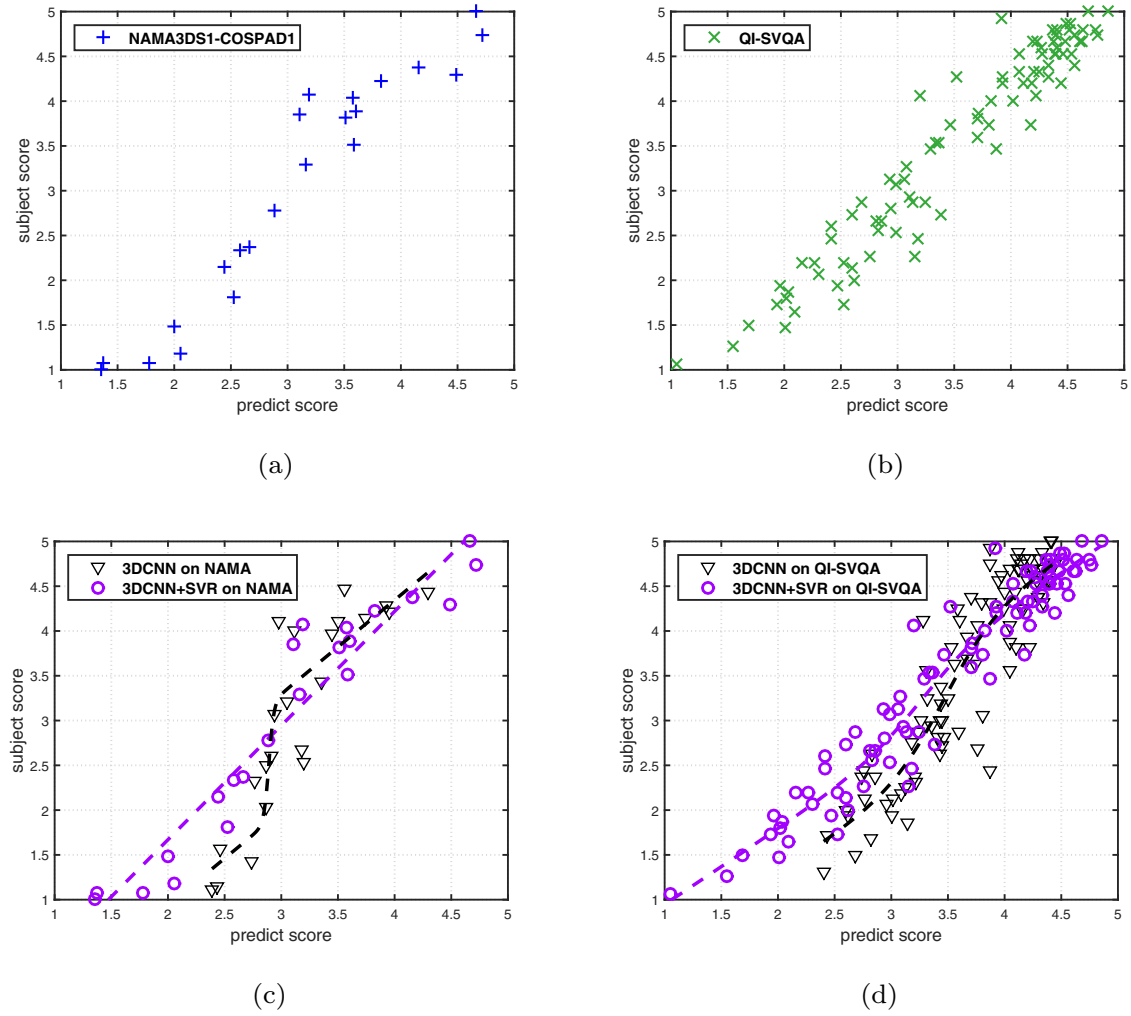
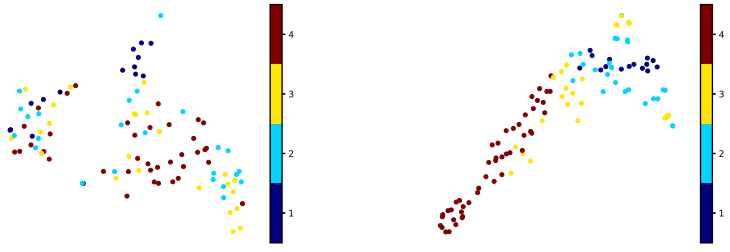


