No-Reference Video Quality Assessment Based on Ensemble of Knowledge and Data-Driven Models

Li Su[1(✉)], Pamela Cosman[2], and Qihang Peng[3]

[1] University of Chinese Academy of Sciences, Beijing, China
suli@ucas.ac.cn
[2] University of California at San Diego, San Diego, USA
pcosman@eng.ucsd.edu
[3] University of Electronic Science and Technology of China, Chengdu, China
anniepqh@uestc.edu.cn

Abstract. No-reference (NR) video quality assessment (VQA) aims to evaluate video distortion in line with human visual perception without referring to the corresponding pristine signal. Many methods try to design models using prior knowledge of people's experience. It is challenging due to the underlying complexity of video content, and the relatively limited understanding of the intricate mechanisms of the human visual system. Recently, some learning-based NR-VQA methods were proposed and regarded as data driven methods. However, in many practical scenarios, the labeled data is quite limited which significantly restricts the learning ability. In this paper, we first propose a data-driven model, V-CNN. It adaptively fits spatial and temporal distortion of time-varying video content. By using a shallow neural network, the spatial part runs faster than traditional models. The temporal part is more consistent with human subjective perception by introducing temporal SSIM jitter and hysteresis pooling. We then exploit the complementarity of V-CNN and a knowledge-driven model, VIIDEO. Compared to state-of-the-art full reference, reduced reference and no reference VQA methods, the proposed ensemble model shows a better balance between performance and efficiency with limited training data.

Keywords: No reference video quality assessment · Neural network
Prior knowledge · Spatial and temporal information

1 Introduction

With the rapid development and wide application of digital media devices, the number of video resources is growing at an explosive rate. Video Quality Assessment (VQA) plays an important role in a broad range of applications, e.g., enhancement, reconstruction, compression, communication, display, registration, watermarking and etc., and has drawn increasing attention from researchers in recent years.

Existing VQA methods can be roughly divided into two categories: subjective and objective. Subjective viewing tests are performed according to standard procedures. However, since Mean Opinion Scores (MOS) need to be obtained from a large number

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11296, pp. 231–242, 2019.
https://doi.org/10.1007/978-3-030-05716-9_19

of observers, measuring subjective video quality can be challenging, time-consuming and expensive. Sometimes, trained experts are required for judging.

Therefore, there has been an increasing demand to build intelligent, objective quality measurement models to predict perceived video quality. These models aim to provide similar results to subjective quality assessment, but are based on automatically measured criteria and metrics. According to the availability of the original video signal (generally not compressed), these methods are classified as Full Reference (FR), Reduced Reference (RR) and No-Reference (NR) Methods. FR metrics compute the quality difference between distorted video and lossless reference video. For RR metrics, the reference video is partially available and usually is in the form of a set of extracted features to help evaluate the quality of the distorted video. NR metrics try to assess the quality of a distorted video without any reference to the original one. Recently, NR-VQA is becoming more important because NR metrics have broader applications than FR and RR metrics [1, 2].

Numerous NR-VQA algorithms have been proposed. The majority attempt to predict the quality of videos that suffer specific types of distortion. Caviedes and Oberti [3] compute a set of blocking, blurring, and sharpness features, and other papers measure blocking and packet loss [4], or blockiness, blur and noise [5]. Later work considered blockiness and blurriness on detected regions of interest [6], or measured the distortion of compressed videos using Laplacian pyramid features [7]. In [8], the authors proposed an NR-VQA algorithm that measures spatial distortion between a video block and its motion compensated block in the previous frame, where temporal distortion is computed as a function of the mean of the motion vectors.

