ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

Spatiotemporal Slice No-Reference Video Quality Assessment Model using 2DLOG Filter and Support Vector Machine

[1] Daniel Oppong Bediako, [2] Maxwell Soubgobiree

[1] Faculty of Engineering and Computer Science, Kaaf University College, Ghana
Corresponding Author Email: [1] danielbediako@stu.xjtu.edu.cn

Abstract*— A key challenge in no reference video quality assessment (NR-VQA) is how to effectively mimic human visual system (HVS) in a data-driven way. A spatiotemporal Slice no-reference video quality assessment model using 2DLOG Filter and Support Vector Machine based on frame-level unsupervised feature learning and temporal pooling is presented in this paper. Given that the spatial and temporal channels of the human visual system both include the second derivative of the Gaussian function, first a two-dimensional LOG (2D LOG) filter was constructed to simulate a human visual filter and to extract perceptual-aware features for the design of VQA algorithms in order to filter the STS images and use support vector machine (SVM) to perform a successful NR video quality evaluation in this dissertation. Secondly, the successful features of STS images of video such as the statistical feature maps, orientation feature map from the gradient magnitude and filtering response of Laplacian of Gaussian were extracted to characterize the motion statistics of videos. Finally, the extracted perceptual features were fed in SVM to perform training and testing. The performance of proposed algorithms shows that the methods are better than that of most mainstream VQA methods.

Index Terms*— No-reference video quality assessment; spatiotemporal slice (STS) images; Human visual system (HVS); two-dimensional LOG (2D LOG).

I. INTRODUCTION

HVS is considered to be well-suited to extracting structural information in order to evaluate the quality of a visual scene [1–4][5][6][7–13] [14–20]. The spatio-temporal properties of the regular visual 205 structures are disrupted by video distortions. As a result, determining the regularity of visual structures includes important information about the visual scene's quality.

The use of HVS models to develop quality indices has been the focus of a significant amount of IQA and VQA research. The basic idea behind these approaches is that in the absence of any knowledge of the distortion process, the best way to predict the quality of an image or video is to use a system similar to the HVS to "see" the image. In an attempt to model the tuning properties of neurons in the front-end of the eye-brain system, standard HVS-based indices use linear transforms separably in the spatial and temporal dimensions to decompose the reference and test videos into multiple channels.

Simulation of HVS processing is required for HVS-based algorithms. Many quality evaluation algorithms employ visual filtering as one of the most efficient HVS processing methods. MOVIE, for instance, employs 3D Gabor filtering. VIS3 employs Gabor filtering, THV-JMG employs the Gabor filter, and 3DLOG-CORR employs video LOG filtering [21]. In addition to the use of visual filtering in VQA, Laplacian of Gaussian (LOG) filtering, which is a technique for extracting first-order and second-order information redundancy in images, has performed well in IQA[22]. Given the importance of visual filtering in perceptual feature extraction, LOG can extract the image's perceived features. Given the above contribution and their importance for video quality assessment (VQA), we design No reference 2D Laplacian of Gaussian (NR2DLOG-VQA) algorithm using LOG-based visual filtering in this paper. The extracted features of video slice images using the designed 2D LOG filter is fed into SVM to train the proposed method.

The rest of the paper is structured as follows: Section 2 presents related works. Section 3 proposed method and the detail features we used. Section 4 presents the experiment setup and result. Section 5 discussion and conclusion.

II. RELATED VQA METRICS

We introduce some well-known performers (VQA) as follows. First, the National Telecommunications and Information Administration (NTIA) [23] developed the Video Quality Metric (VQM) algorithm. The algorithm depends on the loss in computing the spatial gradients of the luminance features components and the color impairment of the VQM. The American National Standard Institute (ANSI) and International Telecommunications Union Recommendation [24] have adopted VQM as a national standard in the VQEG Phase II validation tests. In [25], the authors introduce a model approach to VQA and decompose the video into spatiotemporal features using a 3D gradient that integrates both spatial and temporal slice information to measure video quality. Then, for each group of frames

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

(GOF), a three-dimensional gradient masking is employed to obtain each pixel in all directions. Based on the index, the authors combined the spatiotemporal gradient differencing GSDST between the reference and the distorted video block using the gradient in all directions. The authors believe that using a machine learning strategy can improve performance quality.

