This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

No Reference Quality Assessment of Stereo Video Based on Saliency and Sparsity

Jiachen Yang, *Member, IEEE*, Chunqi Ji, Bin Jiang, Wen Lu, *Member, IEEE*, and Qinggang Meng, *Member, IEEE*

***Abstract*—With the popularity of video technology, stereoscopic video quality assessment (SVQA) has become increasingly important. Existing SVQA methods cannot achieve good performance because the videos' information is not fully utilized. In this paper, we consider various information in the videos together, construct a simple model to combine and analyze the diverse features, which is based on saliency and sparsity. First, we utilize the 3-D saliency map of sum map, which remains the basic information of stereoscopic video, as a valid tool to evaluate the videos' quality. Second, we use the sparse representation to decompose the sum map of 3-D saliency into coefficients, then calculate the features based on sparse coefficients to obtain the effective expression of videos' message. Next, in order to reduce the relevance between the features, we put them into stacked auto-encoder, mapping vectors to higher dimensional space, and adding the sparse restraint, then input them into support vector machine subsequently, and finally, get the quality assessment scores. Within that process, we take the advantage of saliency and sparsity to extract and simplify features. Through the later experiment, we can see the proposed method is fitting well with the subjective scores.**

***Index Terms*—Stereoscopic video quality assessment, saliency, sparse representation, stacked auto-encoder (SAE), sparsity.**

I. Introduction

IN THE contemporary, the 3D technology developed rapidly, such as virtual reality, 3D films, and 3D display devices. Accordingly, the 3D videos are widely used in various domains, and they can bring the more reality sensations than 2D videos [1]. As we all known, the quality related to the videos has a direct effect on the subjective sensation about human eyes. When arising the distortions during the compression, transmission or presentation, the quality about videos will be decreased at the receiving end, and then people will feel uncomfortable even disgusting when watching them. In other words, the users' experience is worse, so we need a metric to measure the distortions. On the other side, the subjective evaluation is tedious and time-consuming [2], so it is necessary to evaluate the quality on the videos by algorithms [3].

The relevant algorithms can be divided into three aspects: full-reference (FR), reduced reference (RR) and no-reference (NR) [4]. As the name suggests, the FR methods acquire the pristine video source, RR algorithms require the partial message about origin videos, when the NR assessments doesn't need any additional information [1]. Obviously it has many restrictions on the FR and RR, because it is difficult to obtain the source video in real life. Therefore, the NR algorithms have more practical significance and applicable value [5]. But from the another aspect, due to lack of undistorted counterpart, there has more obstacles when design an effective NR approaches, so according to the tendency in the future, the NR method will develop faster than other two methods, and it still has plenty of room for growth.

With the help of quality assessments on images, many methods on video's quality evaluations appeared subsequently. The relevant thought appeared at the beginning is simply learned from the practice about the image quality assessment (IQA), such as SSIM [6], VSI [7], BRISQUE [8], DIIVINE [2] and other measures. These algorithms expected to obtain the videos' quality scores by means of evaluating per frames separately, then combined them in the manner of weighted average, treated as video quality assessment (VQA).

However, the final effects of these methods are unsatisfactory. This is imaginable because the approaches abovementioned only pay attention to the spatial features but ignore the movement features. Then several approaches concerned about this point. They added the temporal features and got the better performance. Saad *et al.* [9] proposed a 3D VQA, it utilized the DCT-domain features on each frame to get spatial information, then combined the natural scene statistic model to approach the human visual perception process. Han *et al.* [10] constructed a 3D spatial-temporal structural (3D-STS) model to integrate the peculiarities between spatial and temporal on inter-frames. These ideas pointed out the direction for the further research.

Meantime, Human Visual System (HVS) is also an indispensable factor in quality assessment. Compared with 2D quality assessment, there are a few major points need to be considered [11] in stereovision, such as the disparity between two views, binocular fusion and binocular rivalry phenomenon. Shao *et al.* [12] proposed new feature encoding

Manuscript received August 21, 2017; revised November 20, 2017; accepted December 20, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61471260, and in part by the Natural Science Foundation of Tianjin under Grant 16JCYBJC16000. *(Corresponding author: Bin Jiang.)*

J. Yang, C. Ji, and B. Jiang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 30072, China (e-mail: yangjiachen@tju.edu.dn; jcq_@tju.edu.cn; jiangbin@tju.edu.cn).

W. Lu is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: luwen@xidian.edu.cn).

Q. Meng is with the Department of Computer Science, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: q.meng@lboro.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2018.2789583

0018-9316 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

and similarity measure approaches to model visual properties of the primary visual cortex, which yielded better results. Yu *et al.* [13] divided visual model into binocular fusion portion and binocular rivalry portion, extracted features from each part respectively, finally pooled and got the excellent performance. In [14] and [15], binocular vision model was combined with the saliency model to approach the perception process. In [16], respective fields of the human visual cortex were simulated and support vector regression was used to establish connections between them. These efforts had promoted the development of quality assessment.

In addition, a number of studies based on visual attention and machine learning had achieved rapid development, these work got excellent performance on 3D VQA. Zhao *et al.* [17] constructed evaluation algorithm based on edge differences, visual attention and depth information. Oprea *et al.* [18] estimated the perceptually important areas through salient region detection, then computed distortion measure to obtain perceptual result. In [19], a novel VQA metric was proposed through visual perception and visual attention. Shahid *et al.* [20] presented a model based on LS-SVM (least Square Support Vector Machines), which attained more robust features. In [21], DBN (Deep Belief Nets) was introduced to make more accurate predictions. Narwaria and Lin [22] designed a method on the basis of SVR (Support Vector Regressor) and SVD (Singularly Valuable Decomposition) features to capture structural information for videos. Inspired by previous ideas, we take advantage of visual attention and machine learning in our work.

In this paper, we introduce a method that combines the saliency, sparse representation and stacked auto-encoder (SAE). The contributions in our paper are the following:

1) Taking the advantages of sum-difference theory and 3D saliency, integrating the information an videos in a novel and simple way.

2) Employing the sparse representation to extract video's features, to the best of our knowledge, we are the earliest to adopt the method of sparse representation in SVQA.

3) Putting forward an innovative work that combines the saliency and sparsity to extract features on frames, validating the reasonable and effectiveness of proposed approach.

4) Handling the video's features with SAE based on the principle of sparseness, further simplified features by neural network.

The remainder of this paper is organized as follows. Section II introduces the previous work associated with proposed metric briefly. Section III details the concrete steps of proposed method. In Section IV, the performance of proposed framework and related experiments are presented. Finally, we make a brief summary in Section V.

II. Related Work and Motivation

A. Saliency Map

When people observe a scene, though the visual information they achieve may be complicated, human eye's will be attracted by some preferred areas, which named saliency [23]. It reflects the places where people usually pay more attention to play a important role on visual perception. The related researches grow rapidly as the deeper awareness on visual system [24]. The early work is in [25], Itti *et al*. derived saliency map by means of calculating the images' pixel intensity, orientation and color's contrast. Then many improved algorithms emerged in large numbers, like the method based on natural statistics in [26], the algorithm which mainly utilized the low-level features in [27], the approach which considered the spectral residual in [28], etc. Some VQA methods also viewed saliency as a powerful tool. Yang *et al.* [29] integrated salient region and edge difference into human visual system, Jia *et al.* [30] weighted salient and non-salient region to calculate degradation, Culibrk *et al.* [31] used a multi-scale background-subtraction assessment approach based on salient motion, Wei and Zhang [32] constructed statistics features combined with saliency map for evaluation.

We refer the relevant work in [33], which computed the statistical uncertainty measures to combine the spatial and temporal information. we understand the saliency from another perspective, regard the saliency as an efficient way which reflects the relationship between the spatial and temporal. Different from the related work previous mentioned, we do not directly divide significant and non-significant regions through saliency maps, but process them together to ensure integrity. Besides that, the saliency of the left and right view is combined with sum map in consideration of binocular vision. The details are shown in Section III-A.

B. Sparse Representation

Sparse representation is usually used in the signal processing [34]. The main idea on sparse representation is to express the signals simply and effectively [35] with the linear combination of a small number of elementary signals [36]. The theory about sparse representation also conforms to the human eyes' visual characteristic [37]. When human eyes observe an image, optical signal will be processed on the retina. Cells in the retina encode received visual signal, then information is encoded as complex statistical dependencies among the photoreceptor activities on lateral geniculate nucleus (LGN). And primary visual cortex reduces these statistical dependencies to discover the intrinsic structures, finally form the information of image' features in the brain [38].