The application of those distortion specific methods is restricted because practical distortions are hybrid and complicated. Hence, some distortion non-specific (also called general purpose) NR-VQA methods were put forward recently. Some try to directly predict video quality driven by strong prior knowledge, such as [9], designed according to principles of the human vision system (HVS). The algorithm predicts video quality by modeling subband filter coefficients. Others design learning-based methods, mostly following the approach of first obtaining distortion representation features and then training a regression model. An NR-VQA method was proposed based on natural video statistics in the discrete cosine transform domain by incorporating temporal motion information, then a linear regression model was trained to predict video quality [10]. In [11], the authors proposed a bag-of-words and support vector regression (SVR) model to obtain each frame score, and then a temporal pooling strategy yields the final score for a whole video sequence. In [12], a novel model was based on a 1D convolutional neural network (CNN) and logistic regression. It uses the 3D Shearlet transform to extract features and then puts the features into the CNN and regression sequentially.

There are disadvantages with both the knowledge-driven and learning-based methods. On the knowledge-driven side, video distortions are complicated and people's prior knowledge is limited. The HVS is complex and only partially understood. So, it is difficult to design an algorithm consistent with human perception. Learning-based methods, as a rule, need plenty of data to train a robust model, but existing labeled data is limited and it is expensive to obtain additional labeled video quality data, which restricts the learning ability of these methods.

To tackle these problems, we decide to exploit the complementarity of knowledge-driven methods and learning-based methods. Prediction results of knowledge-driven methods could be more stable because of the model simplicity, whereas learning-based methods could fit the data with better prediction tendency since they use extra data to train models. In this paper, we involve the algorithm in [9] as a representation of knowledge-driven methods. We propose a new learning-based method because existing methods are either too slow or do not learn from the original frames. Our new learning-based method is based on 2D-CNN that learns spatial features from original frames. Then a linear regression is trained to predict video quality incorporating a group of temporal features we devised.

The contributions of the proposed NR-VQA model are summarized as follows:

(1) We exploit the complementarity of knowledge-driven and learning-based methods. The proposed ensemble model achieves a better tradeoff between performance and efficiency with limited training data.
(2) We propose V-CNN, a novel end-to-end learning-based NR-VQA model that adaptively fits distinctive features for universal distortion types.
(3) The proposed V-CNN model is composed of spatial and temporal parts, which benefit the assessment for time-varying video content. The spatial distortion model runs faster than traditional models and fits data well with a shallow neural network. The temporal distortion model is more consistent with human subjective perception by introducing temporal SSIM jitter and hysteresis pooling.

This paper is organized as follows: The two kinds of algorithms are presented in Sect. 2. Experimental results are in Sect. 3 and conclusions in Sect. 4.

2 Algorithm

The framework of the proposed ensemble model is illustrated in Fig. 1. The model is composed of VIIDEO [9], a knowledge-driven method, and our proposed CNN-based algorithm called V-CNN, for the data-driven side.

2.1 Knowledge-Driven Method: VIIEDO

VIIDEO [9] is a representative knowledge-driven algorithm. It is based on the insight that the bandpass filter coefficients of frame differences capture temporal statistical regularities arising from structures such as moving edges. The authors found that such coefficients are more homogeneous for pristine frame differences than for those with distortion. They probe these deviations by analyzing the sample distributions of products of pairs of adjacent coefficients computed along horizontal, vertical and diagonal spatial orientations. The products of neighboring coefficients were shown to be well modeled as following a zero mode asymmetric generalized Gaussian distribution. The model predicts the quality from fine and coarse levels of frame differences.