Freita el [26]. Each set of features is calculated independently of the other. Each set is concatenated to generate a vector feature. The feature vector is used as an input to predict the quality score in a random forest regression (RFR). Although the proposed approach outperforms state-of-the-art video quality metrics for data sets and distortion types, it fails to do the best for the other metrics on the data sets. To continuously monitor the level of exposure, a no-reference metric was developed in [27]. For video conferencing systems, an NR-IQA method was also proposed. A support vector machines (SVM) classifier is used in the method to differentiate between poor and good image sequences [28]. There are many successful NR IQA algorithms for NR quality algorithms, such as blind or referenceless image spatial quality evaluator (BRISQUE) [29], distortion identification-based image verity and integrity evaluation (DIIVINE) [30], blind image integrity notator using DCT statistics (BLIINDS) [31], quality-aware clustering (QAC) [32], and deep CNN-based NR IQA [33–36]. V-BLIINDS [37] and the spatiotemporal feature combine model (STFC) [38] are two NR VQA algorithms'-BLIINDS creates a blind VQA algorithm that correlates highly with human judgments of quality by combining a spatiotemporal natural scene statistics (NSS) model for videos and a motion model that quantifies motion coherency in video scenes. STFC uses vertical STS images to train an SVM model by combining spatial (such as contrast and color) and temporal (such as sharpness and exposure time) features. Among the NR algorithms mentioned above, the STS-based algorithms achieved very good perceptual evaluation results, indicating that the perceptual features extracted in the STS can be used more effectively for NR video quality evaluation. In fact, whether NR algorithm is used, the extraction of perceptual features is important, and its effectiveness determines whether the evaluation algorithm succeeds or fails.

III. IMPLEMENTATION OF NR2DLOG-VQA ALGORITHM

The system model that was utilized to determine the quality of a video is explained in this chapter. Figure 1 shows how we used the 2DLOG Filter and Support Vector Machine (NR 2DLOG) to illustrate the idea of spatiotemporal Slice no-reference video quality assessment. All of the features are extracted from STS images of video, and the 2D-LOG statistical features are fed into a fully connected network called SVM to train evaluation models. SVM is more effective for learning problems with a small sample size when using learning-based methods. As a result, we begin by selecting an SVM to train and test based on the above-extracted features. By utilizing 2D LOG-based features, we first develop a complete representation method. Then, we show how to employ a sophisticated Support Vector Machine (SVM) extension to take into account the uncertainty encapsulated in the representation of the input videos during training. In addition, the experimental results reveal that the proposed method has achieved state-of-the-art quality prediction performance on the largest extant annotated video database, paving the way for additional research in this field. An SVR model is then trained with mean option scores to predict video quality scores (MOS).

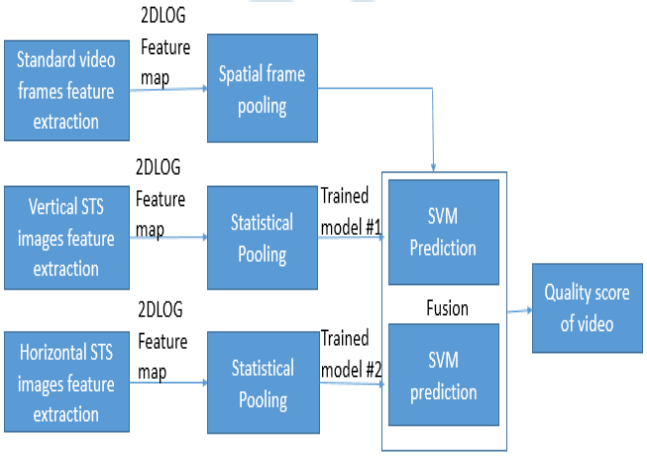

Fig. 1. Framework of 2DLOG-STS model

IV. SUPPORT VECTOR MACHINES

NR VQA techniques based on training/learning frequently rely on a large number of features meant to capture significant elements that affect video quality. The mapping from feature space to image quality is then learned using various regression approaches such as support vector machine (SVM) and neural network. V. N. Vapnik and A. Ya. Chervonenkis (Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia) proposed many of the concepts that are now being developed in the context of Support Vector Machines in the context of the "Generalised Portrait Method" for computer learning and pattern recognition. In 1962, these ideas began to form, and they were initially published in 1964. SVMs are a popular and active area of Machine Learning research right now, and they are perfect example of "kernel approaches." SVMs are a class of supervised learning methods for classification and regression that are all quite similar [39].

They are part of the generalized linear classifier family. Support Vector Machine (SVM) is a classification and regression prediction tool that employs machine learning theory to enhance forecast accuracy while automatically avoiding over-fitting to the data. Support Vector Machines are systems that use the hypothesis space of linear functions in a high-dimensional feature space and are trained with an

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

optimization theory-based learning algorithm that incorporates a learning bias derived from statistical learning theory. The NIPS group first popularized the support vector machine, and it is now widely used in machine learning research around the world. SVM becomes well-known when, given pixel maps as input, it achieves accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task. It is also used in a variety of applications, including handwriting analysis, face analysis, and so on, with a focus on pattern classification and regression-based applications. Vapnik [40] built the foundations of Support Vector Machines (SVM), which have gained prominence because to a number of promising features such as improved empirical performance. The method employs the Structural Risk Risk Minimization (SRM) principle, which has been shown to outperform the traditional Empirical Risk Minimization (ERM) principle utilized in conventional neural networks [41]. ERM minimizes the error on the training data, whereas SRM minimizes an upper bound on the expected risk. SVM has a higher ability to generalize as a result of this difference, which is the purpose of statistical learning. SVMs were originally designed to address classification issues; however they have recently been expanded to solve regression difficulties [42].

