

Paper Type: Original Article



## Deep Attributes and Decisions Fusion for No-Reference Video Quality Analysis

Adil Baig\*

University of Agriculture, Pakistan; adil.baig7862@gmail.com.

Citation:



Baig, A. (2023). Deep attributes and decisions fusion for no-reference video quality analysis. *Big data and computing visions*, 3(3), 91-103.

Received: 11/03/2023

Reviewed: 12/04/2023

Revised: 09/05/2023


Accept: 20/06/2023

### Abstract

Video Quality Assessment (VQA) is a critical component of various technologies, including automated video broadcasting through displaying technologies. Moreover, determining visual quality necessitates a balanced examination of visual features and functionality. Previous research has also shown that features derived from pre-trained models of Convolutional Neural Networks (CNNs) are extremely useful in various image analysis and computer vision activities. Based on characteristics collected from pre-trained models of deep neural networks, transfer learning, periodic pooling, and regression, we created a unique architecture for No Reference Video Quality Assessment (NR-VQA) in this research. We were able to get results by solely employing dynamically pooled deep features and avoiding the use of manually produced features. This study describes a novel, deep learning-based strategy for NR-VQA that uses several pre-trained deep neural networks to characterize probable image and video distortions across parallel. A set of pre-trained CNNs extract spatially pooling and intensity-adjusted video-level feature representations, which are then individually mapped onto subjective peer assessments. Ultimately, the perceived quality of a video series is calculated by combining the quality standards from the various regressors. Numerous researches demonstrate that the suggested approach on two large baseline video quality analysis datasets with realistic aberrations sets a new state-of-the-art. Furthermore, the findings show that combining the decisions of different deep networks can greatly improve NR-VQA.

**Keywords:** Video quality assessment, No reference video quality assessment, Deep neural networks.

## 1 | Introduction

 Licensee Big Data and Computing Visions. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

In the research, determining the content of video streams has become a hot and essential area of research. Before being exhibited, digital videos go through various operations, such as compressing or transfer [1]. Furthermore, each procedure affects the video; in most situations, it produces some form of artifact or noise. The perceived quality of the digital video is degraded by aberrations that can be blurring, geometrical distortion, or blackness artifacts from compressing techniques. Video Quality Assessment (VQA) is classified into two categories in the literature: positivist and interpretivist. Subjective VQA is concerned with gathering quality evaluations from a group of people employing a series of films. The assays were performed in a laboratory [2] or through an internet crowd-sourcing procedure [3]. To represent the detection accuracy of each studied image

 Corresponding Author: [adil.baig7862@gmail.com](mailto:adil.baig7862@gmail.com)

 <https://doi.org/10.22105/bdcv.2023.415895.1165>

sequence, the quality evaluations from observers are combined within one multitude of Median Opinion Scores (MOS).

Furthermore, subjective VQA addresses a variety of areas of VQA, including the identification of test video sequencing, grading scales, time intervals for video delivery for human participants, seeing circumstances, and human participation screening. Consequently, subjectively, VQA generates benchmark datasets [4]–[6] containing video frames and associated MOS values. Multiple objective VQA techniques try to create computational equations for reliably measuring the detection accuracy of video sequences using these datasets heavily as learning or assessment data. Deep learning has historically conquered the fields of object recognition, image analysis, and video analysis. Furthermore, this tendency has a significant impact on the area of NR-VQA. The particular contributions of this work are a fresh, novel deep learning-based method for NR-VQA that uses a collection of simultaneous pre-trained deep neural networks that classify probable image/video abnormalities in both directions.

Further precisely, video-level deep extracted features are constructed from a series of pre-trained Convolutional Neural Networks (CNNs) and transformed onto subjective quality monitoring employing learned linear regression that is spatially pooling and sensitivity scored. Ultimately, the importance of the input video sequences is calculated by fusing the quality ratings from the various regressors. We show that combining the decisions of different deep architectures improves the effectiveness of the NR-VQA substantially. Experimental results with authentic distortion are conducted on two major standard VQA datasets.

### 3 | Problem Statement

Our thesis' main objective is to use contour let transform to accomplish No-Reference Video Quality Assessment (NR-VQA). An image representation will become effective when using unstructured and structured transformations. To describe two-dimensional data or functions that video CT can quickly construct algorithms. Video representation can also be incorporated into CT, including multiresolution, localized, critical samples, directivity, and anisotropic. Compared to FR and RR standard video quality evaluations, NR VQA is less expensive. A technique that can predict video quality in a frame or block component from a specified video quality labeling for regression modeling and represent it in the quality assessment of the full video is necessary to enhance quality assessment. An NR-VQA approach using feature learning is presented to estimate the quality rating frame-by-frame.

### 2 | Literature Review

NR-VQA is a difficult process due to the intricacy of Human Visual System (HVS). As a result, the research on NR-VQA contains a large number of studies and publications. The 3 major categories of methods identified in the research are bitstream-based, pixel-based, and hybrid systems. Bitstream-based approaches, in particular, examine video frame headers and decoded packets to determine the detection accuracy of digital videos. The overall Quality for Networking Video via Preliminary Assessment (QANV-PA) approach represents this category well. The authors recovered the first five video frame layer characteristics, namely compression parameters, frames displaying length, frequency of lost packet, frame category, and bitrate [7]. An aggregating approach of frame-level variables was also developed to define perception video.