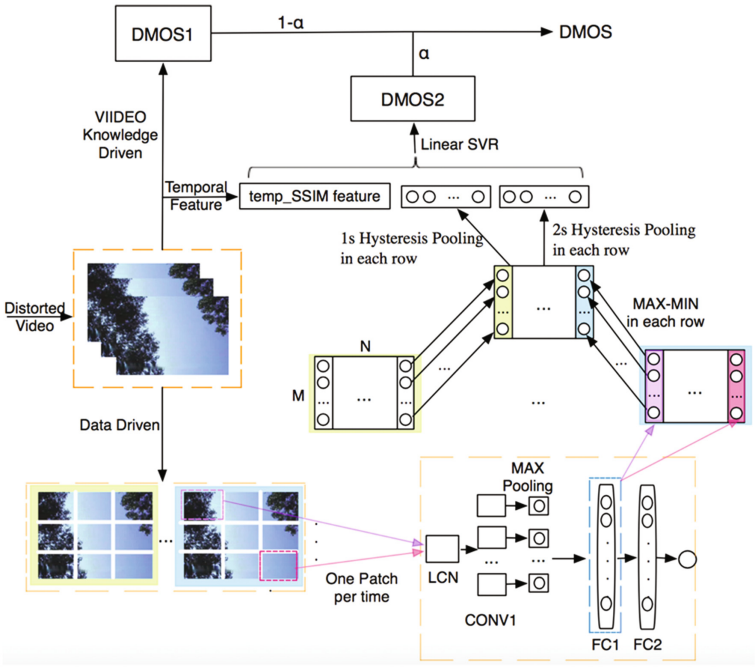


Fig. 1. The framework of our ensemble model.

2.2 Data-Driven Method: V-CNN

In order to adaptively fit the time-varying content of video, the proposed V-CNN algorithm is divided into spatial and temporal parts.

- **Spatial Part**

We reference the network structure of [14], which shows great performance in no-reference image quality assessment (NR-IQA). It has only one convolutional layer (50 kernels of size 7×7) because of speed and because a shallow structure would be stable to fit limited labeled data. Two fully connected layers (FC1 and FC2) both with 1024 nodes are concatenated next. An L1 loss function is used. The input of the network is a 32×32 image patch. Each patch is processed by local contrast normalization to alleviate the saturation problem and make the network robust to illumination and contrast variation. The CNN is pre-trained on the LIVE dataset for image quality assessment (IQA) [15, 16].

When CNN training is complete, we need to organize the patch-level features into video-level features. We select the features of FC1 as the patch-level features. Based on our experiments and those of other papers [11] we first organize patch-level features into frame-level features using the MAX-MIN of local responses, which is beneficial for capturing changes of quality.

- **Temporal Part**

Next, we organize the frame-level features into sequence-level features. Different from conventional linear fusion methods, we propose a temporal pooling strategy to account for the subjective effects.

As discussed in [17], there exists a hysteresis effect in subjective video quality judgment. When a distortion event results in a sharp decrease in video quality, poorer subjective quality scores remain even after the event passes [17]. Hysteresis pooling, shown to be effective for temporal changes of video quality, is used to organize the frame-level features into sequence-level features. With N frames in total, we use $Z(t_c)$ to denote the spatial feature for the t_c^{th} video frame. The pooling feature $a(t_c)$ accounts for the memory effect of the previous T frames. It is a MAX pooling strategy because the worst quality attracts more attention in people's memory. The feature $b(t_c)$ accounts for the propagation effect over the following T frames. To account for the fact that subjects respond strongly to drops in quality, we sort the quality scores in ascending order and combine them using a Gaussian weighting function.

Linear fusion is used to get the weighted feature of the current frame Q_{frame} and the final feature of the whole video Q_{video}. In the following, w_p, w_1, w_2 are empirically-determined parameters controlling the weights.

$$a(t_c) = \begin{cases} Z(t_c), & t_c \leq 1 \\ \max\limits_{t=max(t_c-T,1)}^{t_c}(Z(t)), & t_c \geq 2 \end{cases} \tag{1}$$

$$b(t_c) = w_p \cdot \operatorname*{sort}_{t=t_c}^{min(t_c+T,N)}(Z(t)) \tag{2}$$

$$Q_{frame}(t_c) = w_1 \cdot a(t_c) + w_2 \cdot b(t_c) \tag{3}$$

$$Q_{video} = \frac{1}{N}\sum_{t_c=1}^{N} Score(t_c) \tag{4}$$

Table 1. Performance comparison for introducing different modules in V-CNN

Modules	SROCC	LCC
1s hysteresis pooling	0.587	0.660
2s hysteresis pooling	0.568	0.641
(1s+2s) hysteresis pooling	0.612	0.660
Temp_SSIM	0.260	0.517
1s hysteresis pooling+Temp_SSIM	0.650	0.709
2s hysteresis pooling+Temp_SSIM	0.640	0.697
(1s+2s) hysteresis pooling+Temp_SSIM	**0.671**	**0.713**

Referring to [17], T usually is selected to be two seconds. We believe that the latest memories are more influential. So we add extra one-second hysteresis pooling onto the traditional two-second pooling to emphasize the short-term memory effect. Our experience verified that it further improved the performance, as shown in Table 1.