Initially, machine learning algorithms were designed to learn representations of simple functions. As a result, the goal of learning was to generate a hypothesis that correctly classified the training data, and early learning algorithms were designed to achieve such an accurate fit to the data [43]. Generalization refers to a hypothesis's ability to correctly classify data that is not in the training set. When neural networks overgeneralize easily, SVM performs better. To distinguish professional videos from amateurish ones, the authors in [44] treat the video as a sequence of still images to which they apply a set of visual-based features as well as two additional motion-based features, namely the length of subject region motion and motion stability. They also used kernel SVM, Bayesian classification, and Gentle AdaBoost as well as other learning techniques. [45] describes a more complex method that incorporates a set of features ranging from low- and mid-level attributes to high-level style descriptors, as well as a kernel SVM learning stage. Furthermore, in [46], low- and high-level visual and motion features are extracted at the cell, frame, and shot levels, and a Low Rank Late Fusion (LRLF) scheme is used to fuse the scores produced by a set of SVMs, each of which has been trained with a different aesthetic feature. In [47], the authors assess the effectiveness of motion space, motion direction entropy, and hand shaking (i.e., camera stabilization) on VAQ assessment tasks. They also employ classification techniques such as naive Bayesian, SVM, and AdaBoost. As previously discussed, most researchers have used SVM to model HVS well when compared to other approaches.

Unsupervised learning focuses on extracting hidden structure in unlabeled data. It's assumed that the training data has not manually labeled. It also looks for patterns in the data that can be used to calculate the correct output value for new data instances. Clustering, adaptive resonance theory (ART), hidden markov model (HMM), radial basis function (RBF), conditional random field (CRF), and other unsupervised learning approaches are a few examples. The majority of pattern classification techniques that use numerical inputs are classified as parametric, semi-parametric, or non-parametric.

The two most widely used approaches for supervised learning are support vector machine (SVM) and k-nearest neighbor classifier. Because of its robust performance, Support Vector Machine (SVM) is the most popular application for classification in action recognition and understanding. We devised a learning-based pooling strategy to automatically combine the feature maps to generate a final video quality score due to its high performance.

V. APPLICATION OF NR-VQA TO VIDEO SLICE

Due to the widespread use of multimedia services in the context of wireless communications and telecommunication systems, there has been an increase in interest in the development of NR methods in recent years. The following are some of the areas where NR methods are used:

- Network operators and content providers are keen to objectively quantify the level of service quality provided to end users and within network nodes. NR methods will provide the data required to implement network settings that ensure customer satisfaction and, as a result, reduce churn.
- Due to the involvement of multiple parties between content providers and end users, service-level agreements (SLA) must be established in order to guarantee an agreed level of quality. In this regard, NR methods are a good option for monitoring in-service quality in live systems.
- NR methods, in general, are well suited for performing real-time objective quality assessment in resource constrained environments, such as the frequency spectrum in wireless communications. In these cases, RR methods are limited because an ancillary channel is required to transmit the required features of the original video.
- Quality adaptations using NR methods for collecting statistics of the delivered quality are required for real-time communication and streaming services.

Thus, predicting the quality of compressed and transmission videos is of great interest and NR-VQA are proposed to achieve this goal. The NR2DLOG-VQA algorithm model is proposed to blindly predict the quality of natural videos. In the first stage, each decoded frame of the video sequence is decomposed into spatiotemporal slice in horinzontal and vertical direction. four efficient statistical

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

features, i.e., mean value, standard deviation, and LOG. In the second stage, each frame-level feature is averaged across all frames (temporal pooling); a trained SVM network takes the four features as inputs and outputs a single number as the predicted video quality. NR2DLOG-VQA algorithm model was trained and tested on LIVE VQA database. The results show that the objective assessment of the proposed model has a strong correlation with the subjective assessment.

VI. SPATIOTEMPORAL FEATURE EXTRACTION

Many researchers have successfully used image spatiotemporal features, such as a model of human visual-motion sense and video motion analysis. In this study, we argue that the temporal variation of spatial distortion manifests itself in model design as spatiotemporal dissimilarity, and that these designs can thus be used to estimate video quality. A video can be visualized in spatiotemporal space as a cubiod, with the sides of the columns and rows representing the spatial dimensions x and y, respectively, and the third side indicating the time dimension t, as illustrated in Figure 5.2. The typical view of a video is when all of the frames are displayed in front-to-back sequence when a video is played properly. A video, on the other hand, can be regarded of as being played from left to right or from top to bottom, as indicated by the terms left-right view and top-bottom view. The vertical STS images are made up of extracted slices that depict time in one dimension and vertical space in the other dimension. When a cuboid is cut horizontally, the retrieved slices represent time in one dimension and horizontal space in the other, and are referred to as horizontal STS images. As illustrated in Fig 2, these slices reveal temporal information.

