2025 7th International Conference on Pattern Recognition and Image Analysis (IPRIA), 24–25 September, 2025,
Islamic Azad University–Lahijan Branch, Lahijan, Iran

The Role of Fine-Tuned Feature Map Correlation in Video Quality Assessment

Erfan Asadi
Department of Electrical and Computer Engineering
Faculty of Engineering
Kharazmi University
Tehran, Iran
erfanasadi@khu.ac.ir

Parmida Pourmatin
Department of Electrical and Computer Engineering
Faculty of Engineering
Kharazmi University
Tehran, Iran
parmidapourmatin@khu.ac.ir

Azadeh Mansouri
Department of Electrical and Computer Engineering
Faculty of Engineering
Kharazmi University
Tehran, Iran
a_mansouri@khu.ac.ir

Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran ***Abstract*—Natural videos and user-generated content (UGC) illustrate complex distortions that are difficult to model. As a result, existing Video Quality Assessment (VQA) methods mostly have problem to achieve high performance on these videos. In this paper, we propose a method that utilizes correlation-based features fine-tuned on an image quality assessment dataset to enhance VQA performance. These low-level features are combined with high-level features extracted from the final layers of the network, providing a rich representation of spatial degradations. For temporal pooling, a simple max-pooling operation is applied. Experimental results on two widely used UGC datasets, LIVE-VQA and KoNViD-1k, demonstrate strong performance while maintaining low computational complexity.The implementation of our method is available on GitHub at GitHub Repository.**

***Index Terms*—fine-tuned feature maps, gram matrix, no-reference video quality assessment, deep convolutional neural networks, SVR.**

I. Introduction

Spreading of the multimedia technology has extensively promoted digital visual content usage in all aspects of life. Digital television, video surveillance, and video conference applications have grown so fast that individuals' methods of sharing and receiving content have been transformed [1]. The intersection of social media platforms and portable mobile technology, such as YouTube, Facebook, and TikTok, has rendered video-sharing more mainstream, with YouTube alone hosting 2.6 billion users by 2022 [2]. At the same time, the COVID-19 pandemic considerably increased the use of online media, showcasing the widespread impact that video content has on everyday digital communication [3].

Though it is easy to share and even view video content, maintaining its quality is a concern in terms of compression, network variations, and other deformations that have a tendency to degrade the visual content [1], [4]. Video quality evaluation thus becomes inevitable for improving user experience and confidence in video delivery services [4]. The conventional ways of Video Quality Assessment (VQA) include subjective and objective methodologies. Subjective VQA, though precise, is not feasible for real-time large-scale applications as it utilizes human evaluators [4], [5]. Objective VQA, however, tries to estimate perceived quality without such limitations.

Objective techniques for Video Quality Assessment (VQA) can be categorized into three types: full-reference (FR-VQA), reduced-reference (RR-VQA), and no-reference (NR-VQA) and need different amounts of reference information [5]. Specifically, NR-VQA, or blind quality assessment, has become increasingly significant due to its relevance in environments where a perfect reference is not achievable [1], [5]. Recent advancements in blind assessment approaches tap into the inherent properties of video material, combining Natural Scene Statistics (NSS) with sophisticated machine learning algorithms to effectively estimate video quality from spatio-temporal characteristics [6].

However, current NR-VQA models, predominantly built on top of pre-trained deep Convolutional Neural Networks (CNNs), have been shown to focus mostly on broad patterns at the expense of detail regarding spatial and texture-specific ones [1]. To rectify this limitation, new frame-level feature extraction methods have been explored in recent studies, thus advancing the capacity of NR-VQA models to discern minor quality fluctuations and become more generalizable to other datasets [1]. This paper presents a new algorithm and highlights the need for presenting such new features, thus making a valuable contribution to the evaluation of user-generated content with real distortions. By researching and enhancing such techniques, we open up the possibility of developing more stable and efficient systems for video quality evaluation, which are adaptable to the increasing needs of digital media consumption.

II. Related Works

Recent advances in no-reference video quality assessment (VQA) have increasingly focused on the incorporation of deep learning methods to address the diverse and complex nature of modern video content. Unlike previous methods that relied mainly on manually crafted statistical features, the advent of user-generated content, which is typified by multiple distortions, has prompted the shift towards more sophisticated techniques that utilize deep feature integration.