This phenomenon can be mimicked as describing images with a linear superposition of a small number of basis vectors. Therefore, a larger, over-complete set (namely dictionary) of basis vectors is able to adapt to images and best represent all structures information of images. It also means that basis elements of the dictionary are similar to the receptive field of the cortical simple cells. It is also consistent with the process of sparse representation. Sparse representation decomposes the image signal through dictionary, which is similar to the procedure of encoding light information on the area of retina [39]. In addition, related studies indicated that the natural images' structure had redundancy and would be simplified in receptive field, and sparse coding is a reasonable and valid method to model the biological phenomenon.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

The basic process of sparse representation is as follows: first divide the image into patches $\{X_c\}, \{X_c, c = 1, 2, \ldots n, X_c \in R^{h^2}\}$, h^2 is size of pathes, and n represents the total number of patches. Then the K-SVD [40] algorithm is adopted to solve the problem about dictionary generation and optimization. The first step is fixing the dictionary D and looking for the optimal sparse matrix A. Let $\mathbf{a}_i$ to be the i-th column vector of matrix A, $\mathbf{x}_i$ to be the i-th column vector of batch X, then take the sparse coding as foundation, the optimization of dictionary can be formulated as:

$$D = argmin \sum_i \|\mathbf{x}_i - D\mathbf{a}_i\|_2^2 \, s.t. \, \|\mathbf{a}_i\|_0 \leq T_0 \tag{1}$$

T_0 is a constant which is chosen according to the experience. $\|\mathbf{a}_i\|_0$ is 0-norm of $\mathbf{a}_i$, which represents the number of zero elements in $\mathbf{a}_i$. It is a NP hard problem, Orthogonal Matching Pursuit (OMP) algorithm is usually applied to solve it. The next step is to update dictionary D column by column. The specific approach is fixing the coefficient matrix A and dictionary D, updating the k-th column $\mathbf{d}_k$ of D, multiplying the sparse matrix A and $\mathbf{d}_k$. Each column of result is denoted as $\mathbf{a}_\mathbf{T}^j$, (j = 1, 2, , . . . k) the formula (4) can be rewritten as:

$$\begin{aligned} \|X - DA\|_F^2 &= \left\| X - \sum_{j=1}^{k} \mathbf{d}_j \mathbf{a}_\mathbf{T}^j \right\|_F^2 \\ &= \left\| \left(X - \sum_{j \neq k} \mathbf{d}_j \mathbf{a}_\mathbf{T}^j \right) - \mathbf{d}_k \mathbf{a}_\mathbf{T}^k \right\|_F^2 \\ &= \left\| E_k - \mathbf{d}_k \mathbf{a}_\mathbf{T}^k \right\|_F^2 \end{aligned} \tag{2}$$

Then SVD (singular value decomposition) is used to decompose the difference matrix E_k:

$$E_k = U \nabla V \tag{3}$$

$\mathbf{d}_k$ is updated with the first atom in first column of U, $\mathbf{a}_\mathbf{T}^k$ is replaced with the product of the first column of V and $\nabla(1, 1)$. Iterate these steps, finally we obtain an optimal dictionary, which is used for decomposing per frame into coefficients. The formulation of the decomposition process is as follows:

$$I = \left(\sum_{i=1}^{n} R_i^T R_i \right)^{-1} \sum_{i=1}^{n} \left(R_i^T D \mathbf{a}_i \right) \tag{4}$$

I denotes the images that need to be reconstructed, and R_i^T is a matrix operator which extracts patches from image. Then the OMP algorithm is utilized to solve the equation in a similar manner, attaining a set of coefficients which are used to construct features.

The above is the basic theory of sparse representation. Due to its excellent properties, some IQA methods using sparse represent appear as well. In [41], MUMBLIM model based on sparsity was presented to predict objective scores. Chang *et al.* [42] used sparse feature fidelity (SFF) to measure distortion. Shao *et al.* [43] simulated monocular and binocular visual perception through sparse representation. These methods exhibited good performance on evaluating the quality of stereoscopic images.

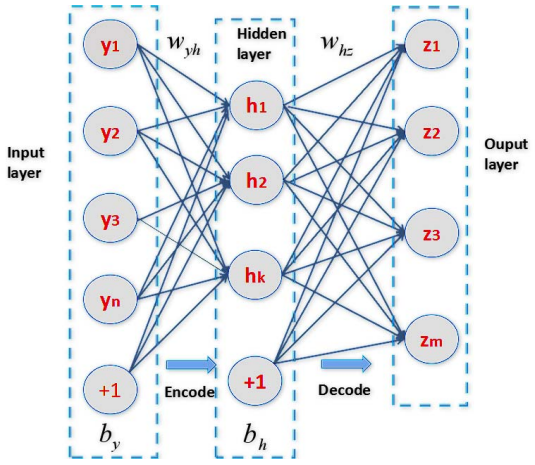


Fig. 1. The structure of auto-encoder.

Inspired by previous work, we perform the strategy on videos to dig the deeper structure message, and the majority of content is referred in [38]. What is different from previous work is that we employ sparse representation on videos rather than images, so the field of application on algorithm is extended from spatial domain to spatial-temporal domain; in addition, we connect sparse theory with association of saliency, the innovative combination can take advantage of two parts.

C. Stacked Auto-Encoder

In the phase of processing features, the selected regression algorithms are different in relevant work. As mentioned earlier, LS-SVM was employed in [20], SVR was chosen in [22], DBN was selected in [21]. In this work, we choose SAE as a tool for feature handling. Compared with previous work, SAE can learn features effectively and express features compactly. And it can enhance the expression ability of the hidden layer by using sparsity, which is in line with visual system.

SAE is a sort of deep-learning (DL) network [44]. It consists of automatic encoders stacked in series. The purpose of stacking multi-layer auto-encoders is extracting high-order feature of input data hierarchically and reducing the dimension of input data during the process. In this way, complex input data is converted into a series of simple high order features. It constructs a more sparse structure, conforming to the processing of signals by human eyes cells to a certain extent, which is better than SVM at this point.

The fundamental unit of SAE named auto-encoder (AE) [45]. It is designed to make the output vector to identical with input vector. It can be divided into three parts, which include input layer, hidden layer and output layer. The structure of AE is shown in Fig. 1, symbol w means the weight between two units and b means bias.

AE minimizes the reconstruction error of the input data by two steps of encoding and decoding, thereby obtaining the best data expression for hidden layers. As the figure shows, denote the training sets as $\{(y_1), (y_2), \ldots (y_n)\}$, when $y_i (i = 1, 2, \ldots n)$ is input in the first hidden layer, hidden units are activated by encoding, and the intermediate quantity

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

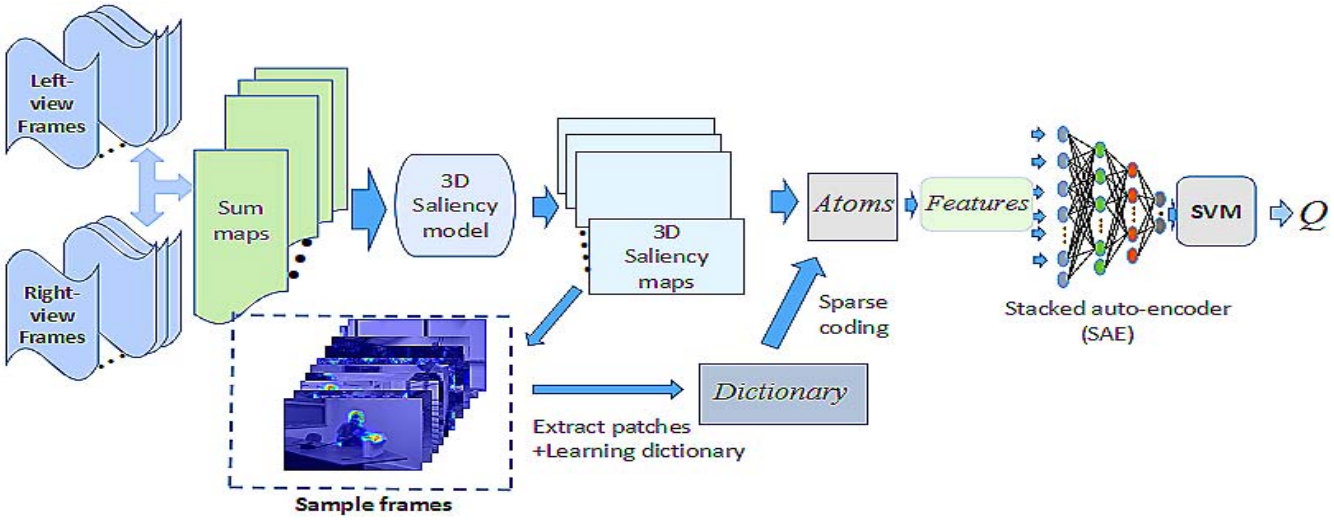


Fig. 2. The framework of proposed method.