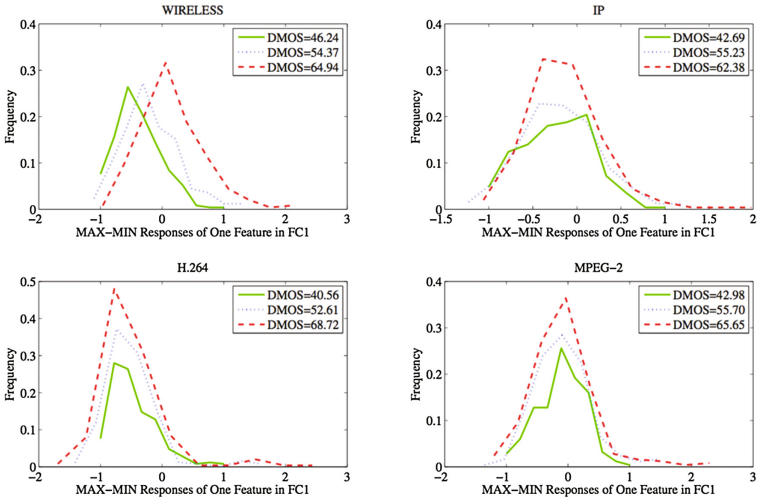


Fig. 2. Example of feature distribution histograms for different types and levels of distortion.

In order to verify the effectiveness of the learned features, we examine distribution histograms for different types and levels of distortions as shown in Fig. 2. We see that the learned features can clearly distinguish the different levels of distortions for universal distortion types.

We also note that local jitter exists in many kinds of temporal distortions. We can model jitter by examining statistics for the motion compensated corresponding blocks across adjacent frames. We divide frames into 16×16 blocks. For each block, a motion vector three-step search algorithm [18] is used to find the reference block. Then we calculate the jitter according to Eq. (5) below.

We devise a group of temporal features based on the idea that the structural similarity index (SSIM) [19] in corresponding blocks across adjacent frames may reflect the temporal quality of a video. The temporal features can be represented as follows:

$$JITTER(t) = Temp\_SSIM\_MAP(t) \tag{5}$$

$$Mean\_M = \frac{1}{N-1}\sum_{t=2}^{N} Mean(JITTER(t)) \tag{6}$$

$$Variance\_M = \frac{1}{N-1}\sum_{t=2}^{N} (Mean(JITTER(t)) - Mean\_M)^2 \tag{7}$$

$$Mean\_V = \frac{1}{N-1}\sum_{t=2}^{N} Variance(JITTER(t)) \tag{8}$$

$$Variance\_V = \frac{1}{N-1}\sum_{t=2}^{N} (Variance(JITTER(t)) - Mean\_V)^2 \tag{9}$$

where $Temp\_SSIM\_MAP(t)$ is a matrix whose entries represent the SSIM value of the corresponding blocks across the t^{th} pair of adjacent frames, N is the total number of video frames, and $Mean()$ and $Variance()$ calculate the expectation and variance of a matrix. To predict video quality, a linear support vector regression (SVR) model is trained using the spatial and temporal features.

2.3 Ensemble of Two Kinds of Methods

Although VIIDEO performs well as a knowledge-driven method, it ignores detail information and its prediction accuracy is not good enough to be used in practice. But its prediction results are stable because of its inherent simplicity. Though there are not enough data for V-CNN as a learning-based method, which leads to unstable prediction results, it still shows good prediction tendencies. Thus, the two kinds of methods have complementary advantages and disadvantages when the dataset is limited.