Fig. 5. Predicted MOS versus subjective MOS on two databases. (a) Scatter plot on NAMA3DS1-COSPAD1. (b) Scatter plot on QI-SVQA. (c) Performance comparison of 3D CNN and 3D CNN combined with SVR on the same data in NAMA3DS1-COSPAD1 database. (d) Performance comparison of 3D CNN and 3D CNN combined with SVR on the same data in QI-SVQA database.



(a) Hand-crafted features extracted from the very recent SVQA method

(b) Features learned by our 3D CNN model

Fig. 6. Feature embedding visualizations of state-of-the-art NR SVQA method [20] and our 3D CNN method on NAMA3DS1-COSPAD1 by t-SNE toolbox [55]. (a) Hand-crafted features extracted from the very recent state-of-the-art SVQA method. (b) Features extracted by our 3D CNN model. After dimensionality reduction, the feature vector of each stereoscopic video is visualized as a point and different colors represent different quality levels. We divide video quality into 4 grades in all.

4.5. Local quality evaluation

After produced from raw videos, numerous small cubic patches are labeled with a quality score consistent with the entire video. We can do this because we hypothesize the quality degrades homogeneously throughout the whole stereoscopic video. Next we will experimentally prove that the small cubic patches have a consistent quality condition with the entire video.

We pick out several stereoscopic videos distorted in different degradation levels of same distortion type from NAMA3DS1-COSPAD1 database, and these videos do not participate in the training stage. First, we divide each stereoscopic video into three segments vertically in spatial dimension and splice the segments

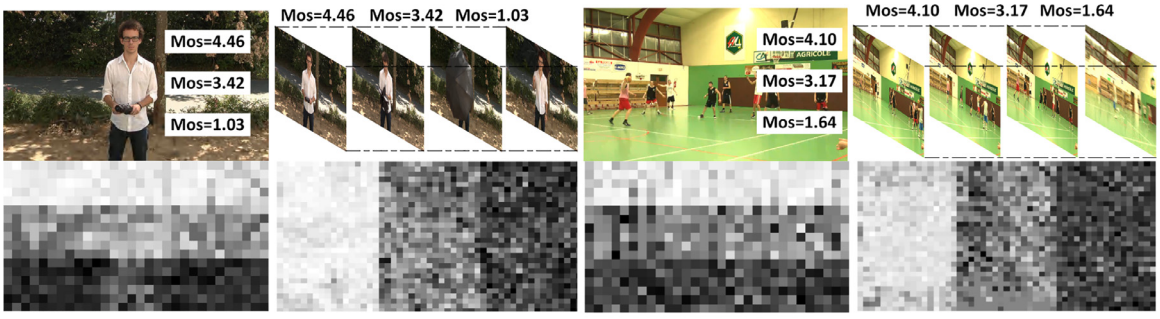


Fig. 7. Spatially and temporally synthetic stereoscopic video examples and the visualization of corresponding local quality predicted scores. The first raw shows synthetic stereoscopic videos generated by splicing video segments distorted in three different degradation levels of (a) JPEG2000 (b) JPEG2000 (c) H.264 (d) H.264. (a) and (c) are synthesized in spatial dimension while (b) and (d) are synthesized in temporal dimension. The second row shows corresponding quality maps and brighter pixels indicate higher quality.

Table 6
The performance comparison with different patch size on NAMA3DS1-COSPAD1 dataset.