At the heart of this development process is the approach that Xu et al. put forward in 2014 and presented V-CORNIA. The approach makes use of unsupervised learning coupled with spatial and temporal feature extraction through modifications of the CORNIA framework and employs max-min pooling methods [7]. Merging these features with support vector regression (SVR) enables precise mapping onto quality scores, effectively encapsulating temporal dynamics.

Varga's framework in 2019 employed CNNs in LSTM models to exploit spatial-temporal dependencies, culminating in a comprehensive no-reference VQA system. In this method, complex video artifacts are elegantly dealt with through temporal pooling and regression strategies, echoing the strength of deep feature extraction [8].

Newer approaches like VSFA, also proposed around the same time, leverage pre-trained CNNs for content-aware features, while GRUs facilitate temporal modeling to manage long-term dependencies of video sequences [11], [12]. Such approaches, inspired by human perception models, strive to align machine ratings with human qualitative ratings, providing greater granularity in quality prediction.

Bakhtiari and Mansouri's work in 2022 highlights the intricate interdependencies among deep feature maps, with CNN architectures particularly designed for video frame degradation identification. Their novel fusion of pooling techniques proves the adaptability necessary for addressing the huge video quality variability space [9].

In 2023, Bakhtiari proposed a new approach that is based on the correlation between feature maps that already pre-exist within pre-trained networks, FMC-VQA. The work emphasizes employing Gram matrices to find correlations in mid-layer feature maps, thereby adding richness to the quality assessment technique. Through investigations of structural attributes such as texture and curvature and the application of Gram matrices as high-level quality features, authors report notable advancements on various datasets, showing improvement in generalizability and efficacy. The paper employs the EfficientNet B4 architecture that proves to be notably superior to conventional approaches in rank correlation coefficients [10].

Also, towards the goal of closing the discrepancy between application and model predictions, approaches that entail established datasets such as LIVE and CSIQ ensure more comprehensive validation environments. The environments allow the development of models that can more accurately capture real-world quality assessment scenarios [6], [11]. Entwining approaches such as PaQ-2-PiQ enhance the precision of quality rating and thereby ensure tests remain in synchrony with evolving video production and consumption scenarios, which change rapidly.

III. Proposed Method

The rapid growth in video consumption has underscored the need for reliable No Reference Video Quality Assessment (NR-VQA) models. Traditional methods, like the Bakhtiyari approach, often struggle to accurately detect spatial distortions and effectively combine spatial and temporal features. To overcome these limitations, we propose a new NR-VQA model that enhances the extraction and integration of spatial features, resulting in a more accurate overall video quality assessment. The flow of our proposed method, from video input to the computation of the Mean Opinion Score (MOS), is clearly depicted in Figure 1. This flowchart emphasizes key stages such as frame extraction, feature processing, and the final prediction step, providing a systematic overview of the model's workflow.

A. Fine-Tuning InceptionV3 for Spatial Feature Extraction

The first step in our approach is to fine-tune the InceptionV3 network using the TID Image Quality Database. This step is key to improving the model's ability to detect subtle spatial distortions that can appear in video frames. By fine-tuning, we adjust InceptionV3's weights, allowing it to more accurately capture quality-related features, leading to a more refined and precise extraction of spatial features that are crucial for assessing video quality.

B. Spatial Feature Extraction from Video Frames

After fine-tuning, InceptionV3 is used to extract spatial features from each video frame. We take advantage of intermediate layers of the network to capture various levels of abstraction which are essential to accurately assess video quality. To further refine the representation, we compute the Gram matrix across these layers, capturing texture information that aligns with perceptual quality. In addition, features of the avgpool layer are included to form a comprehensive spatial feature vector for each frame, ensuring a rich and detailed representation of the spatial characteristics of the video.

C. Spatial-Temporal Feature Synthesis

To capture the temporal dynamics of the video, we combine spatial and temporal characteristics by applying both average pooling and max pooling across the spatial feature vectors of all frames. This dual pooling technique effectively captures both the average and extreme variations in spatial quality over time, offering a well-rounded view of video quality. The pooled features are then concatenated, enriching the feature set with subtle and significant quality changes observed throughout the video sequence, ensuring a more comprehensive representation of its overall quality.

