No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images

Yan, Peng, Mou, Xuanqin

Peng Yan, Xuanqin Mou, "No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images," Proc. SPIE 10817, Optoelectronic Imaging and Multimedia Technology V, 108171D (8 November 2018); doi: 10.1117/12.2503149

SPIE.

Event: SPIE/COS Photonics Asia, 2018, Beijing, China

No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images

Peng Yan and Xuanqin Mou
Institute of Image Processing and Pattern Recognition, Xi'an Jiaotong University, CHINA

ABSTRACT

As video applications become more popular, no-reference video quality assessment (NR-VQA) has become a focus of research. In many existing NR-VQA methods, perceptual feature extraction is often the key to success. Therefore, we design methods to extract the perceptual features that contain a wider range of spatiotemporal information from multi-directional video spatiotemporal slices (STS) images (the images generated by cutting video data parallel to temporal dimension in multiple directions) and use support vector machine (SVM) to perform a successful NR video quality evaluation in this paper. In the proposed NR-VQA design, we first extracted the multi-directional video STS images to obtain as much as possible the overall video motion representation. Secondly, the perceptual features of multi-directional video STS images such as the moments of feature maps, joint distribution features from the gradient magnitude and filtering response of Laplacian of Gaussian, and motion energy characteristics were extracted to characterize the motion statistics of videos. Finally, the extracted perceptual features were fed in SVM or multilayer perceptron (MLP) to perform training and testing. And the experimental results show that the proposed method has achieved the state-of-the-art quality prediction performance on the largest existing annotated video database.

Keywords: video quality assessment, multi-directional video spatiotemporal slices images, no-reference, support vector machine

1. INTRODUCTION

Video quality assessment (VQA) has been a long-standing problem for various video related applications such as video compressing, transmission, and display. Especially now, with the easy availability of video generating and viewing terminals, videos are flooding the world around us, and quality assessment issues are becoming more urgent and important. In general, video quality evaluation is divided into subjective and objective quality evaluation according to whether there is any subjective experimenter's participation. Subjective evaluations that require people to participate in are more time consuming, laborious, and costly than objective algorithms that do not require human intervention. Therefore, researchers have been trying to develop objective VQA algorithms that can replace human subjective evaluation.

Typically, objective VQA algorithms can be classified into full-reference (FR), reduced-reference(RR), and no-reference (NR) according to the degree of availability of the reference information. FR VQA algorithms, which can obtain all the reference information, are a kind of methods that are widely studied and have many research results. RR VQA algorithms that can obtain only part of the feature information of reference videos are used in some bandwidth limited situations. NR VQA algorithms as methods that do not require any reference information will be more widely used in the future. This paper proposes an NR VQA algorithm.

Among FR VQA algorithms, the most typical and widely used are peak signal noise ratio (PSNR) and mean squared error (MSE). Both of them are popular because of their high efficiency and ease of use, however, their inconsistency with subjective perception is often criticized. Therefore, researchers have tried to study algorithms that are consistent with subjective perception and proposed various objective algorithms, such as the FR VQA algorithms Content-weighted VQA (3-SSIM)[1], temporal quality variations (TQV)[2], digital video quality metric(DVQ)[3], perceptual quality index (PQI)[4], temporal trajectory aware video quality (Tetra VQM)[5], video quality metric (VQM)[6], Video structure similarity measure (VSSIM)[7], spatiotemporal most-apparent-distortion measure (ST-MAD)[8], MOTION-based video integrity assessment (MOVIE)[9], VQA analysis via spatial and spatiotemporal slices (ViS3)[10], VQA via gradient magnitude similarity of spatial and spatiotemporal slices (STS-GMSD/SSTS-GMSD)[11], motion structure partition

Optoelectronic Imaging and Multimedia Technology V, edited by Qionghai Dai, Tsutomu Shimura, Proc. of SPIE Vol. 10817, 108171D · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2503149

similarity (MSPS)[12], and the NR algorithm blind natural video quality measure (V-BLIINDS)[13], etc. Some of which will be briefly reviewed below.