$h(y_i)$ is generated. Then through the decoding process, $z(y_i)$ is reconstructed in the output layer. In this way, the data on previous layer is expressed approximately. The encoding and decoding formulas are:

$$h(y_i) = f\left(W_{yh}y_i + b_y\right) \tag{5}$$

$$z(y_i) = g(W_{hz}h(y_i) + b_h) \tag{6}$$

W_{yh}, W_{hz},b_y, b_h are encoding matrix, decoding matrix, encoding bias, decoding bias respectively. f and g are activation functions, sigmoid function is employed in this paper. The target of the network is to minimize the reconstruct error $L(Y, Z)$. $L(Y, Z)$ may be squared error loss or cross entropy loss. We select squared error loss as cost function. Assume θ as optimum parameter, θ can be expressed as:

$$\hat{\theta} = \arg\min_{\theta} L(Y, Z) = \arg\min_{\theta} \frac{1}{2n}\sum_{i=1}^{n}\|y_i - z(y_i)\| \tag{7}$$

Through the above formulas, the similar expression of the input data can be obtained in the output layer. Replacing the output layers with hidden layers, we acquire the basic structure of SAE. First to use origin data for training first layer, then derive activation value $h(y_i)$ by weight matrix and bias, and regard $h(y_i)$ as input for second layers. The same strategy is adopted and repeated. Finally, through the process of fine-tuning, the approximate expression is output. It is noteworthy that we can make the number of nodes in the output layer smaller than the input layer to simplify original features, which compresses the dimensions of data. We attempt to utilize SAE to promote the features' performance. But different from the above process, sparse constraint is added in the reconstruction error, which reflects the sparsity in SAE. The practice is presented in the following section, and it is verified to be reasonable in the experiment.

III. Proposed Framework

Our proposed work is based on the theory abovementioned. Fig. 2 shows a framework which corresponds to the work. As is shown in the figure, the general framework is mainly divided into three parts. First we extract images form videos per 8 frames at left and right view videos, (the number of selected frames can be slightly changed as needed, must guarantee that it doesn't lose too much information on videos), add the left and right views together, obtain their 3D visual saliency map on the basis of the method in [33]. Next, we choose different frames' map at diverse scenes, combine into one map and use it to train a dictionary, then employ sparse coding on all saliency maps of videos. In this procedure we get the coefficients that enable to capture per frames' structure, then extract features based on these coefficients and feed them to a stacked auto-encoder (SAE), finally input processed features into support vector machine (SVM) to get the scores about videos' quality. We will demonstrate the framework and verify the validity on each part in subsequent subsections.

A. Visual Theories and Saliency

For the 3D scene, in addition to the saliency mentioned in Section II, binocular vision is also an important factor that need to be taken into account [46]. So we try to combine saliency with visual theory. When human observes an object, it will produce the binocular vision [47], making the left and right views merge and relate with each other. So it is necessary to mix two eyes' views together. We refer to the work mentioned in [48], which introduced the sum and difference channels with binocular visual characteristic. Denote I_L as left view of a frame, I_R as right view of a frame, by definition, the sum map and difference map are computed as:

$$I_{sum} = (I_L + I_R)/2, \tag{8}$$

$$I_{dif} = |(I_L - I_R)|. \tag{9}$$

These two types of maps can effectively express the image information received by two eyes in a simply way. Therefore, we draw lessons form this theory. But our work is differing from previous work. In [48], it showed difference map play a more important role on visual perception; in this paper, we

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

only adopt the sum maps [49]. Because we find that difference map can't fully reflect the content of each frame. Just as Fig. 9 (e) shows, it retains more edges and contours, in contrast, background or other trivial details are almost neglected. So we try to employ the sum map to solve this problem. On the one hand, we expect to further simplify the structure, on the other hand, it is beneficial for the follow-up work. We hope to get the integrated message on sampled frames, so difference maps can't meet our requirements.

In order to reflect human visual system and preserve original information, sum map is chosen to compute saliency. After obtaining the sum maps from video's frames, we calculate the 3D saliency map on them, treating as the videos' saliency maps. The corresponding definition is:

$$I_{sum_s} = (I_L + I_R)_s/2 \tag{10}$$

We mainly derive the saliency model based on [33]. Convert the color space into YCbCr at first, then divide sum maps into non-overlapping patches. Next calculate the spatial saliency on patch i according to the following formula:

$$S^i_{k(sum)} = \sum_{j \neq i} \left[\frac{1}{\sqrt{2\pi}\sigma_s} e^{-l_{ij}^2/2\sigma_s^2} \right] D^k_{ij} \tag{11}$$

where k indicates the groups of features which include luminance, color and texture, D^k_{ij} denotes the feature difference between the patch i and j, σ_s is a weight of D^k_{ij}, l_{ij} is the spatial distance between two patches. The spatial saliency $S^{(s)}$ is computed as:

$$S^{(s)}_{sum} = \frac{1}{K} \sum_k N(S^k_{sum}) \tag{12}$$

where N is normalize function. Meantime, the temporal saliency about patch i is calculated according to the following formula:

$$S^{(t)}_{sum} = -logp(v) = \alpha logv + \beta. \tag{13}$$

In the formula, $p(v)$ is a power-law function with regard to the prior distribution of motion speed, v is the relative speed of the i-th patch, which is defined as:

$$v_i = \sum_{j \neq i} \left[\frac{1}{\sqrt{2\pi}\sigma_i} e^{-l_{ij}^2/2\sigma_i^2} \right] D^v_{ij} \tag{14}$$

where D_{ij} means the length of the vector difference between the mean absolute motion vectors of patches i and j.

The overall 3D saliency maps consist of spatial maps, temporal maps and uncertainty for each pixel accordingly. The final expression is:

$$S_{sum} = \frac{U^{(t)} S^{(s)}_{sum} + U^{(s)} S^{(t)}_{sum}}{U^{(s)} + U^{(t)}} \tag{15}$$

$U^{(s)}$ and $U^{(t)}$ are uncertainty maps on spatial and temporal respectively, they are calculated through uncertainty computation U on both spatial and temporal saliency maps. U is defined as:

$$U = H_b(p(s|d)) + H_b(p(s|c)) \tag{16}$$

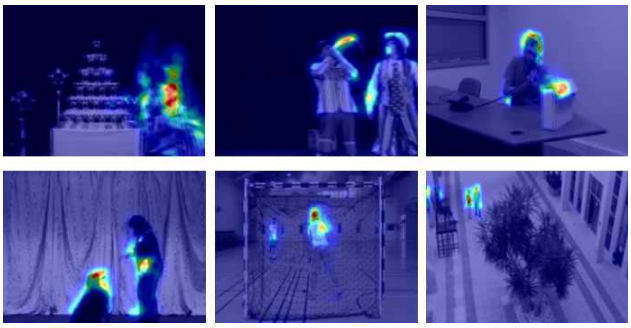


Fig. 3. 3D saliency sum maps of sampled frames in video databases.

where H_b is a binary entropy function, $p(s|d)$ measures the salient probability of a pixel based on the distance d from saliency center, $p(s|c)$ computes the salient probability of a pixel based on the connectedness c from saliency center. More specific description about parameters and concepts are elaborated in [33].

There are two points that need to be explained. Firstly, we are not just focusing on the saliency. Saliency is significant for visual information, but it doesn't mean the background is useless. The combination of background and saliency can show the characteristics of video better. Secondly, we think that saliency is an important expression on the interaction between time and space. In our opinion, spatial and temporal is not mutually independent, the variation of the spatial pixels provides motion information and temporal attention for the time domain; in contrast, the flow of temporal embodies the spatial saliency in the video. So it is more accord with perception to merge the spatial and temporal together. We pick out some sum maps' saliency map of different type of videos, exhibiting them in Fig. 3. Through the group of illustrations, it is clearly that the saliency maps represent more useful message in videos, and the effect of saliency is well on sum maps. The highlight part emphasizes the motion intensity and orientation at saliency position of the frames, the larger percentage of its share in frame, the more movement appears in the videos. It displays the motion information in temporal domain visually.