Patch size	16	24	32	48	64	128	256
PLCC	0.9340	0.9429	0.9478	0.9467	0.9483	0.9480	0.9470
SROCC	0.9224	0.9228	0.9231	0.9237	0.9238	0.9225	0.9227

distorted in different levels to generate a spatially synthetic stereoscopic video. Then the similar process is carried out in temporal dimension to obtain temporally synthetic stereoscopic video. Next, our trained 3D CNN based SVQA method conducts on both spatially and temporally synthetic stereoscopic videos to implement quality evaluation. The estimated quality scores of all cubic patches are normalized into [0,255] and then visualized in a quality map. Fig. 7 presents the predicted quality map of both spatially and temporally synthetic stereoscopic video distorted in JPEG 2000 and H.264. It is clearly observed that proposed method has discrimination capability to evaluate the quality of local cubic patches, and the quality degrades homogeneously throughout the whole stereoscopic video. These results provide a much more reliable basis for our data preprocessing.

In addition, to observe how the patch size affects the overall performance, we conduct a comparative experiment on the NAMA3DS1-COSPAD1 dataset and the experimental results are shown in Table 6. To guarantee that the number of patches per video remains roughly the same when patch size varies, we adopted overlap sampling with fixed sampling stride. As Table 6 shows, as the size of the block increases, there is a very slight fluctuation and no significant improvement in performance. However, using larger video patches will spend more time when do convolution operations. Therefore, the configuration of the patch size is still 32×32 in spatial dimension.

4.6. Runtime analysis

In addition to accuracy, efficiency is also an important criterion to measure a SVQA algorithm. We develop our 3D CNN model by using the python deep learning library Keras on a PC with a single 3.2 GHz CPU and a single GTX1080 GPU. With no complex preprocessing and GPU acceleration, our proposed method is demonstrated computationally efficient. We measure the runtime of our proposed method and compare it with Yang [20], MNSVQM [53] and BSVQE [54], which are the only three NR SVQA models proposed as far as we know. Yang et al. [20] constructed a quality evaluator based on optical flow that is a time-consuming operation. For effective comparison, two implementations are adopted in our experiment to compute the optical flow: CPU implementation in Matlab and GPU implementation in OpenCV. MNSVQM extracted statistical features such as generalized Gaussian distribution (GGD), asymmetric GGD, spatial entropy, spectral entropy associated with two views, and spectral entropy related to depth perception of stereoscopic video. In BSVQE, the binocular summation and difference operations are integrated together with the fusion natural scene statistic measurement and the ARDE measurement to reveal the key influence from texture and disparity. As a result, the runtime of these methods for the whole NAMA3DS1-COSPAD1 dataset is reported in Table 7. It can be clearly observed that our method is far more efficient than the three NR methods above. Therefore, the proposed method is not only effective but also efficient, which is a more feasible solution for real-time application of SVQA.

Table 7
The runtime analysis on NAMA3DS1-COSPAD1 dataset.

Method	Usage	Runtime(h)	Test time per video(s)
Yang et al. [20]	CPU	117.5	250
Yang et al. [20]	GPU	29.4	63
MNSVQM [53]	CPU	67.9	45
BSVQE [54]	CPU	15.4	18
Proposed method	GPU	2.6	2

5. Conclusion

In this paper, we have presented a NR SVQA framework based on 3D CNN, which can effectively model not only local spatiotemporal information but also global temporal information with cubic difference video patches as input. In the framework, we first design a 3D CNN architecture to automatically capture local spatiotemporal features instead of using hand-crafted features and then employ a quality score fusion strategy considering global temporal clues to obtain final video-level predicted scores. Extensive experiments on two challenging stereoscopic video databases have shown that our proposed method correlates highly with human perception and significantly outperforms state-of-the-art methods. In addition, with no complex preprocessing and GPU acceleration, our proposed method is demonstrated computationally efficient compared with previous methods.

Despite deep learning based methods have achieved great success in many challenge tasks, few network architectures are proposed to evaluate the quality of visual information, especially the quality of stereoscopic video. Our work in this paper explores 3D CNN to evaluate the quality of stereoscopic video for the first time. In the future, we will continue to focus on developing deep learning models for SVQA task. Moreover, we plan to establish a large SVQA dataset to address the problem of labeled data scarcity, which definitely meets the needs of deep learning models.