MOVIE[9] is a perceptually driven algorithm that uses a three-dimensional Gabor filter bank to filter video in the spatiotemporal domain and then calculates spatial and temporal differences between filtering outputs. Finally, the spatial and temporal components were combined into an overall quality index. ST-MAD[8] and its improved version ViS3[10] extends image quality assessment (IQA) algorithm named most-apparent-distortion (MAD[14]) to take into account visual perception of motion artifacts, and then estimates video quality by detecting the spatial distortion of video frames and spatiotemporal distortion of spatiotemporal slices (STS) images[15]. ViS3 firstly applies adapted MAD to groups of video frames to capture spatial distortion, and then quantifies the spatiotemporal similarity by measuring spatiotemporal correlation and applying a human vision system (HVS) based model to STS images. At last, it combines the spatial and spatiotemporal distortion. SSTS-GMSD and STS-GMSD, which are also VQA algorithms based on video spatiotemporal structural similarity, have achieved the high precision on the LIVE VQA database than ViS3. They evaluate video quality via detecting the structural similarity of STS images by using GMSD. Furthermore, as moving objects in the video sequence were found to have different influences on the prediction performance in terms of moving speed and track, STS images were divided into simple and complex motion regions, and MSPS[12] was proposed to detect the similarity between motion-partitioning STS images and achieve further better performance than SSTS-GMSD/ STS-GMSD on LIVE VQA database.

Now for NR quality algorithms, there are many successful NR IQA algorithms such as blind or referenceless image spatial quality evaluator (BRISQUE)[16], distortion identification-based image verity and integrity evaluation (DIIVINE)[17], Blind image integrity notator using DCT statistics (BLIINDS)[18], quality-aware clustering (QAC)[19], and deep CNN-based NR IQA[20]–[23], however, there are relatively few NR VQA algorithms. V-BLIINDS[13] and spatiotemporal feature combine model (STFC)[24] are two representatives of NR VQA algorithm.

V-BLIINDS utilizes a spatiotemporal natural scene statistics (NSS) model for videos and a motion model that quantifies motion coherency in video scenes to design a blind VQA algorithm that correlates highly with human judgments of quality. STFC combines spatial features (such as contrast and colorfulness) and temporal features (for example, sharpness and exposure time) extracted from vertical STS images to train an SVM model.

Among the abovementioned FR algorithms, the STS-based algorithms have achieved very good perceptual evaluation results, indicating that the perceptual features extracted in the STS can be more effectively used for FR video quality evaluation. In fact, it is worth noting that, whether it is an FR or NR algorithm, the extraction of perceptual features is important, and its effectiveness largely determines the success or failure of the evaluation algorithm.

Considering that effective FR features can be extracted from video STS images, this paper attempts to extract more perceptual features from video STS images for NR video quality evaluation. At present, quality evaluation based on deep neural networks has become a research hotspot. In IQA, deep neural networks, especially deep convolutional networks (CNN), are getting in full swing, such as the abovementioned IQA algorithms[20]–[23]. However, due to the small subjective video dataset and a large amount of video data, there are fewer VQA algorithms based on deep networks. Researchers are looking for ways to address the challenges of VQA based on deep learning. Especially, Vlad et al. have now published a large, annotated VQA database, which contains more than 1000 videos with real distortion types, named KoNViD-1k[25]. Although the number of videos in this database is much smaller than the amount of data in the large image training set based on deep neural network, such as ImageNet (about 14 million), it still takes a good step forward to solve the problem of the lack of labeled video data.

Inspired by the STS-based FR VQA algorithms, we start to consider extracting features in STS images and training through SVM. And then we perform algorithm performance verification on the existing largest video quality evaluation database with real distortion type, which proves the superiority of the proposed algorithm.

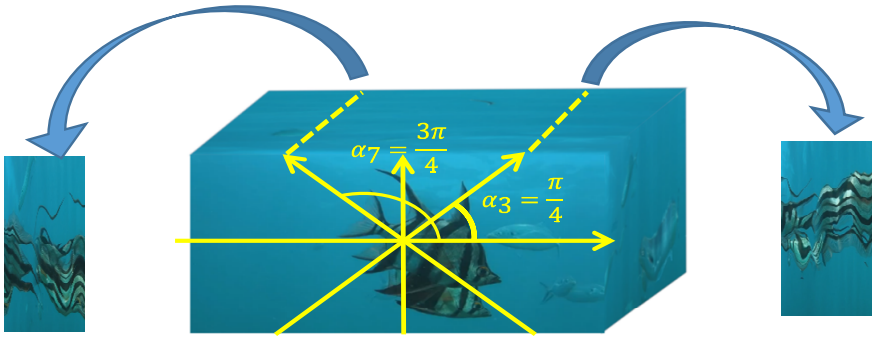
The remainder of the paper is organized as follows: Section 2 describes the methods, Section 3 presents the experiment, results, and discussion, and Section 4 gives the conclusion.