The effectiveness and rationality of our manners will be confirmed in the experimental part, by means of contracting with other synthetic images' saliency map.

B. Saliency and Sparse Representation

The next step is to extract features on saliency maps. We adopt the sparse representation method, it can analyze the information of the frame effectively and reflect the influence of saliency characteristics on visual elements. We select nine frames' 3D saliency map at nine videos, forming an image that used to be trained a dictionary. The source images are shown in the lower left portion of Fig. 2.

In Section II-B, image is divided into n patches of h^2 size (h is 8 in this experiment). But after adding saliency, the effects of saliency on various blocks are different. Obviously, in the saliency part of the image, the change of pixel value is greater. Denote pixels in origin image as I_{io}, pixels in saliency map as I_{is}, and N_I indicates the total number of pixels in image.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

The average variation of pixels value is denoted as $\Delta\bar{I}$, it is formulated as follows:

$$\Delta\bar{I} = \frac{\sum_i^{N_I}(I_{is} - I_{io})}{N_I} \tag{17}$$

Just as mentioned in the previous chapter, the areas highlighted by color (namely the saliency part) symbolize the motion intensity and tendency. The saliency map changes the content of the original image, making the salient part more prominent. It is obvious that the patches which contain much saliency will be different from other 'static' patches, the pixel value of them will be higher than the counterpart in the original image, and the coefficients will be changed as well. And in most cases, background information is the majority in the scene, so most of the regions are not salient. Under this condition, $\Delta\bar{I}$ in the above formula can be an effective criterion to distinguish from salient blocks and non-salient blocks. Denote the salient blocks as X_{c_s}, non-salient blocks as X_{c_n}, original blocks as X_o they are defined as:

$$X_c = \begin{cases} X_{c_s}, & when\ X_c - X_o \geq h^2\Delta\bar{I} \\ X_{c_n}, & when\ X_c - X_o < h^2\Delta\bar{I} \end{cases} \tag{18}$$

Accordingly, the expression of dictionary should also be changed. The following formula is more consistent with the situation of adding saliency:

$$\begin{aligned} \hat{D} = argmin & \left(\sum_{\mathbf{x}_i \in X_{cs}} \left\| \mathbf{x}_i - \hat{D}\mathbf{a}_i \right\|_2^2 + \sum_{\mathbf{x}_i \in X_{cn}} \left\| \mathbf{x}_i - \hat{D}\mathbf{a}_i \right\|_2^2 \right) \\ s.t. \quad & \|\mathbf{a}_i\|_0 \leq T_0 \end{aligned} \tag{19}$$

where $\hat{D}$ is defined to distinguish the dictionary which is derived from ordinary image. By the same reason, Eq.2 is changed as:

$$\left\| \sum X_{c_s} + \sum X_{c_n} - \hat{D}A \right\|_F^2 = \left\| \hat{E}_k - \hat{\mathbf{d}}_k \hat{\mathbf{a}}_{\mathbf{T}}^k \right\|_F^2 \tag{20}$$

$\hat{\mathbf{d}}_k$ and $\hat{\mathbf{a}}_{\mathbf{T}}^k$ are also defined to distinguish the original variables $\mathbf{d}_k$ and $\mathbf{a}_{\mathbf{T}}^k$. Through the aforementioned formula, a series of coefficients which represent image are generated. Based on the conception in [38], the dictionary $\hat{D}$ can represent visual primitive set in theory, so we can conclude that coefficients in sum maps' saliency map are similar with cells on the retina to some extent. Parallel with the process of integrating optimal signal in LGN, we can extract features which fit to the biological vision perception from these coefficients. But due to the number of coefficients is large, and coefficients are too trivial to manifest the valid features in the videos clearly, we decide to get features through coefficients' entropy. Entropy can be used to measure the information quantities on the basis of Shannon theory [50]. It is an effective metric to describe visual perceptual information.

Denote the coefficients as a_k, the probability value is calculated as:

$$p_k = \frac{a_k}{\sum_{k=1}^N a_k}(k = 1, 2, \ldots N) \tag{21}$$

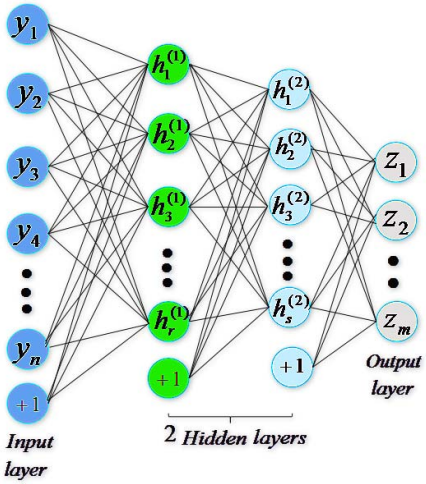


Fig. 4. The structure of SAE in the experiment, which has 2 hidden layers.

By definition, the entropy value is computed through the following formula:

$$H = -\sum_{k=1}^{N} p_k log p_k \tag{22}$$

To explore the regularity on frames' coefficients, we adopt three statistical modes: mean (abbreviates to m), standard deviation (abbreviates to s) and 2-norm (abbreviates to n). Assuming the restriction number of non-zero is L, we attain L entropy value from a frame. If the number of sample frames is M, denote the per frame's entropy as H_{lm}, the corresponding formulas are:

$$m_l = \frac{\sum_{i=1}^M H_{lm}}{M}, (l = 1, 2, \ldots L) \tag{23}$$

$$s_l = \sqrt{\frac{\sum_{1=1}^M (H_{lm} - m_l)^2}{M}}, (l = 1, 2, \ldots L) \tag{24}$$

$$n_l = \left[\sum_{i=1}^M (H_{lm})^2 \right]^{1/2}, (l = 1, 2, \ldots L) \tag{25}$$

We get the $3L$ features from each video. These features measure the magnitude and variation of entropy and represent three statistical structures respectively. At next step, we analyze these features and simplify them based on the idea about sparseness.

C. Sparsity and Deep-Learning Network

Through the procedure abovementioned, we get three types of features, but there may exist correlation between them, that is to say, features can be simplified to a certain degree. Inspired by thought of sparseness, we expect to use less features to give a robust and simple expression, and SAE is adopted to derive a more sparse formation about features. Fig. 4 shows the SAE structure with two hidden layers.

It can see clearly that SAE is a deep neural network model with tightly stacking of AE (auto-encoder). The output of the lower AE is also the input of the upper AE. The principle of AE is explained in Section II.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

SAE gradually realizes the abstraction of features through stacking, eventually gets more compact and useful features. The training of SAE is divided into two stages, which include pre-training and fine-tuning. The first stage is pre-training from bottom to top. The output of the k-th hidden layer of the model will be used as input to train the $k+1$-th hidden layer, encoding matrix W and encoding bias b can be obtained by minimizing the cost function. These two parameters are calculated by gradient descent. For k-th layer, the renewal equations are as follows:

$$W_{ij}(k) = W_{ij}(k) - \varepsilon \frac{\partial}{\partial W_{ij}(k)} L(W, b) \tag{26}$$

$$b_i(k) = b_i(k) - \varepsilon \frac{\partial}{\partial b_i(k)} L(W, b) \tag{27}$$

where ε is learning rate, W_{ij} means weight which is connected with unit j in k-th layer and unit i in $k+1$-th layer, L indicates reconstruction error which is defined in Section II-C. Then parameters distributed in the network is tuned with supervised training after pre-training.

It should be noted that the L has two forms of expression. When the number of neural units in hidden layer is smaller than the input layer, the dimensions of features are reduced through hidden layers. The principle is similar to PCA (principal component analysis) in this case. And we can also set the number of hidden units more than the input layer units, then add sparse constraints to hidden layers. This scheme is more consistent with visual theory, because input signals only stimulate a small amount of neurons in the process of receiving visual signals, while most of the other neurons are inhibited.