Acknowledgments

This work was supported by the Foundation of Pre-Research on Equipment of China (No.61403120103) and the National Natural Science Foundation of China (Nos. 61372130, 61432014).

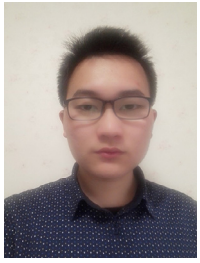
References

[1] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, A.C. Kot, No-reference image blur assessment based on discrete orthogonal moments, IEEE Trans. Cybern. 46 (1) (2016a) 39–50.
[2] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, B. Chen, No-reference quality assessment of deblocked images, Neurocomputing 177 (2016b) 572–584.
[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[4] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Good practice in large-scale learning for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 507–520.
[5] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[7] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[8] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, X. Xue, Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2015, pp. 461–470.
[9] C. Feichtenhofer, A. Pinz, R. Wildes, Spatiotemporal residual networks for video action recognition, in: Proceedings of the Advances in Neural Information Processing Systems, 2016a, pp. 3468–3476.
[10] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016b, pp. 1933–1941.
[11] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740.
[12] W. Zhang, C. Qu, L. Ma, J. Guan, R. Huang, Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network, Pattern Recognit. 59 (2016) 176–187.
[13] C. Hewage, S.T. Worrall, S. Dogan, A. Kondoz, Prediction of stereoscopic video quality using objective quality models of 2D video, Electron. Lett. 44 (16) (2008) 963–965.
[14] S. 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).
[15] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[16] H. Malekmohamadi, A. Fernando, A. Kondoz, A new reduced reference metric for color plus depth 3D video, J. Vis. Commun. Image Rep. 25 (3) (2014) 534–541.
[17] H. Zhu, M. Yu, Y. Song, G. Jiang, A stereo video quality assessment method for compression distortion, in: Proceedings of International Conference on Computational Science and Computational Intelligence, IEEE, 2015, pp. 481–485.
[18] M. Yu, K. Zheng, G. Jiang, F. Shao, Z. Peng, Binocular perception based reduced-reference stereo video quality assessment method, J. Vis. Commun. Image Rep. 38 (2016) 246–255.
[19] C. Galkandage, J. Calic, S. Dogan, J.-Y. Guillemaut, Stereoscopic video quality assessment using binocular energy, IEEE J. Sel. Top. Signal Process. 11 (1) (2017) 102–112.
[20] J. Yang, H. Wang, W. Lu, B. Li, A. Badii, Q. Meng, A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain, Inf. Sci. 414 (2017) 133–146.
[21] C. Li, A.C. Bovik, X. Wu, Blind image quality assessment using a general regression neural network, IEEE Trans. Neural Netw. 22 (5) (2011) 793.
[22] A. Chetouani, A. Beghdadi, S. Chen, G. Mostafaoui, A novel free reference image quality metric using neural network approach, in: Proceedings of the International Workshop Video Process. Quality Metrics Consumer Electron, 2010, pp. 1–4.
[23] L. Li, W. Lin, H. Zhu, Learning structural regularity for evaluating blocking artifacts in JPEG images, IEEE Signal Process. Lett. 21 (8) (2014) 918–922.
[24] H. Tang, N. Joshi, A. Kapoor, Blind image quality assessment using semi-supervised rectifier networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2877–2884.
[25] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[26] H. Tang, N. Joshi, A. Kapoor, Learning a blind measure of perceptual image quality, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 305–312.
[27] D. Ghadiyaram, A.C. Bovik, Blind image quality assessment on real distorted images using deep belief nets, in: Proceedings of the IEEE Global Conference on Signal and Information Processing, IEEE, 2014, pp. 946–950.
[28] W. Hou, X. Gao, D. Tao, X. Li, Blind image quality assessment via deep learning, IEEE Trans. Neural Netw. Learn. Syst. 26 (6) (2015) 1275–1286.
[29] F. Shao, W. Tian, W. Lin, G. Jiang, Q. Dai, Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions, IEEE Trans. Image Process. 25 (5) (2016) 2059–2074.
[30] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, K.-W. Cheung, No-reference video quality assessment with 3D shearlet transform and convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 26 (6) (2016) 1044–1057.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[34] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[35] A. Jain, J. Tompson, Y. LeCun, C. Bregler, MoDeep: a deep learning framework using motion features for human pose estimation, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2014, pp. 302–315.
[36] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[38] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V.C. Mok, L. Shi, P.-A. Heng, Automatic detection of cerebral microbleeds from mr images via 3D convolutional neural networks, IEEE Trans. Med. Imaging 35 (5) (2016) 1182–1195.
[39] 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) (2015a) e0145800.
[40] 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 Rep. 31 (2015b) 138–145.
[41] 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.
[42] L. Ma, X. Wang, Q. Liu, K.N. Ngan, Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment, Neurocomputing 215 (2016) 21–31.
[43] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, N. Garcia, NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences, in: Proceedings of Fourth International Workshop on Quality of Multimedia Experience, IEEE, 2012, pp. 109–114.
[44] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, Handb. Brain Theory Neural Netw. 3361 (10) (1995) 1995.
[45] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[46] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the International Conference on Machine Learning, 2010, pp. 807–814.
[47] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, Comput. Sci. 3 (4) (2012) 212–223.
[48] 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.
[49] P. Joveluro, H. Malekmohamadi, W.C. Fernando, A. Kondoz, Perceptual video quality metric for 3D video quality assessment, in: Proceedings of the 3DTV–Conference: The True Vision-Capture, Transmission and Display of 3D Video, IEEE, 2010, pp. 1–4.
[50] L. Jin, A. Boev, A. Gotchev, K. Egiazarian, 3D-DCT based perceptual quality assessment of stereo video, in: Proceedings of IEEE International Conference on Image Processing, IEEE, 2011, pp. 2521–2524.
[51] 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 Con-