The STS images will appear as random patterns if the video is rapidly changing. In STS images, the randomness of temporal content manifests as spatially random pixels along the dimension that corresponds to time. Because of the joint spatiotemporal relationship of neighboring pixels and the smooth frame-to-frame transition, STS images for normal videos are generally well structured.

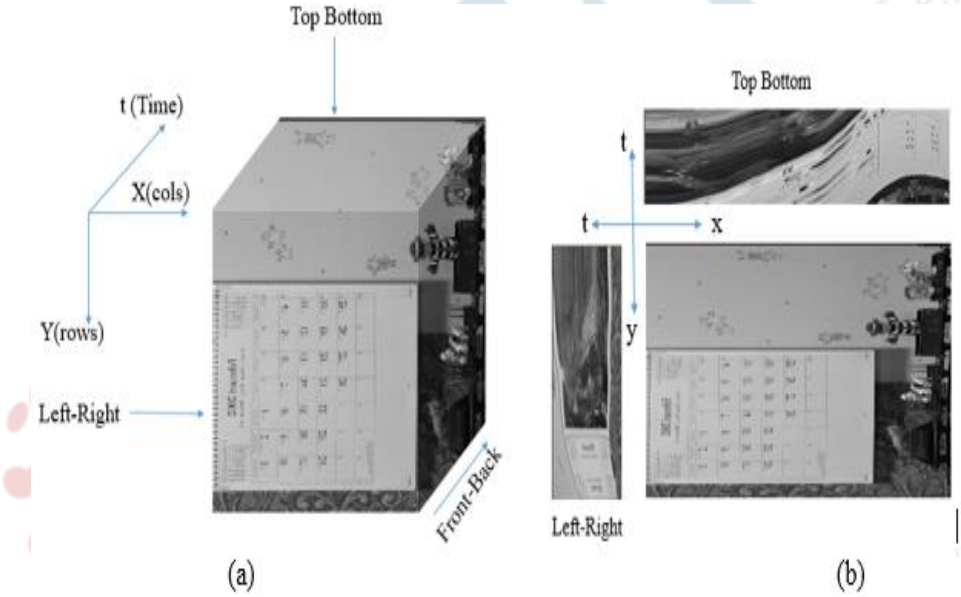

Fig. 2. (a) A video can be viewed as a cubiod in spatiotemporal space, where two of the sides represents the spatial dimensions (x and y) the third side denotes the time dimension (t). (b) Slices from the top-bottom view (top- and right), left-right view (bottomleft) and normal view (bottom-right) of the same video.

Pixels arranged along two spatial and one temporal axis could be considered natural videos. Not only in the spatial domain, but also in the temporal domain, pixels have strong dependencies. Observing the pixel volume from various perspectives, we can see that "images" in the x-t and y-t dimensions exhibit highly structured characteristics similar to those in the x-y dimension, as shown in Fig. 3. In addition to structural distortions in the x-t and y-t dimensions, the distorted video has blocky or blurring artifacts. As we all know, videos not only give us spatial information about natural scenes, but also temporal information about motion. Highly structured features along the temporal axis result from local motion consistency. As a result, we propose that information from patches in the x-t and y-t dimensions be treated in the same way as information from patches in the x-y dimension.

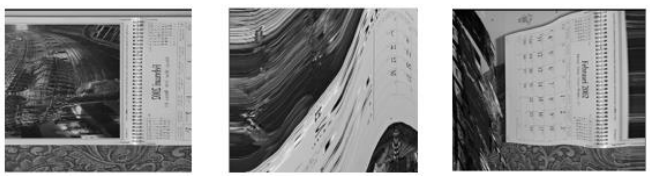

Figure 3. Demonstrative STS images extracted from the LIVE videos database.

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

VII. ALGORITHM FRAMEWORK

Knowledge of HVS will be essential for extracting perceptual-aware features. HVS has some important properties that can help extract important information from complex visual environments and improve survival chances. For starters, center-surrounded receptive fields have been discovered in the retina and LGN. By extracting the edge and contourprofile of objects, the center-surrounded receptive fields can reduce visual redundancy. We successfully extract STS images of video features for assessing video quality using the 2DLOG filter.

Second, in HVS [48–52], there are only three temporal channels that can be simulated using the logarithmic time Gaussian function and its first-order, second-order, or third-order derivative [53,54]. Temporal filters modulate the visual system's sensitivity to temporal patterns. The video signal contains temporal information in addition to spatial information, which is an important part of visual perception. Temporal features of video must be extracted by temporal channels in order to design video quality evaluation methods. We construct a two-dimensional LOG (2D LOG) filter to simulate human visual perception filter and extract the abovementioned perceptual-aware video features for the design of VQA algorithms because the spatial and temporal channels of the human eye both include the second derivative of Gaussian function.

A video signal can be represented as three-dimensional data v (x,y,t), with x and y representing spatial dimensions and t representing time. Note that we are currently only concerned with the luminance component in the video. We use a 2D LOG filter to extract spatial and temporal features from reference videos simultaneously for the video luminance component, and then use the extracted features to design the model for quality indicator.