For the above reason, we choose second strategy, sparse constraint is used in penalty item. It aims at controlling the number of 'activated' neurons in the hidden layer (for sigmoid function, 1 means the output of a neuron is 'activated', while 0 is considered to be 'suppressed'). Average amount of activation is used to indicate the activation level of j-th neuron in the hidden layer, it is defined as:

$$\hat{\rho}_j = \frac{1}{N} \sum_{i=1}^{N} z_j\left(y^{(i)}\right) \tag{28}$$

In order to restrain most of neurons, activation level and penalty term are added in objective function to make $\hat{\rho}_j$ closer to a constant ρ. KL divergence is adopted as a penalty item, it is calculated as:

$$KL\left(\rho || \hat{\rho}_j\right) = \rho log \frac{\rho}{\hat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_j} \tag{29}$$

The deviation degree of KL divergence will become larger with distinction between $\hat{\rho}_j$ and ρ. Especially, $KL(\rho||\hat{\rho}_j) = 0$ when $\hat{\rho}_j = \rho$. Furthermore, the objective function with sparsity constraint can be expressed as:

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta} L_{sparse}(Y, Z) \\ &= \arg\min_{\theta} L(Y, Z) + \gamma \sum_{j=1}^{H_N} KL\left(\rho || \hat{\rho}_j\right) \end{aligned} \tag{30}$$

where γ represents the weight of sparse penalty, H_N indicates the number of hidden layer units. And the corresponding renewal equation is changed as:

$$W_{ij}(k) = W_{ij}(k) - \varepsilon \frac{\partial}{\partial W_{ij}(k)} L_{sparse}(W, b) \tag{31}$$

$$b_i(k) = b_i(k) - \varepsilon \frac{\partial}{\partial b_i(k)} L_{sparse}(W, b) \tag{32}$$

With the sparse constraint, redundancy of features can be reduced to a small degree. And different degrees of sparse constraints will change the result, which is demonstrated in this experiments. Through the SAE, we can obtain the 'deeper' and 'sparse' features, giving a better representation on video's frames structure. In this paper, SAE is designed with two hidden layers, first hidden layer has 60 nodes and the second hidden layer has 50 nodes, input layer has 42 nodes and the output layer has 10 nodes. We have tried to construct SAE with more hidden layers, but the results have not be improved, so we adopt a simpler way. Besides that, SVM is connected with output layer to regress features and subjective scores, finally get the performance metrics.

Fig. 5. The random frame of ten pristine videos in NAMA3DS1-COSPAD1 stereo videos database.

IV. Experiments Results and Analysis

In this section, we compare our metric with other algorithms which have the remarkable effect on two public databases. We validate effectiveness of our proposed method through better performance. Meanwhile, the rationality and essentiality of saliency and sparseness are proved in the manner of contrast experiment.

A. Stereoscopic Video Databases

1) NAMA3DS1-COSPAD1 Stereo Video Database: We carry through contrastive experiments using NAMA3DS1-COSPAD1 stereo video database described in [51]. It is a general database for video quality evaluation. There are 10 original videos and 100 symmetrically distortions videos, 10 kinds of distortion in each video source. The types of distortions include JPEG2000, sharpen, H.264/AVC, reduction of resolution (at different levels of bit rates) and downsample with sharpen. The frames' size of NAMA3DS1-COSPAD1 stereo video database is 1920×1080 and the frame rate is 25 fps. The illustration of NAMA3DS1-COSPAD1 stereo videos database are shown in Fig. 5.

2) QI-SVQA Video Database: To verify the extensive suitability of proposed method, we also test on the another public videos database (we named QI-SVQA video database in this

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 6. The first frame of nine pristine videos in QI-SVQA database.

TABLE I
THE MAIN DIFFERENCE IN NAMA3DS1-COSPAD1 STEREO VIDEO DATABASE AND QI-SVQA DATABASE

video databases	distortion type	characteristic of distortion	number of source videos	total number of videos
NAMA3DS1-COSPAD1	H.264 JPEG2K reduction of resolution, sharpening, downsampling and sharpening	symmetrical	10	100
QI-SVQA	H.264,blur	asymmetric	9	450

paper) in [19]. QI-SVQA video database has 9 pristine source videos and distortions, which be divided into 225 blur videos and 225 H.264 videos. Unlike the NAMA3DS1-COSPAD1 stereo videos database, its distortion is asymmetrical, so it is more challenging in this respect. Fig. 6 shows the illustration of QI-SVQA database.

3) The Difference Between Two Stereo Video Databases: In order to clearly show the difference between the two video databases, we tabulate the key points. The main terms are shown in Table I.

From the table, we can see that there are more samples in QI-SVQA database and more distortion types in NAMA3DS1-COSPAD1 stereo video database. Besides that, the distortion in QI-SVQA database is asymmetrical when the distortion in NAMA3DS1-COSPAD1 stereo video database is symmetrical. Our method is tested on two databases, final results are shown in later subchapter.

B. Performance Metrics

Just like IQA, four indicators which include Pearson linear correlation coefficient (PLCC), Spearman rank correlation coefficient (SROCC) [52], Kendall rank-order correlation coefficient (KROCC) and Root mean squared error (RMSE) are used in VQA. The range of them is from 0 to 1, and 1 indicates the best performance for PLCC, SROCC and KROCC while 0 is the perfect result for RMSE. The four indexes measure the fitting degree with human mean opinion scores (MOS), which

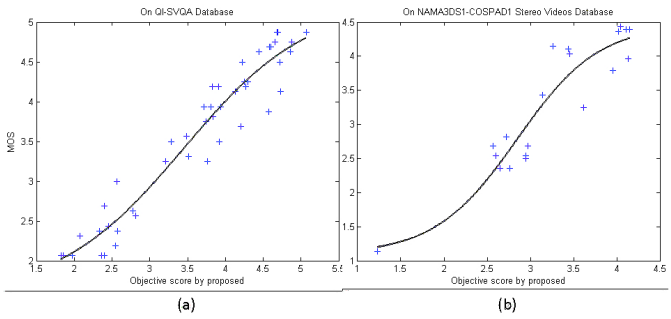


Fig. 7. The scatter plot between the MOS and objective quality scores, (a) shows the results on QI-SVQA database, when (b) shows NAMA3DS1-COSPAD1 stereo video database's.

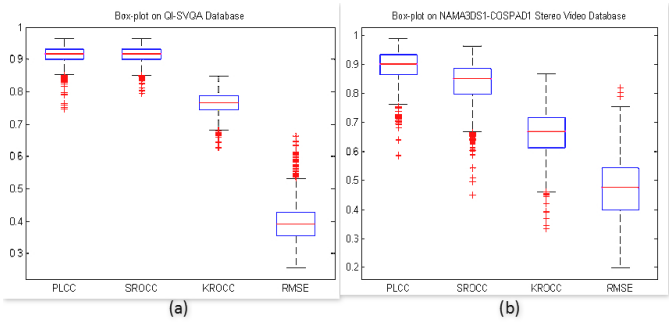


Fig. 8. The box plot between the MOS and objective quality scores, (a) shows the results on QI-SVQA database, when (b) shows NAMA3DS1-COSPAD1 stereo video database's.

is used to represent the videos' perceived quality. Its range is from 1 to 5. For the videos from same source, the higher the evaluation score, the better the subjective quality. The number of times of random subdivision is set to 1200, and the median value is taken as the final result in the experiment.

C. Test and Comparison on Video Databases

We test our method on two different video databases to verify the effectiveness and expansibility, in addition, compare with other objective video quality metrics, include 2D IQA metrics (SSIM [6], PSNR) and some excellent 3D VQA algorithms. With regard to 2D algorithms, apply them on two views and each frame, then take the mean value as the approximate value of VQA. In the experimental part of the paper, we randomly select 80% videos for training and 20% for test, and there is no overlap between the two sets [53]. The run time is set to 1200, and we take median value as the final results. The scatter plot between the MOS and objective quality scores (utilize proposed method) on two databases are shown in Fig. 7, the box plot between the MOS and objective quality scores (utilize proposed method) on two databases are shown in Fig. 8, and the four evaluation indexes on various methods are tabulated in Table II and Table III. Boldface represents the best results among these metrics.

It can be seen intuitively that the prediction score is fitting well with MOS from Fig. 7 and Fig. 8 on two databases, more accurate results are listed in the following two tables.