The characteristics collected using pre-trained models CNN have been demonstrated to be rich and efficient for various computer vision and device learning applications, including content-based information retrieval, NR picture quality evaluation, and clinical image categorization. The significant element and novelty of this research are that we obtain viable NR-VQA responses employing only feature information taken from pre-trained models CNNs (inception-V3 and inception-ResNet-V2), rather than manually picked characteristics. Frame-level deep elements are retrieved from every video frame with a pre-trained model CNN for a certain video series that must be assessed. The frame-level

attributes are then spatially pooled to create a video-level feature representation describing the timeline. Additionally, unlike other publicly available datasets, our design was learned using the published recently Konstanz Natural Video Qualities Data (KoNViD-1k) [8], which includes video segments with true distortion instead of fake distortions. In addition, KoNViD-1k has more films (1200 sequences) than any other available public resource, allowing us to build a deep, temporally pooled framework.

On the other hand, based their theory on three variables: quantized value, bit position, and movement. During research, a packet-layer method was used to assess the perceived qualities of transmission control Internet/Protocol Television Videos (IPTVs). The researchers examined video network packets and retrieved quality-aware characteristics, including bit and packet failure rates. Bit stream-based approaches function well enough in network video surveillance applications like teleconferencing and IPTV but can't be used broadly.

Pixel-based NR-VQA algorithms use the raw video signal as inputs for process improvement. Many Natural Scenario Statistics (NSSs) techniques are quite popular in the research. The basic premise of NSS is that natural photos and videos have patterns in the data that are altered when noise is present. In the research, the Discrete Cosine Transform (DCT) area is widely used to assess divergence from "natural" statistical data [9].

DCT coefficients, for example, have been used to fit various probability distribution functions on these. The dimensions of these PDFs are determined using maximal probability and then used to assess local error. A perceptive Spatiotemporal weighted system was then used to measure total perceived quality. In comparison, the disparity of sequential video frames was first calculated, and then local block-based DCT was added to the differential pictures. The DCT coefficients were then modeled using a modified Generalized Gaussian Distribution (GGD), with the GGD's coefficients serving as quality-aware characteristics. Furthermore, using a Support Vector Regression (SVR) model, those quality-aware characteristics were merged with movement cohesiveness matrices and projected onto quality ratings [9].

Using Three-Dimensional (3D)-DCT for extracting the features, the video content was split into chunks of varied sizes related to spatial and movement activities assessment, identical to the old research. On the other hand, retrieved video frame level characteristics from every video frame. Particularly, DCT was used to build 6 characteristic mappings for each video frame. Next, 5 performance-aware characteristics were selected from the extracted features, aggregated, and combined sequentially to produce video-level relevant features that were then transformed into quality ratings using a neural network. The authors refined this approach. This approach integrates frame classification performance. Other transformation domains, including transform domains, are common in the research. Another area of research compiled quality-aware feature selection by extracting various optical flow characteristics. For example, anomalies in the optical movement were identified at both the picture patch and video frame levels [10].

Intra-patch and inter-patch level abnormalities were detected, and the association among subsequent frames was merged. The magnitude change among two successive images in the consecutive frames was evaluated at the frame levels. The collected features were mapped onto quality ratings with a training SVR, identical to previously discussed approaches. In contrast, extracted features were created by combining spatial data, including contrast and colorfulness, with feature descriptors generated from optical streaming.

Deep learning approaches have suddenly gained a lot of traction in pixel-based algorithms. Furthermore, related topics such as stereoscopic and omnidirectional picture quality evaluation, image super magnification, and stereoscopic VQA have given deep learning a lot of interest. For example, a CNN was constructed from scratch on 3D shearlet transformation parameters retrieved from video frames for subjective VQA. In comparison, to construct quality-aware characteristic vectors for a video sequence merged with hand-crafted and deep characteristics. The generated vector was then regressed onto quality ratings using a frame-to-video component aggregation approach. Image quality parameters, including sharpness, graininess, brightness, and color saturation, were predicted using deep features taken from pre-

trained models CNNs. Frame-level quality ratings were calculated depending on all these quality criteria. The researchers refined the previously reported approach by integrating a sampling method that chooses relevant video frames to remove temporal repetition in video series [11].

Bit stream-based and pixel-based techniques are combined in hybrid approaches. For example, a spatiotemporal feature vector was merged with the mean bit rate and packet loss proportion. The researchers use a non-linear regression approach to estimate the detection accuracy of films sent across the worldwide mobile communication approach by integrating sender bitrate, block error rate, and median burst size. Likewise, video quality via IP networks was tested.

### 3 | Related Work

As previously stated, NR approaches require an input signal and no knowledge of the reference signal. Early noise reduction algorithms were mostly concentrated on distortion-specific techniques. As a result, a technique dependent on the median squared difference among frames was created to measure jerkiness. In comparison, a neural network was trained to simulate the influence of jerkiness on video quality. The error estimation was based on the DCT coefficients for data inside an H.264-specific technique. The movement vectors were collected from the data stream, and subjective quality ratings were calculated using the error estimations. Likewise, they suggested an H.264-specific technique but used DCT coefficients to retrieve the first frame-level information. Furthermore, video-level characteristics were constructed by averaging frame-level features (temporal pooling), and subjective performance ratings were forecasted using a trained neural network. In research, algorithms were also created to evaluate blocking artifacts in distorted videos [12].