VIII. 2DLOG Filtering Response

The Laplacian is a 2-D isotropic measure of the second spatial derivative of an image. It highlights regions of rapid intensity change and is often used for edge detection in 3D digital imaging. The Laplacoian L(x,y) is given as follows:.

$$\Delta^2(x,y)=\frac{\partial^2 I}{\partial x^2}+\frac{\partial^2 I}{\partial y^2} \tag{1}$$

It is particularly sensitive to noise because it is simulating a second derivative measurement on an image. The Laplacian is frequently applied to an image that has already been smoothed with a Gaussian smoothing filter to lower its sensitivity to noise. This combined filter is known as the Laplacian of Gaussian filter. The 2-D LoG (Laplacian of Gaussian) filter function with Gaussian standard deviation σ can be written as:

$$\Delta^2 g(x,y)=\Delta^2 g_{\sigma_x\sigma_y}(x,y) \tag{2}$$

$$=\left(\frac{\partial^2 g_{\sigma_x\sigma_y}(x,y)}{\partial x^2}+\frac{\partial^2 g_{\sigma_x\sigma_y}}{\partial y^2}\right) \tag{3}$$

$$=\frac{g^{''}_{\sigma_x}(x)g_{\sigma_y}(y)}{V-LOG}+\frac{g_{\sigma_x}(x)g^{''}_{\sigma_y}(y)}{H-LOG} \tag{4}$$

Where ∇^2 represents the Laplacian operators, g (x, y) represents the 2D Gaussian function, and σ_x and σ_y represent the Gaussian standard deviation. We must design a discrete convolution kernel that can approximate the Laplacian operator because the image is represented as a series of discrete pixels. We set Gaussian value of σ = 1.4 and get a LoG operator as shown in Figure 4. To make the computation easier, we use a 3x3 Laplace operator, as shown in fig 5.

0	1	1	2	2	2	1	1	0
1	2	4	6	6	6	4	2	1
1	4	6	3	0	3	5	4	1
2	5	3	-12	-24	-12	2	5	2
2	6	0	-24	-40	-24	0	6	2
2	6	3	-12	-24	-12	3	6	2
1	4	5	3	0	3	5	4	1
1	2	4	6	6	6	4	2	1
0	1	1	2	2	2	1	1	0

Figure 4. Discrete approximations LoG (σ = 1.4)

-1	-1	-1
-1	8	-1
-1	-1	-1

Figure 5. Laplace Operators

The computation of the 2D LOG filter includes three parts: the horizontal LOG filter (H-LOG), the vertical LOG filter (V-LOG), and the temporal LOG filter (T-LOG). In the following experiment, the effects of the three parts on video quality will be investigated further.

We utilize this 2D LOG filter to extract spatial and temporal features from natural videos at the same time. The filtering output *f* for a video signal *v* is obtained by convolution of *v(x, y, t)* and 2D LOG filter$\nabla^2 g$ *(x, y)*, as indicated in Equation 5.

$$f=v\otimes\nabla^2 g(x,y)=v(x,y,t)\otimes\nabla^2 g_{\sigma_x\sigma_y}(x,y) \tag{5}$$

$$=v(x,y,t)\otimes(V-LOG+H-LOG) \tag{6}$$

$$=v(x,y,t)\otimes V-LOG+v(x,y,t)\otimes H-LOG \tag{7}$$

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

Inspired by Gaussian, which used in connection with the laplacian to predict the visual quality, here, we used the Laplace operator as simple gradient based metrics magnitude GM. The gradient magnitudes of the reference video are computed as follows:

$$G_m = \sqrt{\left(i_{STS} \otimes L_x\right)^2 + \left(i_{STS} \otimes L_y\right)^2} \quad (8)$$

Where symbol "⊗" denotes the convolution operation, and L_x and L_y are the laplacian filters along horizontal and vertical directions, respectively. i_{STS} is the spatiotemporal slice of images.

The algorithm has now delivered maps, which may be combined to generate a summary video quality score. Apply standard deviation pooling to each VQA map for simplicity of use, resulting in a single score for each map:

$$S_m(d) = \sum_{i=1}^{N} p_m(i,d) / N, \quad (9)$$

The map scores over all indices m and STS dimensions $d \in \{T, H, W\}$ can then be concatenated into a single 1-D vector S. Since the elements of S will fall into different ranges, simple pooling strategies like mean, max or min-pooling will lead to poor results. Since there is no simple way to model or understand the mapping from the elements of S to perceived video quality, we use a learning based pooling strategy in order to predict the video quality score using an SVR model to train with mean opinion scores (MOS).

IX. COMPARISON WITH EXISTING ALGORITHM

The KoNViD-1k videos are encoded at three different frame rates: 27, 25, and 30 frames per second. There are a total of 12 resolutions, with 1280720 pixels accounting for the largest percentage of videos (85%), followed by 19201080 pixels (9 percent). Between the KoNGNiDs and the KoNGNId-125k, the proportion of all of these features is similar.