As shown in Table II and Table III, the proposed metric performs better than other methods. It is noticeable that there are

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
FOUR INDEXES OF DIFFERENT METHODS ON QI-SVQA DATABASE

Metrics	PLCC	SROCC	KROCC	RMSE
SSIM	0.8185	0.8281	0.6418	0.5580
PSNR	0.8496	0.8637	0.6832	0.5122
PQM	0.7852	0.8165	0.6365	0.6158
SFD	0.6483	0.6633	0.5021	0.7571
PHVS-3D	0.7082	0.7195	0.5353	0.7021
3D-STS	0.8311	0.8338	0.6553	0.5520
Feng	0.8415	0.8379	0.6650	0.5372
Proposed	**0.9141**	**0.9111**	**0.7605**	**0.4018**

TABLE III
FOUR INDEXES OF DIFFERENT METHODS ON NAMA3DS1-COSPAD1 STEREO VIDEOS DATABASE

Metrics	PLCC	SROCC	KROCC	RMSE
SSIM	0.7664	0.7492	0.5444	0.7296
PSNR	0.6699	0.6470	0.4800	0.8433
PQM [54]	0.6340	0.6006	0.4391	0.8784
SFD [55]	0.5965	0.5896	0.4025	0.9117
PHVS-3D [56]	0.5480	0.5146	0.3572	0.9501
3D-STS [10]	0.6417	0.6214	0.4544	0.9067
Feng [19]	0.6503	0.6229	0.4575	0.8629
Proposed	**0.9016**	**0.8467**	**0.6690**	**0.4679**

TABLE IV
DIFFERENT TYPES OF DISTORTION ON TWO DATABASES

distortion type	number	PLCC	SROCC	KROCC	RMSE
Q(blur)	225	0.9336	0.9237	0.7847	0.3458
Q(H.264)	225	0.9254	0.9235	0.7863	0.4018
N(JPEG2K)	30	0.9175	0.7914	0.6202	0.4664
N(H.264)	40	0.8926	0.7857	0.6429	0.5017
N(Other 3 types)	30	0.8518	0.6857	0.6113	0.4215

almost FR methods for comparison, but our proposed method is a NR metric, under this circumstance, its performance is still superior to the others, which proves its effectiveness. In addition, comparing with other methods that divide the information into spatial and temporal, the framework is simpler in proposed metric, which demonstrates proposed structure is reasonable and efficient.

In addition, in order to validate the extensibility of our metric, we show the results about two types of distortions on QI-SVQA database (abbreviated as Q) and NAMA3DS1-COSPAD1 stereo video database (abbreviated as N). And 'number' means the quantity of each type of distortion. The results are shown in Table IV.

In Table IV, N (Other 3 types) included reduction of resolution,sharpening, downsampling and sharpening. We put them together due to each type only has 10 samples, which will make the result unstable. From the table we find the performance is good on distortion type of two databases, confirming our proposed metric works well for different types of distortion.

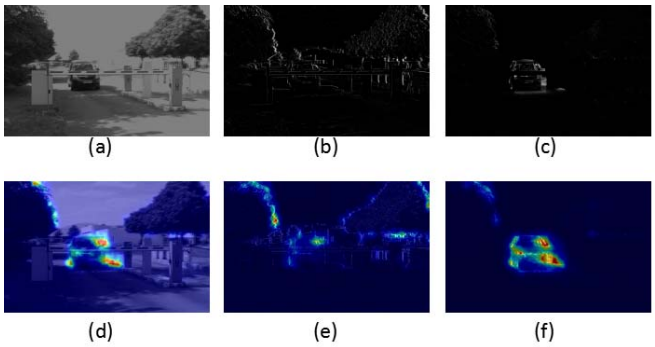


Fig. 9. The sum map,difference map,frame difference map of one frame are shown in (a), (b), (c) respectively, and their corresponding 3D saliency map are (d), (e), (f) respectively.

TABLE V
THE COMPARISON WITH PERFORMANCE AMONG VARIOUS MAPS

Types of maps	PLCC	SROCC	KROCC	RMSE
(1)diff	0.6738	0.6377	0.4689	0.7795
(2)diff+saliency	0.8406	0.7919	0.6050	0.5849
(3)fradiff	0.8055	0.7731	0.5857	0.6546
(4)fradiff+saliency	0.5001	0.4677	0.3545	0.9285
(5)mono-view	0.7915	0.7496	0.5587	0.6746
(6)mono-view+saliency	0.8754	0.8255	0.6351	5408
(7)sum	0.8578	0.8249	0.6579	0.5449
(8)sum+saliency	**0.9016**	**0.8467**	**0.6690**	**0.4679**

D. Compare Performance With Different Maps

In this subchapter, we discuss about the effect of sparsity features with different maps. Then we test 6 types of maps: difference map, which only includes the main spatial information; sum map, which contains almost all information of videos; frame difference map, which retains the variation at spatial and temporal domain. The other three types are saliency maps corresponding to the maps abovementioned. We compare them for the sake of exploring the effect of spatial, temporal, saliency and background. The corresponding six maps of one frame are shown in Fig. 9. Besides that, we also calculate the results under monocular vision, namely take average value between the left and right view. The results of experiment are shown in Table V.

From Table V, we can see that the performance of eighth group is best, which is also our choice. It is because the sum map's 3D saliency map contains almost all the information about frames, whether temporal, spatial domain, saliency or non-saliency objects. It has complete content that none of other maps have. By comparing with results on (1) and (2), (5) and (6), (7) and (8), it is easy to find that adding saliency can improve the evaluation results, and it is comprehensible because the positions with saliency are more attractive. But unexpectedly, the final result decreases severely comparing with (3) and (4). We analyze it is due to the irrational expansion of saliency. Differ from the sum maps and difference maps, the position on saliency may be varying large relatively, result in saliency areas expanding unreasonably. In addition, due to the subtraction operation between two frames,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

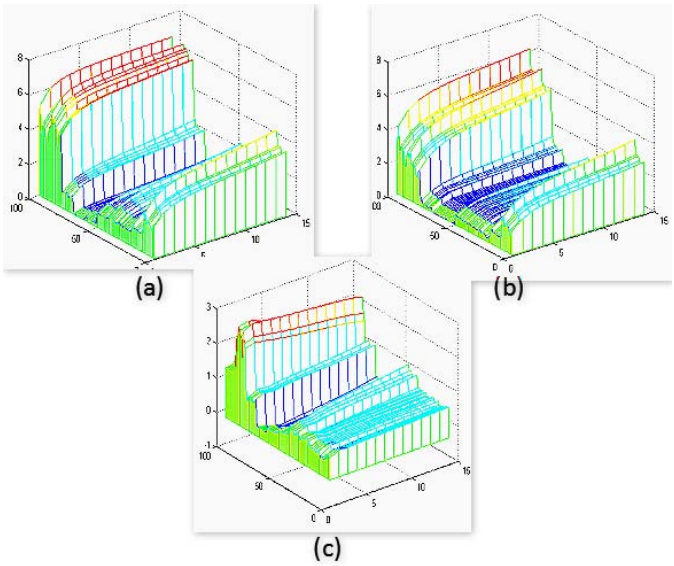


Fig. 10. The distribution of entropy on different maps, (a) corresponds to I_S when (b) corresponds to I_{sum_S}, and (c) is difference between the (a) and (b).

the background and other structure message will be reduced by a large margin, from another perspective, the saliency covers the primitive images' features. It is more intuitive from Fig. 9.

On the other side, comparing (1) with (3), we find that frames' difference maps can express the video's quality better. It is because there exists temporal features in them. In contrast with (5) and (7), (6) and (8), it is easily observed that sum map is beneficial to binocular perception. Through the group of (7), we can see that sum maps are preferable to difference maps between two view of frames and inter-frames. It proves the features on background also play a important role for sparse representation, sufficient images' information improve the effect of sparse representation.

In summary, we draw a conclusion that lacking one of spatial, temporal, saliency or background will degrade performance on sparseness, confirming that our work is reasonable.

E. The Effect of Saliency on Coefficients

We conjecture that the saliency will have an impact on coefficients in the previous chapter, but it has not been verified. In order to valid this assumption, we calculate the value of coefficients' entropy on videos. More concretely, after obtaining the sparse coefficients on per sample frame on videos, we take average value of these frames, regarding as the whole video's entropy. This method is applied on NAMA3DS1-COSPAD1 stereo videos database, so there are 100 videos, and the L (represents the number of entropy per frame) is 14 in the experiment. We have tested on the sum map (denote as I_S) and the saliency map corresponding to it (denote as I_{sum_S}). The intuitive results are shown in Fig. 10.