The focus of subsequent research was on general-purpose algorithms. NSS were used to construct a successful and extensively used feature extraction method, assuming that natural visual signals contain patterns in the data modified by distortion. The video BLind Image Integrity Notator using DCT Statistics (BLINDSs) BLINDS technique was created using a feature of the author's No-Reference Image Quality Assessment (NR-IQA) approach called BLINDS. The collected characteristics are then used to train an SVR using Video BLINDS, which uses a spatiotemporal framework built from the natural image characteristics of the DCT coefficients.

In generic, general-purpose, NR-IQA approaches that do not need some previous knowledge of deformation classes assume that the loss of "naturalness" is a valuable indication for quality evaluation. NSS methods depend on handmade features extracted throughout the spatial and frequency domains. The local spatial normalization brightness coefficients have been used to analyze NSS characteristics. The findings demonstrate that by combining the gradient of the image characteristics, the combined statistic may achieve satisfactory efficiency for the NR-IQA problem [13].

The research proposed a general-purpose NR-IQA measure based on architectural data and gradient intensity, which are significantly associated with human perceptions. Distortion Identification-based image Verity and Integrity Evaluation (DIVINE) is a two-stage technique that requires distortion detection and SVR to provide excellent ratings for corrupted image features. Singular Value Decomposition (SVD) was used to quantify the strength and make in images, and the qualitative prediction was posed as a logistic issue to estimate image rating employing SVR. The Gaussian Process (GP) has been used to estimate the performance of transperineal ultrasound images rather than employing SVR for grade analysis depending on a one-class regression model. Therefore, generating a binary labeling may not have been sufficient for assessing video quality in a naturalistic environment. The abbreviation GP is sometimes used. To develop GP kernels for NRIQA, a deep belief network with a non-linear activation regression model is used. The GP is used to develop an uncertainty-aware analyzer. Kernel density prediction determines the cognitively coherent peers of a testing image. In contrast to GP-based approaches, the Gaussian distribution used in their approach is specialized towards global qualitative characteristic regularization, allowing the dispersion to predict the grade accurately.

Another effective NR-IQA technique used domain characteristics from the DCT for forecast detection performance [14].

As previously stated, NR approaches demand input data and no knowledge of the carrier frequency. Earlier noise reduction algorithms were mostly concentrated on distortion-specific techniques. As a result, a technique depending on the average squared deviations across frames was created to measure jerkiness. In comparison, a network is trained to simulate the effects of jerkiness on video quality. The error estimation was based on the data's DCT coefficients inside an H.264-specific technique. The movement matrices are collected from the bit stream, and sensory excellence ratings are calculated using the error estimations. Likewise, suggested an H.264-specific technique, although they used DCT coefficients to retrieve the first frame-level information. A training neural network was used to construct video characteristics by combining frame-level data [15].

In research, algorithms were also created to evaluate blocking artifacts in distorted films. The focus of later research was on general-purpose algorithms. Natural scene characteristics were used to construct a robust and extensively used feature extraction technique, assuming that natural visual signals include patterns in the data modified by distortion. The Video BLINDS method was created using a component of the Study's NR-IQA approach named BLINDS. The collected characteristics are then used to retrain an SVR using Video BLINDS, which uses a spatiotemporal framework built from the natural scene characteristics of the Transform coefficients. The 3D-DCT area was eventually added to this approach.

Unlike other techniques, the Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) does not need any knowledge of the forms of distortions or human visual quality assessments. Rather, it's argued that perfect video series have inherent patterns in the data and that departures from these can be employed to anticipate perceptual quality ratings. The main aspect of this approach is that assuming the video is of excellent quality, localized metrics linked to frame discrepancies produced using mean elimination and division properly performed must follow a modified Gaussian distribution. Using the NR-IQA CORNIA approach, they also introduced Video CORNIA, an opinion-unaware NR-VQA approach in which frame-level features are selected first by unsupervised extraction and subsequently used to train an SVR. Lastly, the perceived accuracy rate of the video is calculated using temporal pools of frame-level data. Likewise, suggested an opinion-free architecture for HEVC encrypted videos, in which performance is anticipated using a motion vector extractor and spatial data generated from the video content types [16].

The suggested system was trained using the KoNViD-1k database that contains several video sequences with actual deformation, as opposed to prior work that used intentionally distorted films. They used a combination of six spatial and 3 temporal variables to define a video sequence. Following that, such characteristics were mapped onto subjective peer assessments using a trained SVR. Deep learning methods are used in different areas of study. Although few NR-VQA approaches use deep learning, deep learning-based NR-IQA algorithms have suddenly gained wide acceptance. Weakly supervised learning was used to train a CNN, with the associated labeling for the video blocks produced using a complete reference metric. The extracted features were then recovered and converted into subjective performance scores for the training CNN [17].

VQA has recently received a lot of attention, especially in terms of assessing compressing and transmitting defects. Due to the optical flow data, the author created the NR-VQA framework. Statistically, abnormalities of consecutive frames at the patches and frame levels are measured to represent the effect of distortions on an optical flow that is then integrated with the SVR to forecast perceived video quality. To aggregate the quality rating, the author created an NR-VQA by integrating 3D shearlet transforms and deep learning. V-MEON is an NR-VQA approach that uses a 3D convolution operation to feature extracted. Combining spatial-temporal characteristics could contribute to improved prediction quality. The author retrieved Lower Complexity Features from the video sequence and higher complexity features [18].