2. METHODS

Humans can extract efficient, accurate, and high-level semantic perceptions from images and videos with complex and sophisticated HVS. In image or video quality assessment, researchers try to simulate the perception process of HVS, such as contrast normalization, visual masking, motion silencing, and so on. After various visual processing, the

perceptual features of images or videos are obtained. Perceptual feature extraction is undoubtedly a key step in the process of quality evaluation. And good perceptual features will result in good evaluation performance.

For the spatial feature extraction of videos, classic IQA algorithms such as GMSD[26], VIF[27], MAD[14], SSIM[28], FSIM[29], NLOG[30], VSI[31], PAMSE[32], NSER[33], NLOG[30] etc. have performed well. For the temporal feature extraction and computing, most quality evaluation algorithms often use simple average pooling of image quality, frame difference features, and features come from a small group of frames [1-3],[6]. Compared with the previous temporal feature extraction methods, STS images provide the convenience of extracting perceptual features from a larger time scale, which will facilitate the design of effective quality evaluation algorithms, and the success of FR MSPS[12]and ViS3[10] have illustrated this point. Inspired by the success of STS-based FR-VQA in extracting features and evaluating video quality, this paper designs a method to obtain large temporal scale track features of video motion as much as possible by using multi-directional video STS images. Multi-directional video STS images, that is, the section of multi-directional cuts on video data parallel to the time axis. Multi-directional video STS images are shown in Figure 1.



(a) (b) (c)

Figure 1. Multi-directional video STS images. (a) shows a STS image extracted from $\frac{3\pi}{4}$ rad,(b) A sample video from KoNViD-1k Database, (c) shows the STS image extracted from $\frac{\pi}{4}$ rad.

We extracted the STS images in 8 directions according to Equation (1) and arranged them into a whole slice as shown in Figure 2. Note that only the pixel values of the positions cut by the multi-directional section are extracted into the multi-directional video STS images. After extracting multi-directional video STS images, we perform a perceptual feature extraction design.

$$\alpha_i = \begin{cases} 0, & i = 1 \\ \arctan(0.5), & i = 2 \\ \pi/4, & i = 3 \\ \arctan(2), & i = 4 \\ \pi/2, & i = 5 \\ \pi/2 + \arctan(0.5), & i = 6 \\ 3\pi/4, & i = 7 \\ \pi/2 + \arctan(2), & i = 8 \end{cases} \quad (1)$$

Figure 2. Splicing of multi-directional video STS images.

2.1 The moment-related features

First, in statistics, moments are often used to characterize the characteristics of sample data. Similarly, we first directly extract moment-related features (the mean value, standard deviation, skewness, kurtosis) of the grayscale multi-direction video STS images as a feature representation of the slice itself. In addition, considering the gradient is an important perceptual feature, we compute the spatial and temporal gradient of the multi-direction video STS images according to Equations (2) and (3), and then obtain the gradient map and the angle map of gradient of the multi-directional video STS images according to the Equation (4) and (5). On this basis, the mean value, standard deviation, skewness, kurtosis of gradient map and angle map are also extracted by Equations (6)-(9).

$$dx(i,j) = \frac{I(i,j+1) - I(i,j-1)}{2} \tag{2}$$

$$dt(i,j) = \frac{I(i+1,j) - I(i-1,j)}{2} \tag{3}$$

$$Gmap = \sqrt{dx^2 + dt^2} \tag{4}$$

$$GangleMap = arctan(\frac{dt + eps}{dx + eps}) \tag{5}$$

Note that I represents a multi-directional video STS image; $dx(i,j)$ and $dt(i,j)$ represent the spatial gradient and temporal gradient at position (i,j), respectively. *Gmap* denotes the gradient map of multi-directional video STS images, and *GangleMap* is the angle map of gradient of multi-directional video STS images. *eps* is a small constant to prevent dividing zeros

$$K1 = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

$$K2 = \sqrt{\sum_{i=1}^{n} \frac{(x_i - K1)^2}{n}} \tag{7}$$

$$K3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - K1)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - K1)^2\right)^{\frac{3}{2}}} \tag{8}$$

$$K4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - K1)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - K1)^2\right)^{2}} \tag{9}$$

Where x_i denotes a one-dimensional vectorized representation of the map, K1, K2, K3, and K4 represent the mean value, standard deviation, skewness, and kurtosis of a vector, respectively. N is the length of a vector.