From three figures we can clearly see that the amplitude of videos' entropy increased after applying the saliency on sum maps, and the regularity of the change is universal, though

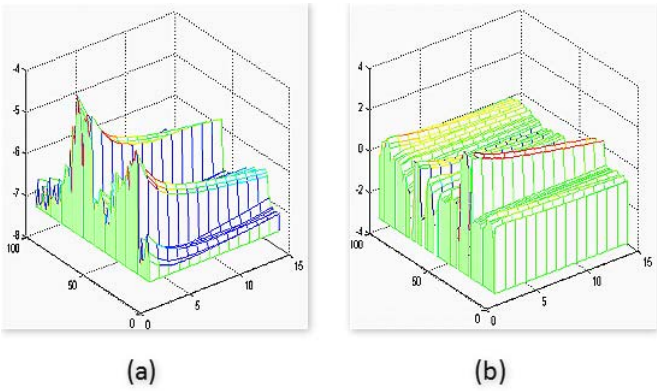


Fig. 11. The distribution of entropy on different maps, (a) corresponds to the difference between the I_{dif} and its saliency map when (b) corresponds to the difference between the I_{fradif} and its saliency map.

the degrees of variations are not identical. It can demonstrate iconically the influence of saliency on coefficients through the sparse representation, which conforms to the characteristic of HVS. From the view of sparse coding, the parts of saliency change the pixels value in image blocks; from the perspective of visual perception, optical signal from salient parts will be handled differently by cells on retinas, so the sum map's model that considers the saliency has better performance, which explains the results on above subchapters.

Beyond that, the difference map (denote as I_{dif}) and frames' difference map (denote as I_{fradif}) are processed in the same manner, which is used to compare with the previous experiment. Intuitive results can be seen from figures in Fig. 11.

Surprisingly, from Fig. 11, the results on two kinds of difference map are nearly opposite to sum map's. Adding the saliency to I_{dif} will lead to a decline of amplitude on entropy, but the variations are consistent on all coefficients. As for I_{fradif}, the change on salient is uncertain. This might be the principal reason for the decrease of metrics. It is obviously there is somewhat different on frame-difference map, its amplitudes have both positive and negative when the amplitudes of other two maps both are positive or negative. Therefore, it is reasonable to think such distinction leads to different results.

We further analyze the reasons. The reason for this result may be different characteristics between maps. Sum map and difference map are generated from left and right view. When people observe a sense, the distinction between two views is not large, so the salient region is similar; but for frame-difference map, it also includes the temporal variation of salient regions, which makes two salient regions (saliency of previous frame and current frame) with larger difference appear on one map, it is apparently unreasonable. Besides that, salient regions are usually brighter (namely pixels' value is higher), so unreasonable salient region also masks the original information. These factors result in different variations and make bad performance on saliency for frame-difference maps.

Through a series of tests, we draw the conclusion that the consistency of variation about coefficients' entropy plays an important role on VQA, which is affected by saliency. On the other side, the fluctuation on amplitude of entropy can not account for the performance is good or bad.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VI
PERFORMANCE METRICS UNDER DIFFERENT COMPONENTS OF FEATURES

features	PLCC	SROCC	KROCC	RMSE
(1)f1	0.7358	0.6933	0.5142	0.7329
(2)f2	0.7369	0.6970	0.5264	0.7311
(3)f3	0.7483	0.7078	0.5230	0.7149
(4)f1+f2	0.8713	0.8245	0.6311	0.4976
(5)f1+f3	0.7780	0.7281	0.5407	0.6955
(6)f2+f3	0.8812	0.8257	0.6442	0.4810
(7)f1+f2+f3	0.8893	0.8309	0.6518	0.4731

F. Redundancy Between Features

As mentioned earlier, features extracted from videos are made up of three components: mean, standard deviation and 2 norm. For the sake of understanding the contribution of each part, we perform experiments on features of mean, standard deviation and 2 norm component respectively. In addition, any two types of features are combined for testing (note that there is no sparse constraint here). Denote mean component as $f1$, standard deviation component as $f2$, 2 norm component as $f3$. The performance metrics are listed in Table VI.

From Table VI, something can be observed. Through (1), (2), (3), we can see the results on a single type of features are relativity low, and the performance metrics are close to each other. Besides that, when different types of features are combined, the effect may be improved, but the levels of improvement are different. For instance, (4) and (6) show the improved effect is significant, while (5) is lower. Compare (7) with (4) and (6), the improvement is also a little. This may be able to analyze in terms of mathematics. The correlation between mean and 2 norm is higher, which leads combination (5) to have more redundant information. By contrast, standard deviation has low correlation with mean and 2 norm, so the features can complement each other.

According to the analysis above, we can find there is redundant information between the three types of features. Meanwhile, sparsity can reduce redundancy and further improve the results, so a variable which represents sparse constraint is introduced in the next experiment.

G. Sparsity and SAE

In this section, we mainly explore the effect of sparsity on improving feature performance. As we mentioned in the previous chapter, it is more consistent with the visual perception theory that introducing sparse constraint in the structure of SAE. In Section IV-F, it is proved that there exist redundancy between features, and sparsity is beneficial to reduce the redundancy. so we conduct experiment on sparse constraint. Parameter S is introduced as a variable. It represents the target of sparse degree, namely expectation on the sum of weights in each layer. The degree of sparsity is higher when the value of S is smaller. We set different values for S and results are shown in Fig. 12.

From Fig. 12, the general changing tendency can be analyzed. We can clearly see that within a certain range, the performance increases gradually with the decrease of S, which means the increase of sparse degree will lead to the performance improvement of SAE. But when the value of S is too small, performance will decline instead. This can be explained by the characteristics of sparsity. On one hand, when S is large, namely the expectation of sparsity is small, sparse constraints is not fully reflected, which causes the decline of ability about learning features. On the other hand, when the value of S is too small, only a very small number of hidden layer units are working. In this case, the hidden layers lack in ability of expressing features, leading to decline of network performance. It follows that the influence of sparsity on performance is not monotonically increasing, only suitable value can achieve the best effect. In this paper, S is set to 0.3. Besides that, comparing the red points (with suitable sparse constraint) with red horizontal dashed lines (without sparse constraint), we find that an appropriate sparse degree can further improve performance, which proves the effect of sparsity.

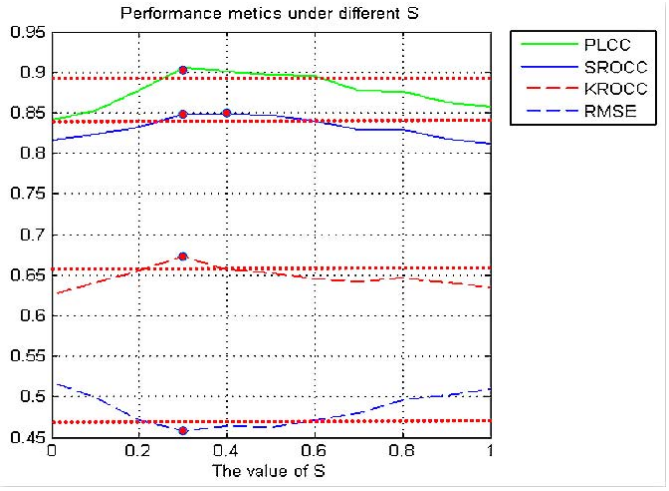


Fig. 12. Performance metrics under different S. The red points indicate the appropriate value point, four red horizontal dashed lines corresponding to the four performance metrics without sparse constraint.

It should be noted that S only indicates an expected target, instead of the sum of network's weights on each layer. For example, when S is valued as 0, obviously the sum of weights on each layer won't be 0, in this case it means that all nodes in the hidden layer are not be activated, which is impossible.

H. Number of Hidden Layers on SAE

In the previous description, we constructed a SAE with two hidden layers, but this structure is shallower compared with general deep learning network. So the networks which have more hidden layers are tested in the experiments. We have chosen several numbers, which are ranged from 1 to 10. The performance metrics are detailed in Table VII.

From the table, we find that there is not much improvement for SAE which has more hidden layers, which means that 2 hidden layers are enough. It is probably because the number of samples input is small, so shallower network is able to deal with the problem, while too many hidden layers will degrade

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
PERFORMANCE METRICS ON SAE WITH DIFFERENT HIDDEN LAYERS

Network	PLCC	SROCC	KROCC	RMSE
SAE(1 layer)	0.8776	0.8217	0.6319	0.4793
SAE(2 layers)	0.9016	0.8467	0.6690	0.4679
SAE(3 layers)	0.8939	0.8401	0.6546	0.4640
SAE(5 layers)	0.8373	0.7551	0.6175	0.5476
SAE(10 layers)	0.7802	0.6836	0.5460	0.6239

the network performance. Moreover, adding hidden layers will increase the number of parameters, making it more difficult to adjust parameters. All things considered, 2 hidden layers are fitting to this work.