To aggregate their advantages, we propose an ensemble model composed of the two methods. In our ensemble model, we merge their predicted quality scores:

$$QS_{final} = \alpha \times QS_{V-CNN} + (1-\alpha) \times QS_{VIIDEO} \tag{10}$$

where QS_{final} is the final predicted video quality score, $QS_{V\text{-}CNN}$ and QS_{VIIDEO} are the predicted scores of the proposed V-CNN and VIIDEO respectively, and $\alpha = 0.25$ is an empirically-determined parameter controlling the weight of the two algorithms.

3 Experiments

3.1 Dataset and Evaluation Protocol

Most popular NR-VQA methods such as V-BLIINDS and VIIDEO are tested on the LIVE VQA dataset [20, 21]. For the sake of comparison, we conduct experiments on it as well. The dataset includes 160 videos, with ten uncompressed high-quality videos as reference videos. A set of 150 distorted videos are created from these reference videos (15 distorted videos per reference) using four different distortion types: MPEG-2 compression, H.264 compression, and simulated transmission of H.264 compressed bit-streams through error-prone IP networks and through error-prone wireless networks. The differential mean opinion score (DMOS) in the range from 0 to 100 is used. A higher DMOS denotes worse quality.

Like most VQA research, we employ the linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SROCC) to evaluate performance. For both, a

higher value denotes better performance. There are 10 distinct video contents in the dataset; we used 8 for training and 2 for testing, and there are 45 such combinations. All experiments are repeated 45 times, and the median LCC and SROCC values are presented as final results.

Additionally, since the prediction quality score of VIIDEO is in the range of 0 to 1, and the ground truth score is in the range of 0 to 100, in V-CNN, we first normalize the ground truth score into the range of 0 to 1 as follows:

$$QS_i = (QS_i - QS_{min})/(QS_{max} - QS_{min}) \tag{11}$$

where QS_i denotes the ground truth score of the i^{th} video, and QS_{min} and QS_{max} denote the minimum and maximum scores across all videos.

3.2 Results and Discussion

We first tested on every specific distortion dataset and then on the whole dataset. We compare our algorithm with four FR-VQA methods, one RR-VQA method and two NR-VQA methods: PSNR and SSIM [19] are evaluated for image quality and we use mean pooling across frame scores in our experiment. STMAD [22] and MOVIE [23] are recent VQA methods with top performance. STRRED [24] is a typical RR-VQA method, and V-BLIINDS [10] and VIIDEO [9] are popular NR-VQA methods with state-of-the-art performance.

- **Proposed temporal modules in V-CNN**

As shown in Table 1, by introducing hysteresis pooling and the temporal SSIM jitter module, the proposed V-CNN model fits temporal quality features well and is more consistent with human subjective perception.

Table 2. Median SROCC correlations for different VQA methods on the LIVE database.

Methods		Distortion types				
		Wireless	IP	H.264	MPEG2	Mix
FR	PSNR	0.691	0.600	0.714	0.643	0.677
	SSIM [19]	0.691	0.543	0.881	0.786	0.650
	MOVIE [23]	0.786	**0.771**	0.881	0.905	0.807
	STMAD [22]	**0.810**	**0.771**	**0.952**	**0.929**	**0.834**
RR	STRED [24]	0.762	0.771	0.905	0.905	0.826
NR	V-BLIINDS [10]	0.691	0.600	0.643	0.667	0.735
	VIIDEO [9]	0.548	0.600	0.762	0.571	0.651
	V-CNN	0.690	0.600	0.738	0.738	0.671
	V-CNN+VIIDEO	**0.738**	**0.657**	**0.786**	**0.786**	**0.751**

Table 3. Median LCC correlations for different VQA methods on the LIVE database.

Methods		Distortion types				
		Wireless	IP	H.264	MPEG2	Mix
FR	PSNR	0.798	0.733	0.698	0.696	0.722
	SSIM [19]	0.634	0.726	0.851	0.805	0.625
	MOVIE [23]	**0.920**	0.895	0.919	**0.955**	0.852
	STMAD [22]	0.904	**0.901**	**0.947**	0.942	**0.861**
RR	STRED [24]	0.806	0.816	0.892	0.904	0.725
NR	V-BLIINDS [10]	0.844	0.852	**0.956**	**0.949**	0.790
	VIIDEO [9]	0.740	0.848	0.886	0.872	0.701
	V-CNN	0.808	0.914	0.892	0.871	0.713
	V-CNN+VIIDEO	**0.874**	**0.923**	0.870	0.876	**0.794**

Table 4. Average runtime for different NR-VQA methods on the LIVE database.