2.2 Joint statistics of gradient magnitude and Laplacian of Gaussian features

Joint statistics of gradient magnitude and Laplacian of Gaussian features (GM-LOG) are very effective natural image statistical features, and it has shown a good evaluation correlation with the subjective evaluation in IQA. Gradient magnitude is first order statistic feature and LOG are second order statistic feature. GM-LOG is the joint statistics of the image gradient. The detailed calculation and specific application of GM-LOG in IQA can refer to Ref [34]. The motion track displayed in multi-direction video STS images are of the form of a gradient and show certain regularity. In view of its excellent natural image statistical representation ability, we try to extract the GM-LOG statistical features of multi-directional video STS images to extract the feature of motion track. In all, we extracted a total of forty GM-LOG features

for each video by using the methods from Ref [34]. By means of GM-LOG features, the motion characteristics of multi-direction video STS images are better statistically recorded

2.3 Spatiotemporal energy feature

Spatiotemporal energy[35] once appeared as a perceptual feature in the FR VQA named STME[36]. In STME, the spatiotemporal energy of the reference video and the distorted video is extracted, and then the spatiotemporal energy similarity is obtained as an effective VQA index. The spatiotemporal model used in STME are shown in Figure 3. Since there is no reference information in NR VQA, therefore, we directly characterize video motion features by extracting spatiotemporal energy features from multi-directional video STS images. Spatiotemporal energy characterizes the overall direction and velocity of motion in the video STS images. And the speed of motion often has a great relationship with the perception of video quality. For example, when the video motion is intense, we can't easily find the distortion in it. Therefore, here we extract spatiotemporal energy as a video quality-aware feature.

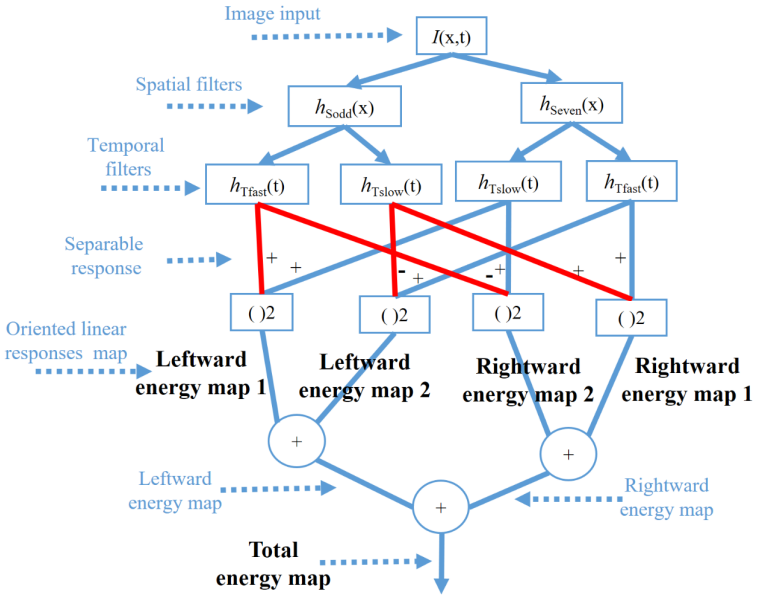


Figure 3. The spatiotemporal model used in STME.

For the extraction of spatiotemporal energy, we apply the method of Ref.[35] to multi-directional video STS images. As shown in Figure 3, the leftward energy map 1(LE1), leftward energy map 2 (LE2), rightward energy map 1(RE1), rightward energy map 2 (RE2) are oriented liner responses maps, and total energy map (TE) are the point-wise sum of all the oriented energy maps. First, the mean oriented energy was computed according to Equations (11)-(14), and then the net motion energy RL was computed according to equation (15). Therefore, for each multi-direction video STS image, a total of five spatiotemporal energy features is computed.

$$T = sum(TE) \tag{10}$$

$$Rm1 = \frac{sum(RE1)}{T} \tag{11}$$

$$Rm2 = \frac{sum(RE2)}{T} \tag{12}$$

$$Lm1 = \frac{sum(LE1)}{T} \tag{13}$$

$$Lm2 = \frac{sum(LE)}{T} \tag{14}$$

$$RL = Rm1 + Rm2 - Lm1 - Lm2 \tag{15}$$

Where *sum*() denotes the sum of the values of all points in the feature map.