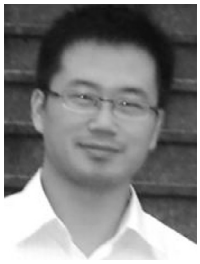
ference: The True Vision-Capture, Transmission and Display of 3D Video, IEEE, 2009, pp. 1–4.
[52] J. Han, T. Jiang, S. Ma, Stereoscopic video quality assessment model based on spatial-temporal structural information, in: Proceedings of the Visual Communications and Image Processing, IEEE, 2012, pp. 1–6.
[53] G. Jiang, S. Liu, M. Yu, F. Shao, Z. Peng, F. Chen, No reference stereo video quality assessment based on motion feature in tensor decomposition domain, J. Vis. Commun. Image Rep. 50 (2018) 247–262.
[54] Z. Chen, W. Zhou, W. Li, Blind stereoscopic video quality assessment: from depth perception to overall experience, IEEE Trans. Image Process. PP (99) (2018) 721–734.
[55] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

Jiachen Yang received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2005 and 2009, respectively. He is currently a professor at Tianjin University. He is a visiting scholar with the department of computer science, school of science, Loughborough University, U.K. His research interests include multimedia quality evaluation, stereo vision research and pattern recognition.

Yinghao Zhu received the B.S. degree in communication and information engineering from Hohai University, Nanjing, China, in 2016. He is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include multimedia quality evaluation, stereo vision research and pattern recognition.



Chaofan Ma received the B.S. degree in materials science and engineeringfrom Tianjin University, Tianjin, China, in 2016. He is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include stereo vision research, pattern recognition and deep learning.



Wen Lu received the M.S. and Ph.S. degrees in electrical engineering from Xidian University, China, in 2006 and 2009, respectively. He is currently a professor at Xidian University. His research interests include image and video understanding, visual quality assessment, and computational vision.

Qinggang Meng received the B.S. and M.S. 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 a Senior Lecturer with the Department of Computer Science, Loughborough University, U.K. His research interests include biologically and psychologically inspired learning algorithms and developmental robotics, service robotics, robot learning and adaptation, multi-UAV cooperation, drivers distraction detection, human motion analysis and activity recognition, activity pattern detection, pattern recognition, artificial intelligence, and computer vision. He is a Fellow of the Higher Education Academy, U.K.