## 4 | Shortcomings of Current Methods

Despite all the advancements in recent years, the majority of these techniques still have some potential weaknesses. The first issue with present methodologies is how to respond to a question requiring a lengthy chain of deductions. Additionally, none of these systems are particularly adept at handling queries requiring quick recall, such as integer equality. Another example of a question that might be quite difficult for the majority of the models we described in this paper is one concerning counting the number of a particular object in the image. There have recently been initiatives to solve these difficulties as well. The authors approached the numbering problem as serial decision-making and solved it using a reinforced learning strategy. The items that go into each count are also identified by this method. Future models can enhance the effectiveness of present techniques by expanding on them, such as co-attention or modularity networks, while also discussing the problems raised here, perhaps by employing a solution specifically designed to handle them.

## 5 | Proposed Method for Video Analysis

Fig. 1 shows the proposed NR-VQA algorithm's high-level workflow. This picture demonstrates that various pre-trained CNNs are used to recover deep frame-level selected features from each video frame, which are then combined to create multiple simultaneously pooled video-level feature vectors. Fig. 1 depicts the topology of our suggested deep feature pooling technique. The frame-level deep characteristics are retrieved first with the pre-trained models CNNs for a certain video series, which must be analyzed. Such frame-level extracted features are then temporally combined to form a video-level feature representation that describes the entire video. Lastly, using a trained SVR, the sequentially pooled video-level characteristics are translated into subjective performance ratings. The training and test database generation techniques are described. Pre-trained models CNN are used to retrieve frame-level features. Lastly, we go over how to extract video-level features.

## 6 | Composition of Databases

Several video quality databases, such as LIVE VQA, LIVE mobile video quality database, and MCL-V, are open to the public. We used the KoNViD-1k natural video quality database to train and test our system in this work. KoNViD-1k provides natural films with realistic distortions, unlike most earlier available data datasets comparable to LIVE-VQC. The videos are also drawn from the Yahoo flickr creative commons 100 million (YFCC100m) data source. The CrowdFlower platform was used to obtain the subjective performance scores. This data collection has a pixel density of  $960 \times 540$  pixels and a frame rate of 25, 27, or 30 frames per second. Moreover, video segments last at least 7 and 8 seconds [19], [20].

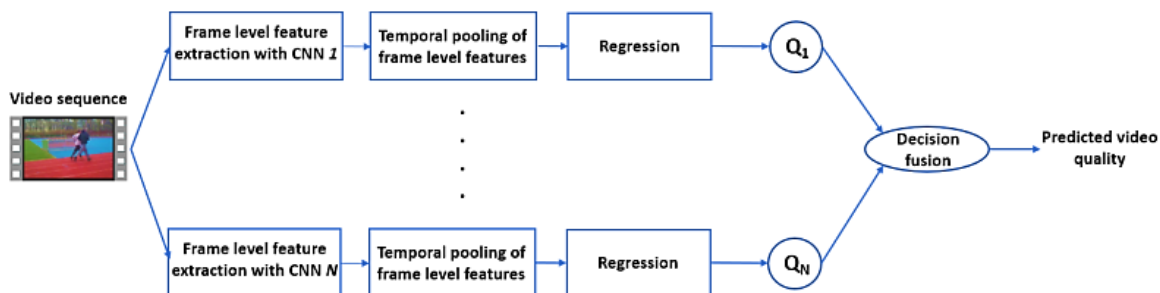


Fig. 1. The suggested algorithm's high-level processing.

## 7 | Retrieval of Frame-Level Features

Fig. 2 depicts the feature extraction method at the frame level. As earlier stated, frame-level extracted features are extracted separately from one another using a different variety of pre-trained CNNs. ResNet 18, ResNet50, GoogLeNet, GoogLeNet-Places365, and InceptionV3 are optimal parameters. All architecture is pre-trained on ImageNet, which includes over one billion photos and 1,000 meaningful classifications, except GoogLeNet-Places365. GoogLeNet-Places365, on the other hand, is learned using the Places-365 dataset with 17 million training photos from 366 scenario types. Saliency-Weighted Global Average Pooling (SWGAP) layering, a component of this research, is coupled to only certain components of the basic algorithms to collect frame-level characteristics. Because CNNs collect visual information at numerous levels, integrating various levels of deep features can enhance perceived quality assurance. The size of the collected extracted features and the components evaluated by the implemented pre-trained CNNs are summarized in Table 1. In the context of AlexNet and VGG16, it must be said that the characteristics of the convolution components are employed. In contrast, the characteristics of the residue and Inception modules are being used.

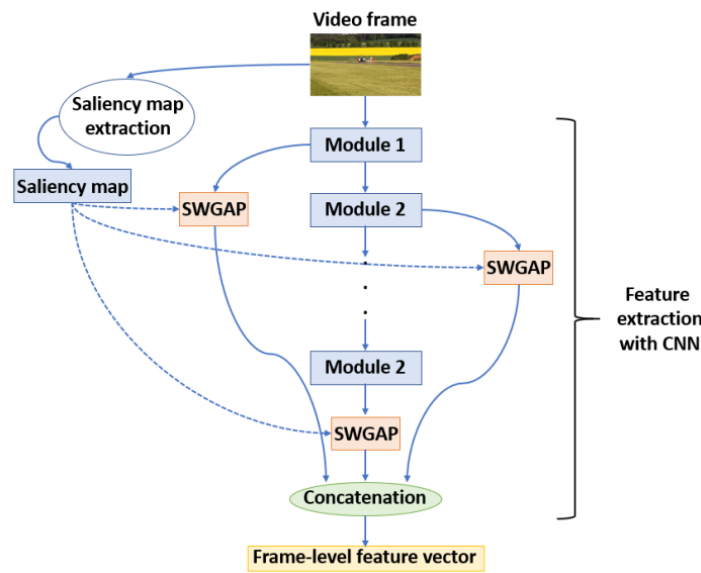


Fig. 2. Shows the extraction of frame-level features.