2.4 Features of the spatiotemporal filtering responses

Finally, we performed spatiotemporal filtering on the multi-direction video STS images. Spatiotemporal filtering is used to compute the spatiotemporal response similarity between reference and distorted video and increases correlation with subjective evaluation in ViS3. We performed the same log-Gabor filtering as ViS3 on the multi-directional video STS images, and obtained the filter coefficients in two scales and four spatial directions. Next, we calculated the amplitude and angle feature map of the filtering coefficients, and extracted the mean and standard deviation of the feature map as perceptual features for later training use.

$$Am = \sqrt{real(C_{\log\_Gabor})^2 + imag(C_{\log\_Gabor})^2} \tag{16}$$

$$Ang = arctan\left(\frac{imag(C_{\log\_Gabor})}{real(C_{\log\_Gabor})}\right) \tag{17}$$

Where Am and Ang denote the amplitude and angle feature map of the filtering coefficients, respectively. C_{log_Gabor} represent the filtering responses of log_Gabor, and they are complex numbers. Now to sum up, Table 1 summarizes the abovementioned features extracted from multi-directional video STS images.

Table 1. Summary of features extracted from multi-directional video STS images

	Feature Maps	Extracted features	Number of features
1	Multi-directional video STS image, Gradient map of multi-directional video STS, Angle map of multi-directional video STS gradient.	Mean value, standard deviation, skewness, kurtosis.	3*4=12
2	GM map and LOG filtering map of multi-directional STS image.	The marginal distributions and the independency distributions of GM and LOG (40)	1*40=40
3	Five energy maps of multi-directional STS image oriented.	Normalized energy feature (4), net energy.	5*1=5
4	Amplitude map of multi-directional STS image after spatiotemporal log-Gabor filtering (4 filters), Angle map of multi-directional STS image after spatiotemporal log-Gabor filtering (4 filters).	Mean value, standard deviation.	2*4*2=16
			total 73

2.5 Algorithm framework

The overall framework of the proposed NR algorithm is shown in Figure 4:

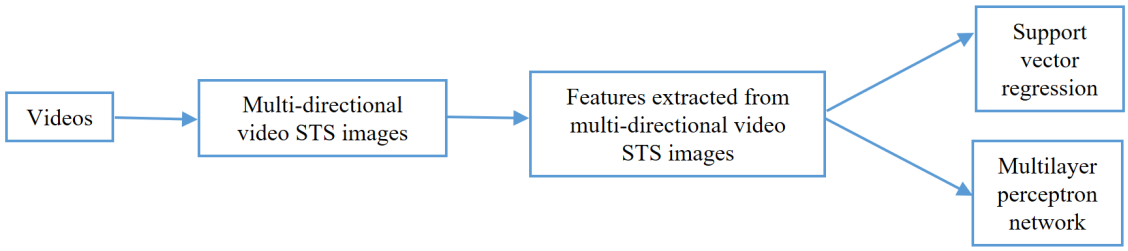


Figure 4. Overall algorithm flowchart

Figure 4 shows that all the features are extracted from multi-directional video STS images first, and then the extracted features are fed in SVR or multilayer perceptron network (MLP), which is a fully connected network, to train evaluation models. For learning-based methods, SVR is more effective for learning problems with a small sample size. Therefore, based on the above-extracted features, we first select an SVR to train and test. Meanwhile, given the slightly larger

sample size of KoNViD-1k VQA database, we constructed a shallow fully connected network for regression attempts. The neural network structure and specific network parameters are shown in Figure 5.