Methods	Runtime (s)
STMAD	667.57
V-BLIINDS [10]	709.14
VIIDEO [9]	160.94
V-CNN	175.63
V-CNN+VIIDEO	336.57

- **Proposed NR-VQA model versus state-of-the-art VQA models**

As shown in Tables 2 and 3, both the SROCC and LCC coefficients of our ensemble algorithm are the best among NR-VQA methods and are also better than FR methods PSNR and SSIM on three single distortion subsets and on the whole dataset. Also, the proposed ensemble model runs much faster but achieves comparable performance with all state-of-the-art FR and RR methods, as shown in Table 4.

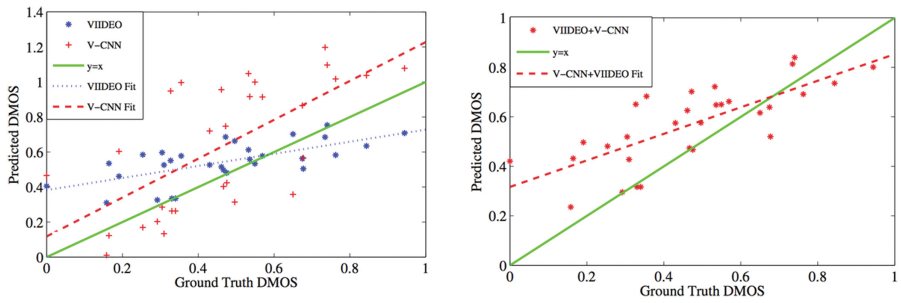
- **V-CNN versus knowledge-driven NR-VQA methods**

From Tables 2 and 3, we see that V-BLIINDS performs best among all knowledge-driven NR-VQA methods. It even outperforms the proposed data-driven method V-CNN. However, the increased running time for V-BLIINDS is large, as shown in Table 4. That is because the feature extraction of V-BLIINDS references more comprehensive prior knowledge, which contributes to final performance but results in higher runtime.

V-CNN is trained on limited training data, which restricts its learning ability. However, V-CNN fits the data better than the knowledge-driven model VIIDEO and costs similar runtime with a shallow neural network. That is to say, V-CNN achieves comparable results with V-BLIINDS with less runtime.

- **Ensemble versus Separate**

As shown in Tables 2 and 3, the ensemble of data-driven model V-CNN and knowledge-driven model VIIDEO improves the performance dramatically compared to the two algorithms run separately. The ensemble algorithm runs much faster than V-BLIINDS as shown in Table 4. Therefore, the ensemble method keeps better balance between performance and efficiency compared to the state-of-the-art.



(a) Separate (b) Ensemble

Fig. 3. The regression tendency of separate two models and the ensemble model on LIVE.

We have conducted another experiment to verify the complementarity of the two methods. As shown in Fig. 3(a), the prediction results of VIIDEO are more stable. It demonstrates the simplicity and universal adaptability. However, V-CNN has better prediction tendency. Figure 3(b) indicates the ensemble method aggregates their advantages and gains better performance. These findings support our ensemble motivation. Figure 4 shows the three algorithms' distributions of absolute residual values between predicted DMOS and ground truth DMOS. The residual values of the ensemble method are smaller. It also indicates that the complementarity of the two methods improves the performance of the ensemble model.

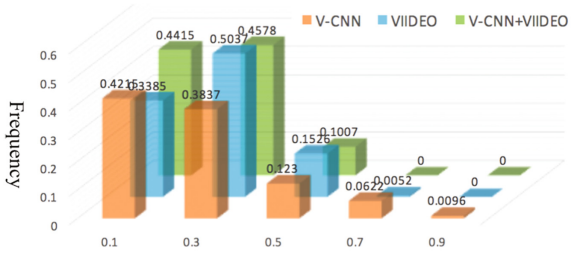


Fig. 4. Distribution of absolute residual value between the predicted DMOS and the ground truth DMOS.