In total, 30 features are extracted from two scales of STS video images of video. Six gradient features, four 2D LOG features, and two gradient orientation features are included in each scale. The entire data set, on the other hand, was randomly divided into two parts: 80 percent for training and 20% for testing. Ten times the training and testing process is performed. The following are the details of the test: A total of Nx1200 patches with the size of 32x32 are extracted from each entire STS image (N=100, 200). Table 1 summarizes and lists these characteristics. We use Support Vector Regression (SVR) to combine the features into a single quality score to derive a score for overall video quality. A portion of the dataset is designated as training videos, while the remainder is designated as test videos. We partition a training video into separate frames and utilize each frame as a training instance to extract features for training during the training process. We also divide a test video into separate frames throughout the testing procedure so that the trained SVR model can integrate the features of each frame into a score. The quality score of a testing video is then computed by averaging the scores of all frames in the video.

Table 1 Summary of features extracted by our proposed model

Feature type	Feature description
Gradient Features of STS images	Standard deviation and entropy of Gradient maps
2D LOG Feature of STS images	2D LOG map statistics including Mean and standard deviation
Gradient Orientation Features of STS images	Mean value, standard deviation

X. EXPERIMENT RESULTS

All of the strategies we employed in our experiments were validated on the KoNViD-1k [55]. The overall performance was evaluated using k-fold (k=5) cross validation. We use the results from [56] for other approaches (e.g., Video BLIINDS, VIIDEO, STFC Model, and FC Model). We utilize three measures to assess the performance of VQA algorithms: SROCC, PLCC, and RMSE. The monotonicity, linearity, and consistency of objective prediction and subjective evaluation are measured by SROCC, PLCC, and RMSE, respectively. Table 5.6 shows the median values of SROCC, PLCC, and RMSE in the test results of the SVR algorithms.

From Table 2, the proposed 2DLOG-STS algorithm performs well in terms of SROCC and PLCC and rank fourth in terms of RMSE. In terms of PLCC, STS-SVR is ranked second, as it employs spatiotemporal filtering on multi-direction STS images of videos. The proposed 2DLOG-STS algorithm may further improve prediction performance by adding more spatiotemporal information, as demonstrated by the best performance with the highest PLCC and SROCC in video CORNIA. However, our 2DLOG-STS model, on the other hand, is substantially simpler, with a run duration of 56 seconds on average per video.

We also evaluated the framework's performance on the KoNViD-1k video database while employing the Laplace Operator, which we adapted into 2DLOG-STS instead of the LoG filter. Fig. 5.7 depicts the comparing findings in terms of SRCC. It can be shown that the LoG filtered map produces better correlations against human subjectivity than the other created maps.

ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 11, Issue 3, March 2024

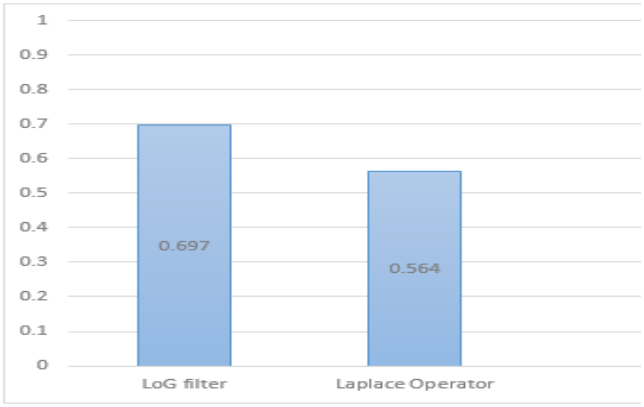

Fig. 5.7. Performance comparison of 2DLOG-STS WHEN USING LoG filter and Laplace Operator, respectively

Table 2. Comparison of the performance by PLCC, SROCC and RMSE on the KoNVid-1k Videos

	SROCC	PLCC	RMSE
Video BLINDS	0.572	0.565	0.526
VIDEO	0.031	-0.015	0.639
V.CORNIA	0.747	0.765	0.412
STFC Model	0.605	0.639	0.425
FC model	0.492	0.472	0.556
STS-SVR	0.673	0.680	0.489
2DLOG-STS	0.679	0.659	0.491

XI. SUMMARY

The proposed method requires the extraction of a complete STS features using 2DLOG filter. The propose methods is to fast extract the perceptual features from STS images of video and obtain learning based VQA models through SVR. Although the robustness to incomplete has been raised by this method, but the main purpose is to establish distinguishable algorithm that can extract good features based on localized spatial and temporal variation along different part of directions. However, higher degree is obtained with the help of well design 2DLOG filter. The verification results on the existing largest VQA database with real distortion types show that these NR features extracted from STS images of video 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.