The lengths of the retrieved frame-level extracted features and the applied components in extracting the features are reported.

Table 1. A description of the CNNs used is shown.

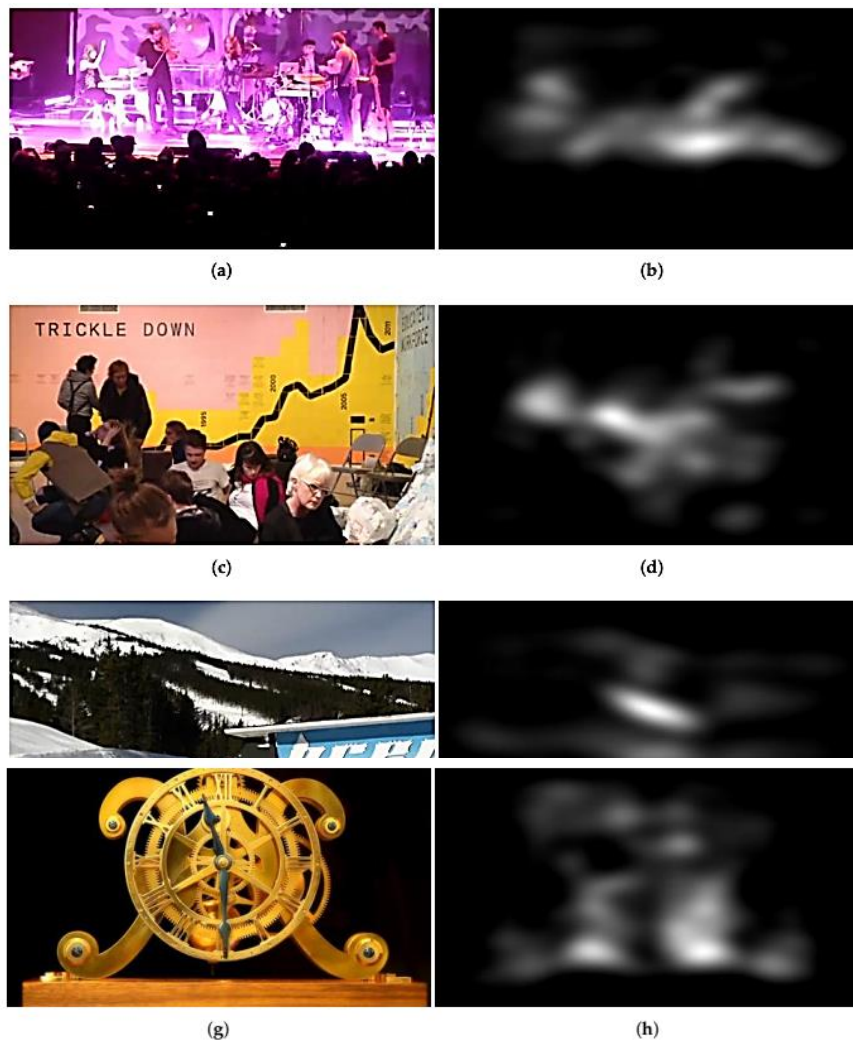
Base CNN	Module	Length of Feature Vector
AlexNet	Convolutional	1275
VGG16	Convolutional	4112
ResNet 18	Residual	1752
ResNet 50	Residual	15369
GoogLeNet	Inception	5287
GoogLeNet-Place 365	Inception	5875
Inception V3	Inception	11,251

In CNN, Global Average Pooling (GAP) levels are commonly employed to impose congruence across convolution layers and the number of semantic classifications, allowing networks to be trained on pictures of varied resolutions. A further frequent GAP application uses a CNN to extract frequency-independence graphic resources from photos. We upgrade GAP to SWGAP for extraction of features leveraging visual significance in this study. Visual saliency algorithms, for example, are concerned with locating the most noticeable areas of a digital image from a perception standpoint. Humans tend to fixate on specific portions of the image throughout the first 3 seconds of assessment, which is particularly important in estimating

perceived performance. Due to the above findings, SWGAP is suggested for extracting features to demonstrate specific regions prominent to human visual perception. SWGAP implements a weighted arithmetic function across a CNN's  $F(i,j)$  feature map and the input image's scaled  $S(i,j)$  saliency map (bilinear interpolation is used). It can be expressed in formal terms as

$$\sigma = \frac{\sum_{i=1}^M \sum_{j=1}^N S(i,j).F(i,j)}{\sum_{j=1}^N S(i,j)}$$

Where  $\sigma$  indicates the result worth of SWGAP for one element map; further,  $M$  and  $N$  separately represent the level and width of the element map.  $I$  and  $j$  mean the directions of the component maps and the resized saliency map. In this review, the strategy was applied to decide the saliency guide of a video outline because of its low computational expenses. *Fig. 3* portrays a few video outlines and their saliency maps.