4 Conclusion

In this paper, we propose V-CNN, a learning-based VQA model, and exploited the complementarity of V-CNN and the well-known knowledge-driven model VIIDEO. Experiments show that V-CNN achieves comparable performance and runs fast. The proposed ensemble of two models further improves the performance when there is limited labeled training data. It also keeps a better balance between performance and efficiency compared to state-of-the-art approaches.

Acknowledgement. This work was supported in part by the China Scholarship Council Program and by the National Natural Sciences Foundation of China: 61472389, 61332016 and 61301154.

References

1. Fang, Y., Yan, J., Li, L., Wu, J., Lin, W.: No reference quality assessment for screen content images with both local and global feature representation. IEEE Trans. Image Process. **27**(4), 1600–1610 (2018)
2. Wu, Q., Li, H., Meng, F., Ngan, K.N.: Generic proposal evaluator: a lazy learning strategy toward blind proposal quality assessment. IEEE Trans. Intell. Transp. Syst. **19**(1), 306–319 (2018)
3. Caviedes, J.E., Oberti, F.: No-reference quality metric for degraded and enhanced video. In: Visual Communications and Image Processing, International Society for Optics and Photonics, pp. 621–632 (2003)
4. Babu, R.V., Bopardikar, A.S., Perkis, A., Hillestad, O.I.: No-reference metrics for video streaming applications. In: International Workshop on Packet Video, pp. 10–11 (2004)
5. Farias, M.C., Mitra, S.K.: No-reference video quality metric based on artifact measurements. In: IEEE International Conference on Image Processing, vol. 3, pp. III–141 (2005)
6. Lin, X., Tian, X., Chen, Y.: No-reference video quality assessment based on region of interest. In: 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), pp. 1924–1927 (2012)
7. Zhu, K., Hirakawa, K., Asari, V., Saupe, D.: A no-reference video quality assessment based on Laplacian pyramids. In: 20th IEEE International Conference on Image Processing (ICIP), pp. 49–53 (2013)
8. Yang, F., Wan, S., Chang, Y., Wu, H.R.: A novel objective no-reference metric for digital video quality assessment. IEEE Signal Process. Lett. **12**(10), 685–688 (2005)
9. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Trans. Image Process. **25**(1), 289–300 (2016)
10. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Trans. Image Process. **23**(3), 1352–1365 (2014)
11. Xu, J., Ye, P., Liu, Y., Doermann, D.: No-reference video quality assessment via feature learning. In: IEEE International Conference on Image Processing (ICIP), pp. 491–495 (2014)
12. Li, Y., et al.: No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Trans. Circ. Syst. Video Technol. **26**(6), 1044–1057 (2016)

13. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2013)
14. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1733–1740 (2014)
15. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
16. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
17. Seshadrinathan, K., Bovik, A.C.: Temporal hysteresis model of time varying subjective video quality. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1153–1156 (2011)
18. Li, R., Zeng, B., Liou, M.L.: A new three-step search algorithm for block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **4**(4), 438–442 (1994)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: The SSIM index for image quality assessment. MATLAB implementation, vol. 23, p. 66 (2003). http://www.cns.nyu.edu/lcv/ssim
20. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Trans. Image Process. **19**(6), 1427–1441 (2010)
21. Seshadrinathana, K., Soundararajanb, R., Bovik, A.C., Cormack, L.K.: A subjective study to evaluate video quality assessment algorithms. In: IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, p. 75270H (2010)
22. Vu, P.V., Vu, C.T., Chandler, D.M.: A spatiotemporal most-apparent-distortion model for video quality assessment. In: 18th IEEE International Conference on Image Processing (ICIP), pp. 2505–2508 (2011)
23. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. Image Process. **19**(2), 335–350 (2010)
24. Soundararajan, R., Bovik, A.C.: Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Trans. Circ. Syst. Video Technol. **23**(4), 684–694 (2013)