XII. CONCLUSION

Various contributions to the field of no-reference (NR-VQA) video quality assessment have been given in this paper. A new NR-VQA model based on a Laplacian of Gaussian decomposition has been given, which has a lower computational cost than the previously described DCT-based model. The proposed model has two primary steps: measuring distortion and predicting video quality. Each frame of the distorted video sequence was first decomposed into a spatiotemporal slice, after which the response was convolved with a 2DLOG filter. The statistical features were completely utilized. The features were fed into the prediction model as inputs. It produced a single score representing the predicted video quality. The performance of the proposed method was evaluated on the KoNViD-1k video database. The predicted quality scores were well correlated with the MOS associated with the subjective assessments, per the findings. Finally, when compared to another well-known method, the performance of VQA algorithms was significantly improved.

REFERENCES

[1] W. Lin, C.-C. J. Kuo, Perceptual visual quality metrics: A survey, Journal of Visual Communication and Image Representation 22 (4) (2011) 297–312.

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of subjective and objective quality assessment of video, IEEE transactions on image processing 19 (6) (2010) 1427–1441.

[3] M. Lin, D. Chenwei, K. N. Ngan, L. Weisi, Recent advances and challenges of visual signal quality assessment, China Communications 10 (5) (2013) 62–78.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, L. J. Karam, Objective video quality assessment methods: A classification, review, and performance comparison, IEEE transactions on broadcasting 57 (2) (2011) 165.

[5] A. C. Bovik, Automatic prediction of perceptual image and video quality, Proceedings of the IEEE 101 (9) (2013) 2008–2024.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[7] Z. Wang, L. Lu, A. C. Bovik, Video quality assessment based on structural distortion measurement, Signal processing: Image communication 19 (2) (2004) 121–132.

[8] Y. Wang, T. Jiang, S. Ma, W. Gao, Novel spatio-temporal structural information based video quality metric, IEEE transactions on circuits and systems for video technology 22 (7) (2012) 989–998.

[9] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Transactions on Image Processing 21 (4) (2012) 1500–1512.

[10] D. Liu, Y. Xu, Y. Quan, Z. Yu, P. Le Callet, Directional regularity for visual quality estimation, Signal Processing 110 (2015) 211–221.

[11] J. Zhang, T. M. Le, S. H. Ong, T. Q. Nguyen, No-reference image quality assessment using structural activity, Signal Processing 91 (11) (2011) 2575–2588.

[12] M. Narwaria, W. Lin, A. E. Cetin, Scalable image quality assessment with 2d mel-cepstrum and machine learning approach, Pattern Recognition 45 (1) (2012) 299–313.

[13] Q. Wu, H. Li, F. Meng, K. N. Ngan, S. Zhu, No reference image quality assessment metric via multi-domain structural information and piecewise regression, Journal of Visual Communication and Image Representation 32 (2015) 205–216.

[14] X. Li, Q. Guo, X. Lu, Spatiotemporal statistics for video quality assessment, IEEE Transactions on Image Processing 25 (7) (2016) 3329–3342.

[15] Q. Li, W. Lin, Y. Fang, No-reference quality assessment for multiply-720 distorted images in gradient domain, IEEE Signal Processing Letters 23 (4) (2016) 541–545.

[16] Q. Li, W. Lin, J. Xu, Y. Fang, Blind image quality assessment using statistical structural and luminance features, IEEE Transactions on Multimedia 18 (12) (2016) 2457–2469.

[17] G. Yue, C. Hou, K. Gu, N. Ling, B. Li, Analysis of structural characteristicsmfor quality assessment of multiply distorted images, IEEE Transactions on Multimedia 20 (10) (2018) 2722–2732.

[18] P. G. Freitas, W. Y. L. Akamine, M. C. Farias, No-reference image quality assessment using orthogonal color planes patterns, IEEE Transactions on Multimedia 20 (12) (2018) 3353–3360.

[19] B. Wandell, Foundations of vision. 1995, Sinauer, Sunderland, MA.

[20] R. Blake, R. Sekuler, Perception, 5th ed. New York, NY, USA: McGrawHill, 2006.

[21] P. Yan, and X. Mou. ‘’Video quality assessment based on LOG filtering of videos and spatiotemporal slice images’’ Optoelectronic Imaging and Multimedia TechnologyVI,v.11187,pp.1118709.[doi: 10.1117/12.2536872]

[22] Mou, X., Xue, W., Chen, C., Zhang, L., Paper, C., Society, T. I., Engineering, O., Xi, X. M., Ontario, W., Mou, X., Xue, W., Chen, C. and Zhang, L., “LoG acts as a good feature in the task of image quality assessment,” 1–7 (2014).

[23] Recommendation ITU-R BT.1907 - Objective perceptual video quality measurement techniques for broadcasting applications using HDTV in the presence of a full reference signal,” 26 (2012)

[24] M. H. Pinson, N. Staelens, and A. Webster, “The history of video quality model validation,” IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 458–463 (2013) [doi:10.1109/MMSP.2013.6659332].

[25] W. Lu et al., “A spatiotemporal model of video quality assessment via 3D gradient differencing,” Information Sciences 478, 141–151 (2019) [doi:10.1016/j.ins.2018.11.003].