D. Support Vector Regressor for Quality Prediction

The concatenated spatial-temporal features are then fed into a Support Vector Regressor (SVR), which is trained to predict the Mean Opinion Score (MOS) for video quality. The SVR uses the rich, combined feature representation to map these features to subjective quality scores, resulting in a robust prediction model. This model is capable of handling a wide range of video content, ensuring accurate quality assessments in various types of videos.

E. Implementation Details

Our proposed NR-VQA model is built on the InceptionV3 architecture, which is fine-tuned using the TID Image Quality Database to enhance its ability to extract spatial features essential for video quality assessment. This fine-tuning is carried out in PyTorch, a popular deep learning framework known for its dynamic computation graph and flexibility in model customization. PyTorch's strong support for automatic differentiation and GPU acceleration makes it ideal for efficient training and experimentation.

For the video quality prediction task, we use a Support Vector Regressor (SVR) implemented through the scikit-learn library. This combination ensures computational efficiency while maintaining high accuracy in predicting video quality, making the model adaptable to various video datasets. The implementation is designed to be both scalable and portable, enabling easy adjustments and improvements for different video quality assessment scenarios.

F. Evaluation/Validation

We validate our NR-VQA model using two widely recognized user-generated content (UGC) video datasets: KoNVid-1k and LiveVQA. These datasets cover a broad range of video types, ensuring a thorough evaluation of our model's performance across diverse content.

To measure performance, we use two key metrics: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC). These metrics are essential for assessing how closely our model's predictions align with subjective human evaluations of video quality.

In addition to standard validation, we conduct cross-dataset testing to assess the model's generalizability. We train the model on one dataset (e.g., KoNVid-1k) and test it on the other (e.g., LiveVQA), and vice versa. This approach helps demonstrate our model's robustness and its ability to perform well across different video datasets with varying contexts and distributions.

The results show that our proposed method consistently achieves high PLCC and SROCC values in both data sets, outperforming existing NR-VQA models. This underscores the model's adaptability and effectiveness in delivering accurate quality predictions, regardless of the video content.

In summary, the proposed method makes a significant contribution to the field of No Reference Video Quality Assessment by improving spatial feature extraction and integrating a dual pooling strategy. These advancements result in notable gains in both the accuracy and robustness of video quality predictions. Looking to the future, further research will aim to expand the model's applicability across a wider variety of video content scenarios. Additionally, there are plans to incorporate additional contextual factors that could affect perceived video quality, further enhancing the model's ability to predict quality in diverse contexts.

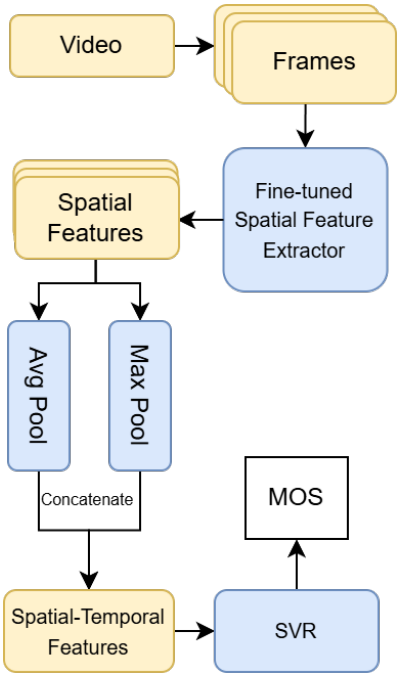

Fig. 1. From Video to MOS: The flowchart of the proposed No Reference Video Quality Assessment method. This diagram outlines the transformation process starting from video input to the final Mean Opinion Score prediction, detailing each crucial stage of frame conversion, feature extraction, and spatial-temporal feature synthesis.

Experimental Results

The median SROCC and PLCC values over 100 test times (20 rounds of 5-fold cross-validation) are presented in Tables I and II for KONVID-1K and LIVE-VQC, respectively. The best-performing architectures are highlighted in bold face. The results clearly show that the Mixed-5b layer with a Linear Kernel outperforms other structures.