**Fig. 3. Saliency map retrieval: input video segments (a, c, e, g) and saliency mappings of the source video sequence (b, d, f, h) [6].**

## 8 | Datasets

This paper uses two enormous genuine VQA data sets KoNViD-1k and LIVE VQC, to assess the proposed strategy and other cutting-edge calculations. The recordings of KoNViD-1k were gathered from the YFCC100m [8] data set and assessed in a huge scope publicly supporting examination, including 642 human evaluators who produced no less than 50 quality evaluations for each video. The recordings' goal is  $960 \times 540$ , and the MOS goes from 1 to 5. A VQA data set containing 585 remarkable video successions with bona fide contortions caught by 80 distinct clients with 101 unique camera gadgets. Comparably to KoNViD-1k, the recordings were assessed in a huge scope publicly supporting

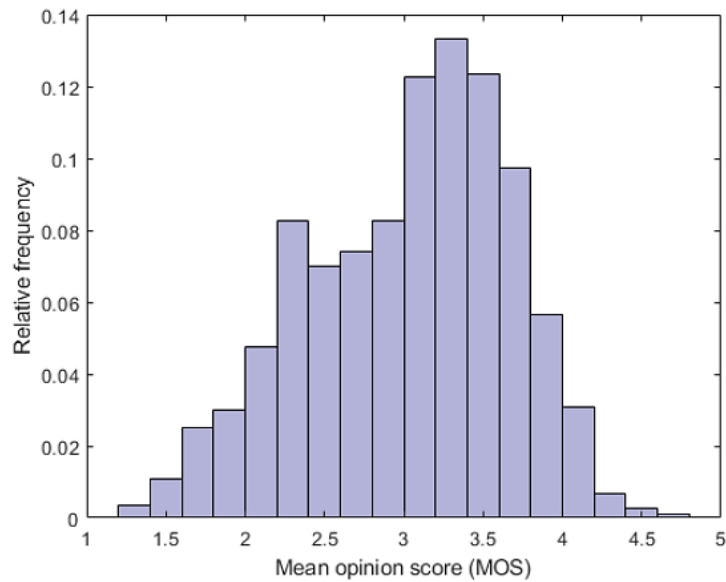


trial including 4776 human eyewitnesses who created more than 205,000 quality evaluations. Rather than KoNViD-1k contains recordings with different picture goals, and the MOS goes from 0 to 100. Not at all like KoNViD-1k, LIVE VQC has no proper picture goal.

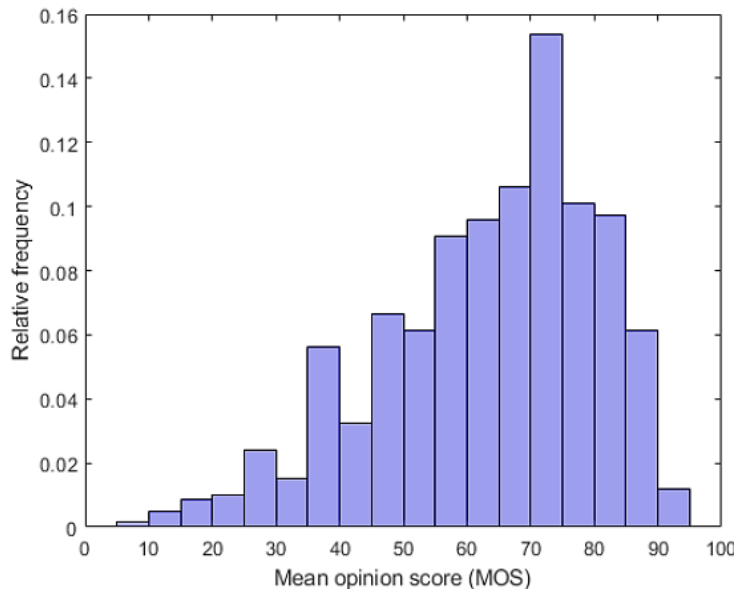
**Table 2. Provides an overview of the VQA databases that have been used.**

Attribute	KoNViD	LIVE VQC
Videos	1100	5451275
Devices	> 142	4754
Test subjects	521	MP4
Format	MP4	Authentic
Distortion	Authentic	Crowd-sourcing
Test environment	Crowd-sourcing	Crowd-sourcing

Table 3 summarizes the key characteristics of the used VQA databases. Figs. 4 and 5, respectfully, show the MOS distribution observed in KoNViD-1k and LIVE VQC. Numerous videos from the KoNViD-1k VQA test dataset are shown in Fig. 6. Similarly to Fig. 6, Fig. 7 shows several LIVE VQC videos.



**Fig. 4. Shows the experimental MOS distributions using KoNViD-1k.**



**Fig. 5. MOS dispersion with LIVE VQC detailed empirical.**

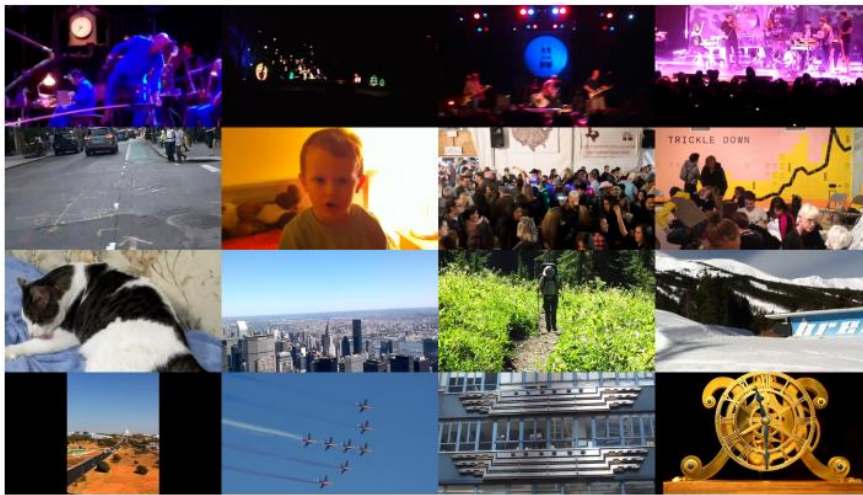


Fig. 6. Shows a selection of films from the KoNViD-1k VQA benchmark dataset [6].