[26] P.G. Freitas, W. Y. L .Akamine and M. C .Q. Farias, "Using Multiple Spatio-Temporal Features to Estimate Video Quality," Signal Processing: Image Communication, 64 (1-10) [doi:10.1016/j.image.2018.02.010]

[27] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, \A no reference metric for the quality assessment of videos affected by exposure distortion," in IEEE International Conference on Multimedia and Expo, Barcelona, Spain, Jun. 2011.

[28] L. Yu, X. Tian, T. Li, and J. Tian, \No-reference image quality assessment based on SVM for video conferencing system," in International Conference on Network Computing and Information Security, vol. 345, Shanghai, China, Dec. 2012, pp. 555-560

[29] 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).

[30] J. Y. Lin, C.-H. Wu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, “Evqa: An ensemble-learning-based video quality assessment index,” IEEE Int’l Conf. Multim. & Expo Worksh., pp. 1–6, 2015.

[31] 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).

[32] 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).

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

[34] 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).

[35] 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).

[36] 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).

[37] 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).

[38] 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).

[39] D. Peijun, B. Xuyu, T. Kun, X. Zhaohui, S. Alim, X. Junshi, L. Erzhu, S. Hongjun and Liu, Wei. ‘’ Advances of Four Machine Learning Methods for Spatial Data Handling: a Review’’ Journal of Geovisualization and Spatial Analysis, vol. 4, pp. 1-25, 2020. [doi:10.1007/s14651-020-00048-5]

[40] C. Cortes, V. Vapnik. ‘’Support-vector networks’’. Mach Learn 20(3): 273–297(1995)

[41] M. Shahid, A. Rossholm, and B. Lovstrom. ‘’A no-reference machine learning based video quality predictor’’ 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. {176--181} {2013}

[42] Z. Lin, G. Zhongyi, L. Xiaoxu, L. Hongyu and L. ‘’Jianwei. Training quality-aware filters for no-reference image quality assessment’’ IEEE MultiMedia. Vol 21,pp. 67—75, 2014

[43] H. Rui, Z. YunHao, Hu. Yang and L. Huan, ‘’ No-reference video quality evaluation by a deep transfer CNN architecture’’, Signal Processing: Image Communication. Vol 83, pp. 115782, 2020

[44] Y. Luo and X. Tang, “Photo and video quality evaluation: Focusing on the subject,” in Proc. of the 10th European Conference on Computer Vision (ECCV), Marseille, France, pp. 386–399. Springer, 2008.

[45] Y. Wang, Q. Dai, R. Feng, and Y.-G. Jiang, “Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data,” in Proc. of the 21st ACM Int. Conf. on Multimedia. ACM, 2013, pp. 369–372.

[46] Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, “Towards a comprehensive computational model for aesthetic assessment of videos,” in Proc. of the 21st Int. Conf. on Multimedia. ACM, 2013, pp. 361–364.

[47] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, and C.-S. Chen, "Video aesthetic quality assessment by temporal integration of photo-and motion-based features," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 1944– 1957, 2013.
[48] Cass, J. and Alais, D., "Evidence for two interacting temporal channels in human visual processing," Vision Res. 46(18), 2859–2868 (2006).
[49] Hess, R. F. and Plant, G. T., "Temporal frequency discrimination in human vision: Evidence for an additional mechanism in the low spatial and high temporal frequency region," Vision Res. 25(10), 1493–1500 (1985).
[50] Mandler, M. B. and Makous, W., "A three channel model of temporal perception," 1881–1887 (1984).
[51] Watson, a. B. and Robson, J. G., "Discrimination at threshold: Labelled detectors in human vision," Vision Res. 21(7), 1115–1122 (1981).
[52] Hess, R. F. and Snowden, R. J., "Temporal properties of human visual filters: Number, shapes and spatial covariation," Vision Res. 32(1), 47–59 (1992).
[53] Johnston, a. and Clifford, C. W. G., "A unified account of three apparent motion illusions," Vision Res. 35(8), 1109–1123 (1995).
[54] Adelson, E. H. and Bergen, J. R., "Spatiotemporal energy models for the perception of motion," 284–299 (1985)
[55] Hosu, V. et al., "The Konstanz natural video database (KoNViD-1k)," in 2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017 (2017).
[56] Y. LeCun, B. Boser, J .S. Denker D. Henderson, R. E. Howard, W. Hubbard and L.D. Jackel "Backpropagation applied to handwritten zip code311 recognition," Neural computation 1(4), 541–551 (1989) [doi:101162/neco.1989.1.4.541]

AUTHORS DETAILS

Daniel Oppong Bediako received his Bachelor of Engineering (BEng) in Electrical Electronic Engineering from Accra Institute of Technology, Accra, Ghana, in 2013 and his MS and Ph.D degree in Information and Communication Engineering from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2017 and 2023 respectively. His research interests include image quality assessment and video quality assessment. (Phone: +233-243950731; E-mail: danielbediako@stu.xjtu.edu.cn)