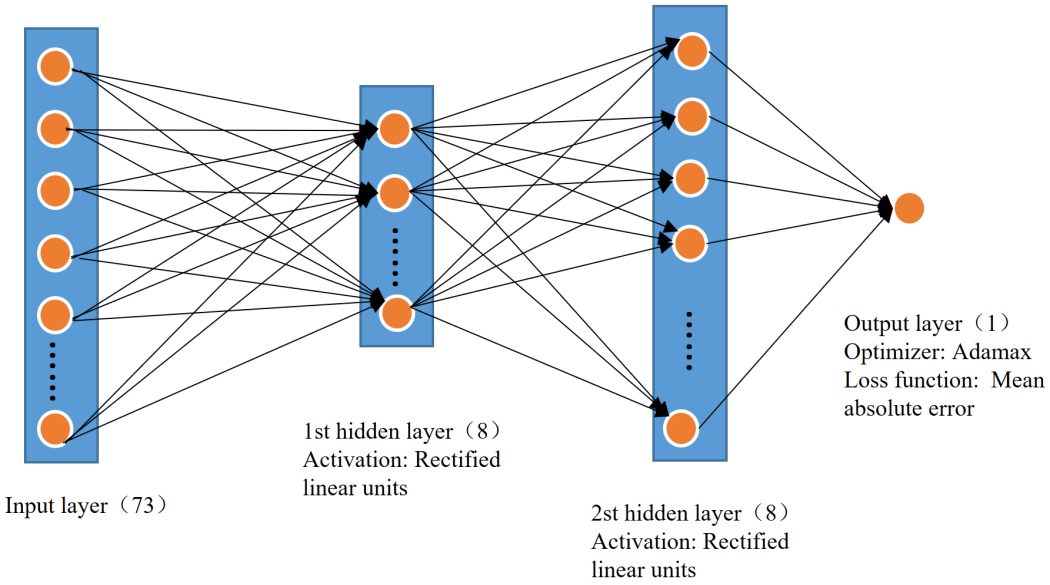


Figure 5. MLP network structure

3. EXPERIMENTS AND RESULTS

3.1 Experiments

Most subjective VQA databases are generated by adding various artificially generated noise to reference videos. Such as the well-known CSIQ[37], LIVE[38], IVP[39] and so on. Objective VQA models proposed or trained on these artificially generated databases are difficult to justify in the absence of validation on subjective VQA databases containing real distortions. In the past, there has been no large subjective VQA database with real distortion types. Fortunately, Vlad et al. have published the large subjective KoNViD-1k VQA database that contains 1200 videos with real distortion types. The database was designed to solve the problem of inability to perform deep neural network based VQA models training with limited data. Now, is a good database to validate existing NR VQA algorithms.

In order to verify the validity of the proposed algorithm, we chose to perform verification on KoNViD-1k database. First, the entire data set was randomly divided into 80% for the training and 20% for the testing. Next, all extracted features are fed into the SVR for regression. In the experiment, the kernel function of SVR is set as the radius basis kernel function. Finally, the training and verification process is repeated 1000 times. Furthermore, for comparison, we also input the extracted features into the fully connected network shown in Figure 3 for training, and testing.

3.2 Results and discussions

To evaluate the performance of VQA algorithms, we use three metrics: SROCC, PLCC and RMSE. SROCC, PLCC, and RMSE measure the monotonicity, linearity, and consistency of objective prediction and subjective evaluation, respectively. The median values of SROCC, PLCC and RMSE in 1000 test results of the SVR and deep network based algorithms are listed in Table 2.

Note that for the being compared methods (i.e., Video BLIINDS, VIIDEO, Video. CORNIA, STFC Model and the FC model), we use the results from [40].and [24]. For Video CORNIA, it was trained with the average CORNIA features over all frames in[24].

Table 2. Performance comparison of the proposed algorithm in term of SROCC, PLCC, and MSE.

	PLCC	SROCC	RMSE
Video BLIINDS[13]	0.565	0.572	0.526
VIIDEO[41]	-0.015	0.031	0.639
Video CORNIA[42]	0.747	0.765	0.412
FC Model[24]	0.492	0.472	0.556
STFC Model[24]	0.639	0.606	0.425
STS-MLP	0.407	0.420	0.610
STS-SVR	0.680	0.673	0.489

From Table 2, the proposed STS-SVR algorithm performs second in terms of SROCC and PLCC and rank third in terms of RMSE. And Video CORNIA rank first, it uses IQA CORNIA frame by frame to video and average pooling them. Inspired by this, in future work, the proposed STS-SVR algorithm can further improve prediction performance by adding more spatiotemporal features. The proposed STS-MLP perform badly among all the compared algorithms, which indicates that given a small sample size, the deep neural network is hard to train. Therefore, a neural network based VQA need new strategies and large samples to train.