V. CONCLUSION

In this paper, we proposed NR-SVQA method based on the saliency and sparsity. The entire procedure embodies the ideas on saliency and sparseness. We utilize the sum maps of 3D saliency maps to integrate the information on spatial, temporal, saliency and background, instead of separating the videos into spatial and temporal domain; in addition, verifying the validity of this procedure in the experiment subsequently. Then we extract features by means of combining saliency maps and sparse representation. Afterwards, handling the features with SAE in the light of idea on sparse. Sparse coding lessen the superfluous structure, decomposing per frame into coefficients, analogy with the process about receiving light signals on human eyes' cells; meanwhile, SAE reduces the redundancy and makes features sparse. Experimental results indicate that the sparse is meaningful and feasible for VQA. The overall framework of our model is simple, but the performance is well.

There are other questions worth exploring in the future work. For instance, we only employ the sum maps to model the binocular vision, but it is not enough for stereoscopic perception. And we don't further highlight the role of saliency in sparse coding, there is still room for improvement. We are going to study in these directions.

REFERENCES

[1] Q. Xu *et al.*, "Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence," *Inf. Sci.*, vol. 278, pp. 736–756, Sep. 2014.

[2] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[3] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Process. Image Commun.*, vol. 28, no. 8, pp. 870–883, 2013.

[4] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synth. Lectures Image Video Multimedia Process.*, vol. 2, no. 1, pp. 1–156, 2006.

[5] J. Wu, W. Lin, G. Shi, L. Li, and Y. Fang, "Orientation selectivity based visual pattern for reduced-reference image quality assessment," *Inf. Sci.*, vol. 351, pp. 18–29, Jul. 2016.

[6] Z. Wang and A. C. Bovik, "Image and multidimensional signal processing-a universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[7] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[9] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[10] J. Han, T. Jiang, and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, 2012, pp. 1–6.

[11] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 591–602, Apr. 2014.

[12] F. Shao, W. Chen, G. Jiang, and Y.-S. Ho, "Modeling the perceptual quality of stereoscopic images in the primary visual cortex," *IEEE Access*, vol. 5, pp. 15706–15716, 2017.

[13] M. Yu, K. Zheng, G. Jiang, F. Shao, and Z. Peng, "Binocular perception based reduced-reference stereo video quality assessment method," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 246–255, Jul. 2016.

[14] Y. Liu *et al.*, "Stereoscopic image quality assessment method based on binocular combination saliency model," *Signal Process.*, vol. 125, pp. 237–248, Aug. 2016.

[15] J. Yang *et al.*, "Quality assessment metric of stereo images considering cyclopean integration and visual saliency," *Inf. Sci.*, vol. 373, pp. 251–268, Dec. 2016.

[16] F. Shao, W. Chen, W. Lin, Q. Jiang, and G. Jiang, "Simulating receptive fields of human visual cortex for 3D image quality prediction," *Appl. Opt.*, vol. 55, no. 21, pp. 5488–5496, 2016.

[17] W. Zhao, L. Ye, J. Wang, and Q. Zhang, "No-reference objective stereo video quality assessment based on visual attention and edge difference," in *Proc. IEEE Adv. Inf. Technol. Electron. Autom. Control Conf. (IAEAC)*, 2015, pp. 523–526.

[18] C. Oprea, I. Pirnog, C. Paleologu, and M. Udrea, "Perceptual video quality assessment based on salient region detection," in *Proc. 5th Adv. Int. Conf. Telecommun. (AICT)*, 2009, pp. 232–236.

[19] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal Image Video Process.*, vol. 10, no. 4, pp. 737–744, 2016.

[20] M. Shahid, A. Rossholm, and B. Lövström, "A no-reference machine learning based video quality predictor," in *Proc. 5th Int. Workshop Qual. Multimedia Exp. (QoMEX)*, 2013, pp. 176–181.

[21] D. C. Mocanu *et al.*, "No-reference video quality measurement: Added value of machine learning," *J. Electron. Imag.*, vol. 24, no. 6, 2015, Art. no. 061208.

[22] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 347–364, Apr. 2012.

[23] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[24] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[26] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "Sun: Top-down saliency using natural statistics," *Vis. Cogn.*, vol. 17, nos. 6–7, pp. 979–1003, Aug. 2009.

[27] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, Sep. 2007.

[28] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[29] X. Yang, W. Zhao, L. Ye, and Q. Zhang, "A novel no-reference objective stereoscopic video quality assessment method based on visual saliency analysis," in *Proc. 9th Int. Conf. Digit. Image Process. (ICDIP)*, vol. 10420, 2017, Art. no. 104203S.

[30] C. Jia, W. Lu, L. He, and R. He, "Spatiotemporal saliency detection based video quality assessment," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2016, pp. 340–343.

[31] D. Culibrk *et al.*, "Salient motion features for video quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 948–958, Apr. 2011.

[32] Q. L. Wei and Y. Zhang, "Video objective quality evaluation system based on the visual saliency map," *Appl. Mech. Mater.*, vols. 411–414, pp. 1362–1367, Sep. 2013.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[33] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.

[34] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.

[35] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1146–1153.

[36] T. Guha, E. Nezhadarya, and R. K. Ward, "Sparse representation-based image quality assessment," *Signal Process. Image Commun.*, vol. 29, no. 10, pp. 1138–1148, 2014.

[37] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *J. Opt. Soc. America A Opt. Image Sci. Vis.*, vol. 15, no. 8, pp. 2036–2045, 1998.

[38] F. Qi, D. Zhao, and W. Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2338–2344, Dec. 2015.

[39] J. Zhang, S. Ma, R. Xiong, D. Zhao, and W. Gao, "Image primitive coding and visual quality assessment," in *Proc. PCM*, 2012, pp. 674–685.

[40] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2691–2698.

[41] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, "Learning sparse representation for no-reference quality assessment of multiply distorted stereoscopic images," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1821–1836, Aug. 2017.

[42] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, Oct. 2013.

[43] F. Shao, K. Li, G. Jiang, M. Yu, and C. Yu, "Monocular–binocular feature fidelity induced index for stereoscopic image quality assessment," *Appl. Opt.*, vol. 54, no. 33, pp. 9671–9680, 2015.

[44] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *Algorithmic Learning Theory*. Heidelberg, Germany: Springer, 2011, pp. 18–36.

[45] J. Zabalza *et al.*, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, Apr. 2016.

[46] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full–reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1940–1953, May 2013.

[47] S. Grossberg and F. Kelly, "Neural dynamics of binocular brightness perception," *Vis. Res.*, vol. 39, no. 22, pp. 3796–3816, 1999.

[48] S. Henriksen and J. C. A. Read, "Visual perception: A novel difference channel in binocular vision," *Current Biol.*, vol. 26, no. 12, pp. R500–R503, Jun. 2016.

[49] G. R. Engel, "The visual processes underlying binocular brightness summation," *Vis. Res.*, vol. 7, nos. 9–10, pp. 753–767, 1967.

[50] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[51] M. Urvoy *et al.*, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Proc. 4th Int. Workshop Qual. Multimedia Exp. (QoMEX)*, Jul. 2012, pp. 109–114.

[52] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity Oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.

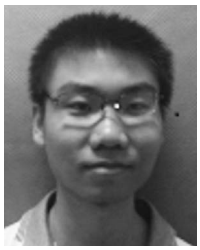
[53] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[54] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando, and A. M. Kondoz, "Perceptual video quality metric for 3D video quality assessment," in *Proc. 3DTV Conf. True Vis. Capture Transm. Display 3D Video (3DTV CON)*, Jun. 2010, pp. 1–4.

[55] F. Lu, H. Wang, X. Ji, and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model," in *Proc. 3DTV Conf. True Vis. Capture Transm. Display 3D Video*, May 2009, pp. 1–4.

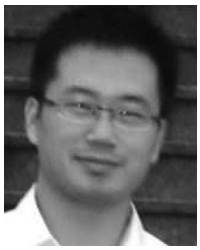
[56] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 2521–2524.

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, where he is currently a Professor. 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.



Chunqi Ji received the B.S. degree in electrical and information engineering from the Hebei University of Technology, Tianjin, China, in 2016. He is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin. His research interests include multimedia quality evaluation, stereo vision research, and pattern recognition.

Bin Jiang received the B.S. and M.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. His research interests include multimedia quality evaluation, stereo vision research, and pattern recognition.



Wen Lu received the M.S. and Ph.D. degrees in electrical engineering from Xidian University, China, in 2006 and 2009, respectively, where he is currently an Associate Professor. 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.