Figures 2 and 3 illustrate the scatter plots of the predicted quality scores versus subjective scores using the best model on the KonVid-1k and LIVE-VQC datasets, respectively. The results clearly demonstrate that the proposed method achieves a strong correlation with subjective scores, particularly for KonVid-1k.

We conducted cross-database validation by selecting each video as the training set and evaluating performance on the other. The SROCC scores for both databases are presented in Table 3. The top 3 results are indicated in bold face. The results clearly show that the proposed method achieves acceptable performance. Moreover, the method performs well with a low-complexity approach that simply calculates feature map correlations using a fine-tuned network. Overall, the presented method delivers competitive results while maintaining efficiency compared to more complex approaches.

Conclusion

The proposed method exploits the correlation between deep features as a frame-level features for UGC quality assessment. These features capture structural information at various

TABLE I
CORRELATION OF PREDICTED MOS WITH GROUND TRUTH MOS - INCEPTIONV3 - KONVID-1K

Layers			Kernel	SROCC	PLCC
avgpool	Mixed_5b	Mixed_5c			
✓	✓	✓	linear	0.7983	0.8067
			rbf	0.7924	0.7822
✓	✓	-	linear	0.8145	0.8076
			rbf	0.7895	0.7797
✓	-	✓	linear	0.7987	0.7964
			rbf	0.7679	0.7769
✓	-	-	linear	0.5841	0.5737
			rbf	0.6433	0.6498
-	✓	-	**linear**	**0.8179**	**0.8118**
			rbf	0.7838	0.7727
-	✓	✓	linear	0.8139	0.8052
			rbf	0.7798	0.7846
-	-	✓	linear	0.7809	0.7780
			rbf	0.7939	0.7904

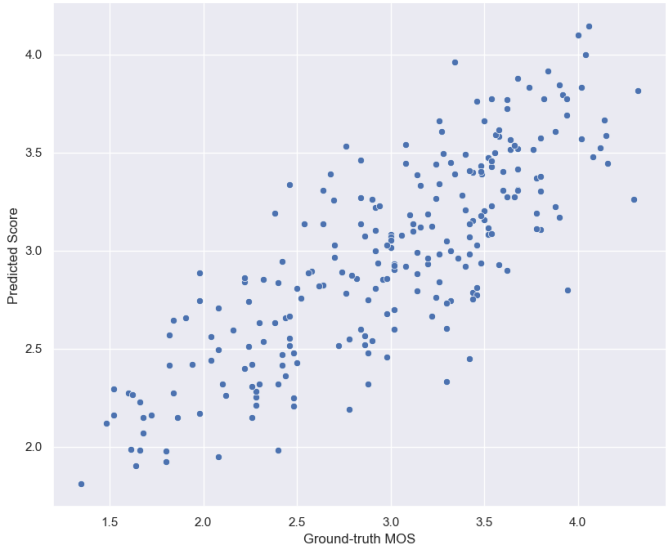

Fig. 2. Scatter plot of subjective MOS vs objective scores on the KoNViD-1k dataset

granularities. The extracted features are derived from a network fine-tuned on TID-2013, a well-known image quality assessment dataset; then combined with high-level features extracted from the final layers of the network and temporally pooled. The impact of fine-tuning on the image dataset is reflected in spatial quality-aware feature vectors. Experimental results demonstrate this effect in both single and cross-dataset validations.

TABLE II
CORRELATION OF PREDICTED MOS WITH GROUND TRUTH MOS - INCEPTIONV3 - LIVEVQC

Layers			Kernel	SROCC	PLCC
avgpool	Mixed_5b	Mixed_5c			
✓	✓	-	linear	0.7842	0.7923
			rbf	0.7672	0.7918
✓	-	✓	linear	0.7923	0.8139
			rbf	0.7259	0.7690
✓	-	-	linear	0.5624	0.6241
			rbf	0.5274	0.6001
-	✓	-	**linear**	**0.7939**	**0.8125**
			rbf	0.7867	0.7626
-	✓	✓	linear	0.7704	0.8027
			rbf	0.7690	0.7768
-	-	✓	linear	0.7572	0.8007
			rbf	0.7607	0.8121
✓	✓	✓	linear	0.7868	0.7910
			rbf	0.7611	0.8098