In terms of features extraction times, the Ref. [24] only mentions that Video CORNIA takes about 254 seconds per video, and STFC needs about 56 seconds per video in KoNViD-1K VQA database, without introducing a specific computer configuration. Table 3 gives a summary of average time-consuming of feature extraction per video in KoNViD-1K VQA database. Given the computer configuration listed in the last row of Table 3, our proposed algorithm only takes about 3.13 seconds per video, which has significant time-consuming advantages.

Table 3. Average time-consuming summary of feature extraction per video in KoNViD-1K VQA database.

	times	Computer configuration
Video CORNIA[42]	254s	unknown
STFC Model[24]	56s	unknown
STS-MLP STS-SVR	3.13s	OS: Windows 10 (64 bits) CPU: Intel(R) Core i7-7700K @ 4.20GHz Memory: SanDisk DDR4@ 2400MHz 16 GB

4. CONCLUSIONS

In this paper, we propose methods to fast extract the perceptual features from multi-directional video STS images and obtain learning based VQA models through SVR and neural network. The verification results on the existing largest VQA database with real distortion types show that these NR features extracted from multi-directional video STS images are great for VQA designing to achieve better prediction performance by means of SVR, which provides a new idea for the feature designing of NR VQA algorithms. In addition, it is recommended that the NR VQA algorithms be validated on the VQA databases containing real distortion types. [1]

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China (No. 2016YFA0202003) and the National Natural Science Foundation of China (No. 61571359).

REFERENCES

[1] Li, C., Bovik, A. C., "Content-weighted video quality assessment using a three-component image model," J. Electron. Imaging Papers **19**(1), 011003 (2010).

[2] Narwaria, M., Lin, W., and Liu, A., "Low-complexity video quality assessment using temporal quality variations," IEEE Trans. Multimed. Papers **14**(3 PART1), 525–535 (2012).

[3] Watson, A. B., "Digital video quality metric based on human vision," J. Electron. Imaging Papers **10**(1), 20 (2001).

[4] Zhao, Y. et al., "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," IEEE Trans. Circuits Syst. Video Technol. Papers **21**(12), 1890–1902 (2011).

[5] Barkowsky, M. et al., "Temporal trajectory aware video quality measure," IEEE J. Sel. Top. Signal Process. Papers **3**(2), 266–279 (2009).

[6] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," IEEE Trans. Broadcast. Papers **50**(3), 312–322 (2004).

[7] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assessment based on structural distortion measurement," Signal Process. Image Commun. Papers **19**(2), 121–132, Elsevier (2004).

[8] Vu, P. V., Vu, C. T., and Chandler, D. M., "A spatiotemporal most-apparent-distortion model for video quality assessment," in Proceedings - International Conference on Image Processing, ICIP, pp. 2505–2508 (2011).

[9] Seshadrinathan, K. and Bovik, A. C., "Motion tuned spatio-temporal quality assessment of natural videos," IEEE Trans. Image Process. Papers **19**(2), 335–350 (2010).

[10] Vu, P. V. and Chandler, D. M., "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," J. Electron. Imaging Papers **23**(1), 013016 (2014).

[11] Yan, P., Mou, X., and Xue, W., "Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices," Proc. SPIE 9411, 94110M (2015)

[12] Yan, P. and Mou, X., "Video quality assessment based on motion structure partition similarity of spatiotemporal slice images," J. Electron. Imaging Papers **27**(03), 1 (2018).

[13] Saad, M. A., Bovik, A. C., and Charrier, C., "Blind prediction of natural video quality," IEEE Trans. Image Process. Papers **23**(3), 1352–1365 (2014).

[14] Chandler, D. M., "Most apparent distortion: full-reference image quality assessment and the role of strategy," J. Electron. Imaging Papers **19**(1), 011006 (2010).

[15] Ngo, C. W., Pong, T. C., and Zhang, H. J., "Motion analysis and segmentation through spatio-temporal slices processing," IEEE Trans. Image Process. Papers **12**(3), 341–355 (2003).

[16] Mittal, A., Moorthy, A. K., and Bovik, A. C., "Blind/referenceless image spatial quality evaluator," in Conference Record - Asilomar Conference on Signals, Systems and Computers, pp. 723–727 (2011).