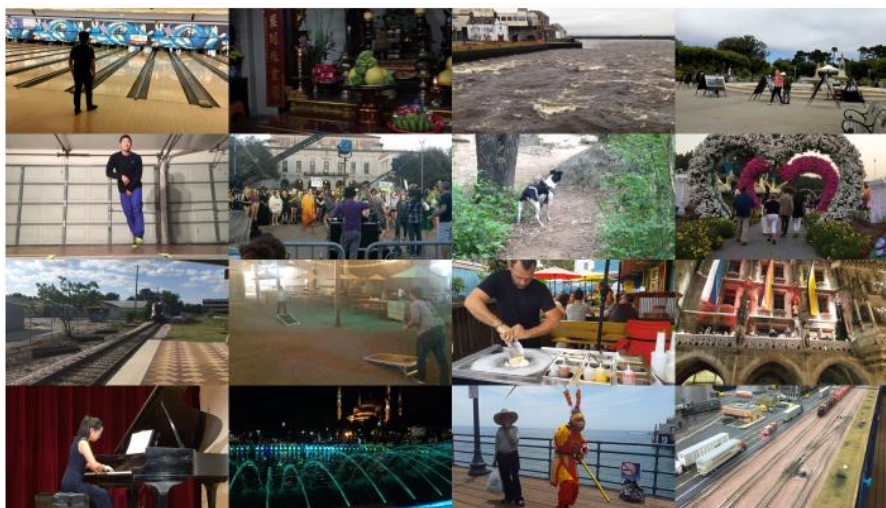


Fig. 7. Shows a selection of videos from the LIVE VQC VQA benchmark database [4].

## 9 | Evaluation Protocol

The assessment of VQA calculations depends on deciding the relationship between the ground-truth scores of a VQA data set and the anticipated scores given by the calculation. Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank-Order Connection Coefficient (SROCC) are applied in the writing. As currently referenced, KoNViD-1k and LIVE VQC are utilized to evaluate the proposed and other cutting-edge techniques. To this end, a VQA information base is haphazardly partitioned into a preparation set (~80% of recordings) and a test set (~20% of recordings) to prepare a VQA technique. This cycle is rehased multiple times.

Further, middle PLCC and SROCC are accounted for in this paper. As proposed, non-direct planning between the anticipated and the ground-truth scores is executed previously in the computation of PLCC. In particular, a calculated capability with five boundaries is utilized to this end.

## 10 | Results

This subsection introduces a removal study to reason the plan decisions of the proposed strategy for various element extraction and relapse methods. Also, we show that the chosen combination of numerous profound models essentially works on the exhibition of NR-VQA. To this end, the KoNViD-

1k data set was applied in this removal, concentrating on utilizing the assessment convention. The outcomes are summarized in *Tables 3 to 5*. From these outcomes, it very well may be seen that GPRs with normal quadratic piece capability give a lot better presentation than SVRs with Gaussian bit capability in every conceivable case.

Regarding the choice combination technique, we can see that the basic normal is better for the assessment execution than taking the middle of the regressors' results. More importantly, it is obvious that combining the outputs of numerous CNNs enhances predictive accuracy significantly. Furthermore, replacing GAP layers with the suggested SWGAP layers can increase performance since SWGAP uses a visual saliency weighted average rather than a basic arithmetic mean, allowing picture regions that are important to the HVS to be emphasized. As a result, in the suggested technique, which is software SWDF-DF-VQA in the following, GPRs with irrational quadratic basis functions, SWGAP levels, and mathematical averaging as decision fusing have been used.

Use GAP layers for extracting features and SVRs with Gaussian kernel functions for regression. Over 1000 random train–test divisions, the median PLCC and SROCC are calculated.

**Table 3. Results of several base designs and data fusion techniques.**

Base CNN	PLOC	SROCC
AlexNet	0.852	0.743
VGG16	0.741	0.852
ResNet 18	0.745	0.746
ResNet 50	0.522	0.766
T6GoogLeNet	0.786	0.744
GoogLeNet-place 365	0.743	0.852
Inception V3	0.788	0.798
All-median	0.791	0.812
All-average	0.789	0.815

**Table 4. Results of several base architectures and decision fusion approaches.**

Base CNN	PLOC	SROCC
AlexNet	0.752	0.701
VGG16	0.721	0.841
ResNet 18	0.740	0.751
ResNet 50	0.500	0.784
T6googlenet	0.788	0.745
GoogLeNet-place 365	0.843	0.810
Inception V3	0.858	0.784
All-median	0.702	0.855
All-average	0.778	0.860

An architecture that uses GAP layers for feature extraction and GPR with reasonable quadratic kernel functions for regression (*Table 4*). Over 1000 random parts of the train test, average PLCC and SROCC have been calculated.

**Table 5. Results of various base architectures and decision fusion approaches.**

Base CNN	PLOC	SROCC
AlexNet	0.80	0.725
VGG16	0.742	0.758
ResNet18	0.645	0.745
ResNet50	0.612	0.714
T6GoogLeNet	0.726	0.748
GoogLeNet-place 365	0.703	0.755
Inception V3	0.718	0.820
All-median	0.702	0.820
All-average	0.721	0.842

Use SWGAP layers to extract features and SVRs with Gaussian kernel functionality for prediction (Table 5). Over 1000 random train–test splits, the average PLCC and SROCC are calculated.