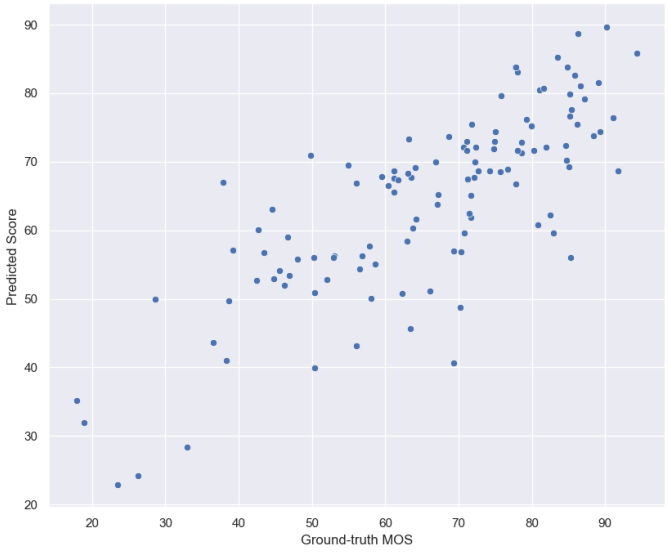

Fig. 3. Scatter plot of subjective MOS vs objective scores on the LiveVQC dataset

REFERENCES

[1] D. Li, T. Jiang, and M. Jiang, "Recent Advances and Challenges in Video Quality Assessment," *ZTE Commun.*, 2019. [Online]. Available: https://doi.org/10.12142/ZTECOM.201901002

[2] "10 YouTube Statistics That You Need to Know in 2022," [Online]. Available: https://www.oberlo.com/blog/youtube-statistics (accessed Jul. 23, 2022).

[3] "Facebook Video Stats to Know in 2022 — 99firms," [Online]. Available: https://99firms.com/blog/facebook-video-statistics/ (accessed Jul. 23, 2022).

[4] Nidhi and N. Aggarwal, "A review on Video Quality Assessment," in *Proc. Recent Adv. Eng. Comput. Sci. (RAECS)*, Mar. 2014, pp. 1–6. doi: 10.1109/RAECS.2014.6799645.

[5] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011, doi: 10.1109/TIP.2011.2147325.

[6] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 491–495. doi: 10.1109/ICIP.2014.7025098.

TABLE III
COMPARATIVE ANALYSIS OF VIDEO QUALITY ASSESSMENT MODELS TRAINED ON LIVEVQC AND KONVID-1K AND TESTED ON EACH OTHER USING PLCC AND SROCC METRICS

Train on	LiveVQC		KonVid-1K	
Test on	KonVid-1k		LiveVQC	
Method	PLCC	SROCC	PLCC	SROCC
CNN-TLVQM (2020, MM)	0.631	0.642	0.752	0.713
VIDEVAL (2021, TIP)	0.621	0.625	**0.841**	0.669
DisCoVQA (2022, arxiv)	**0.785**	**0.754**	0.787	**0.737**
GST-VQA (2021, TCSVT)	**0.7074**	**0.685**	0.777	**0.732**
MDTVSFA (2021, JCV)	**0.711**	0.645	0.816	0.716
Bakhtiar (2022, DFMP)	0.647	**0.679**	**0.823**	**0.756**
Proposed Method	0.6715	0.6698	**0.821**	0.729

[7] J. Xu, et al., "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Process.*, 2014.

[8] D. Varga and T. Szirányi, "No-reference video quality assessment via pretrained CNN and LSTM networks," *Signal, Image and Video Process.*, 2019.

[9] A. H. Bakhtiari and A. Mansouri, "No-Reference Video Quality Assessment by Deep Feature Maps Relations," in *Proc. 12th Int. Conf. Comput. Know. Eng. (ICCKE)*, 2022.

[10] A. H. Bakhtiari and A. Mansouri, "Feature Maps Correlation-based Video Quality Assessment," *Multimedia Tools Appl.*, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-18068-w

[11] P. Chen, et al., "RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment," in *Proc. ACM Int. Conf. Multimedia*, 2020.

[12] Y. Wang, et al., "Rich features for perceptual quality assessment of UGC videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.