[17] Mittal, A., Moorthy, A. K., and Bovik, A. C., "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process. Papers **21**(12), 4695–4708 (2012).

[18] Saad, M. A., Bovik, A. C., and Charrier, C., "A DCT statistics-based blind image quality index," IEEE Signal Process. Lett. Papers **17**(6), 583–586 (2010).

[19] Xue, W., Zhang, L., and Mou, X., "Learning without human scores for blind image quality assessment," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 995–1002, IEEE (2013).

[20] Kim, J. and Lee, S., "Fully Deep Blind Image Quality Predictor," IEEE J. Sel. Top. Signal Process. Papers **11**(1), 206–220 (2017).

[21] Li, Y. et al., "No-reference image quality assessment with deep convolutional neural networks," in International Conference on Digital Signal Processing, DSP, pp. 685–689, IEEE (2017).

[22] Kang, L. et al., "Convolutional Neural Networks for No-Reference Image Quality Assessment," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740, IEEE (2014).

[23] Bosse, S. et al., "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," IEEE Trans. Image Process. Papers **27**(1), 206–219 (2016).

[24] Men, H., Lin, H., and Saupe, D., "Spatiotemporal Feature Combination Model for No-Reference Video Quality Assessment," in International Conference on Quality of Multimedia (2018).

[25] Hosu, V. et al., "The Konstanz natural video database (KoNViD-1k)," in 2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017 (2017).

[26] Xue, W. et al., “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” IEEE Trans. Image Process. Papers **23**(2), 668–695 (2014).
[27] Sheikh, H. R. and Bovik, A. C., “Image information and visual quality,” IEEE Trans. Image Process. Papers **15**(2), 430–444 (2006).
[28] Wang, Z. et al., “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process. Papers **13**(4), 600–612 (2004).
[29] Zhang, L. et al., “FSIM: A feature similarity index for image quality assessment,” IEEE Trans. Image Process. Papers **20**(8), 2378–2386 (2011).
[30] Xue, W. and Mou, X., “Image quality assessment with mean squared error in a log based perceptual response domain,” in 2014 IEEE China Summit and International Conference on Signal and Information Processing, IEEE ChinaSIP 2014 - Proceedings, pp. 315–319, IEEE (2014).
[31] Zhang, L., Shen, Y., and Li, H., “VSI: A visual saliency-induced index for perceptual image quality assessment,” IEEE Trans. Image Process. Papers **23**(10), 4270–4281 (2014).
[32] Xue, W. et al., “Perceptual fidelity aware mean squared error,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 705–712, IEEE (2013).
[33] Zhang, M., Mou, X., and Zhang, L., “Non-shift edge based ratio (NSER): An image quality assessment metric based on early vision features,” IEEE Signal Process. Lett. Papers **18**(5), 315–318 (2011).
[34] Xue, W. et al., “Blind image quality assessment using joint statistics of gradient magnitude and laplacian features,” IEEE Trans. Image Process. Papers **23**(11), 4850–4862 (2014).
[35] Adelson, E. H. and Bergen, J. R., “Spatiotemporal energy models for the perception of motion,” J. Opt. Soc. Am. A Papers **2**(2), 284, Optical Society of America (1985).
[36] Yan, P. and Mou, X., “Video quality assessment based on correlation between spatiotemporal motion energies,” in Proceedings of the SPIE 9971, 997130 (2016).
[37] Larson and D. M. Chandler, “Databases: CSIQ Image Quality Database” (accessed 12 September 2018).
[38] Seshadrinathan, K. et al., “Study of subjective and objective quality assessment of video,” IEEE Trans. Image Process. Papers **19**(6), 1427–1441 (2010).
[39] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. N., “IVP subjective quality video database,” 2011 (accessed 29 July 2018).
[40] Men, H., Lin, H., and Saupe, D., “Empirical evaluation of no-reference VQA methods on a natural video quality database,” in 2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017, pp. 1–3 (2017).
[41] Mittal, A., Saad, M. A., and Bovik, A. C., “A completely blind video integrity oracle,” IEEE Trans. Image Process. Papers **25**(1), 289–300 (2016).
[42] Peng Ye et al., “Unsupervised feature learning framework for no-reference image quality assessment,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1098–1105 (2012).