## 11 | Conclusions

In this research, we describe a novel deep learning-based strategy for NR-VQA that extracts features in parallel using a series of pre-trained CNNs. The basic idea behind this design was that a group of pre-trained CNNs might catch possible picture distortion more effectively than a single system. With the help of various CNNs, spatially pooling and intensity weighted deep extracted features are created. Following that, such extracted features are transformed into subjective quality monitoring, and a data fusion procedure is used to produce the overall video sequence's quality score. With thorough scientific results, we showed that combining the deep segmentation method and decision can significantly increase prediction performance compared to single neural network designs. On two large benchmark VQA datasets featuring actual distortions, the suggested technique is contrasted with other recent NR-VQA algorithms. Extensive tests have shown that the suggested technique presented scheme is a new state-of-the-art in the field of authentic distortions. Based on the findings, further study could go into several areas. For example, properly combining motion and deep features to identify video distortions is worth investigating. Furthermore, a feature-level fusion of CNNs can be a useful way to chop training time and computing expenses.

## References

- [1] Li, X., & Qiu, J. (2021). A multi-parameter video quality assessment model based on 3D convolutional neural network on the cloud. *ASP transactions on internet of things*, 1(2), 14–22.
- [2] Bianco, S., Celona, L., Napoletano, P., & Schettini, R. (2018). On the use of deep learning for blind image quality assessment. *Signal, image and video processing*, 12, 355–362.
- [3] Varga, D. (2019). No-reference video quality assessment based on the temporal pooling of deep features. *Neural processing letters*, 50(3), 2595–2608.
- [4] Chen, P., Li, L., Wu, J., Dong, W., & Shi, G. (2021). Contrastive self-supervised pre-training for video quality assessment. *IEEE transactions on image processing*, 31, 458–471.
- [5] Varga, D. (2022). No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion. *Sensors*, 22(6), 2209.
- [6] Hong, C., Chen, X., Wang, X., & Tang, C. (2016). Hypergraph regularized autoencoder for image-based 3D human pose recovery. *Signal processing*, 124, 132–140.
- [7] Xue, J., Yin, L., Lan, Z., Long, M., Li, G., Wang, Z., & Xie, X. (2021). 3D DCT based image compression method for the medical endoscopic application. *Sensors*, 21(5), 1817.
- [8] Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., ... & Saupe, D. (2017). The konstanz natural video database (konvid-1k). *2017 ninth international conference on quality of multimedia experience (qomex)* (pp. 1–6). IEEE. <https://doi.org/10.1109/QoMEX.2017.7965673>
- [9] Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of medical imaging*, 3(3), 34501.
- [10] Vranješ, M., Rimac-Drlje, S., & Vranješ, D. (2018). Foveation-based content adaptive root mean squared error for video quality assessment. *Multimedia tools and applications*, 77, 21053–21082.
- [11] Koike, M., Urata, Y., & Yamagishi, K. (2022). Bitstream-quality-estimation model for tile-based VR video streaming services. *IEICE transactions on communications*, 105(8), 1002–1013.
- [12] Li, J., Zou, L., Yan, J., Deng, D., Qu, T., & Xie, G. (2016). No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks. *Signal, image and video processing*, 10, 609–616.
- [13] Li, X., Guo, Q., & Lu, X. (2016). Spatiotemporal statistics for video quality assessment. *IEEE transactions on image processing*, 25(7), 3329–3342.
- [14] Li, Y., Po, L. M., Cheung, C. H., Xu, X., Feng, L., Yuan, F., & Cheung, K-W. (2015). No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE transactions on circuits and systems for video technology*, 26(6), 1044–1057.

- [15] Villaret, M., & others. (2021). Efficient fundus image gradeability approach based on deep reconstruction-classification network. *Artificial intelligence research and development: proceedings of the 23rd international conference of the catalan association for artificial intelligence* (Vol. 339, p. 402). IOS Press. [https://books.google.com/books?id=LYxJEAAAQBAJ&lr=&source=gbs\\_navlinks\\_s](https://books.google.com/books?id=LYxJEAAAQBAJ&lr=&source=gbs_navlinks_s)
- [16] Koike, M., Urata, Y., Egi, N., & Yamagishi, K. (2021). Extension of itu-t p. 1204.3 model to tile-based vr streaming services. *2021 ieee international workshop technical committee on communications quality and reliability (cqr 2021)* (pp. 1–6). IEEE. <https://doi.org/10.1109/CQR39960.2021.9446237>
- [17] Saad, M. A., Bovik, A. C., & Charrier, C. (2011). DCT statistics model-based blind image quality assessment. *2011 18th ieee international conference on image processing* (pp. 3093–3096). IEEE. <https://doi.org/10.1109/ICIP.2011.6116319>
- [18] Saad, M. A., Bovik, A. C., & Charrier, C. (2014). Blind prediction of natural video quality. *IEEE transactions on image processing*, 23(3), 1352–1365.
- [19] Saupé, D., Hahn, F., Hosu, V., Zingman, I., Rana, M., & Li, S. (2016). Crowd workers proven useful: a comparative study of subjective video quality assessment. *QoMEX 2016: 8th international conference on quality of multimedia experience*. konstanzer online-publikations-system (KOPS). <http://nbn-resolving.de/urn:nbn:de:bsz:352-0-371921>
- [20] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85–117.