



# No-Reference Video Quality Assessment Based on the Temporal Pooling of Deep Features

Domonkos Varga<sup>1</sup>

Published online: 12 April 2019  
© The Author(s) 2019

## Abstract

Video quality assessment (VQA) is an important element of various applications ranging from automatic video streaming to display technology. Furthermore, visual quality measurements require a balanced investigation of visual content and features. Previous studies have shown that the features extracted from a pretrained convolutional neural network are highly effective for a wide range of applications in image processing and computer vision. In this study, we developed a novel architecture for no-reference VQA based on the features obtained from pretrained convolutional neural networks, transfer learning, temporal pooling, and regression. In particular, we obtained solutions by only applying temporally pooled deep features and without using manually derived features. The proposed architecture was trained based on the recently published Konstanz natural video quality database (KoNViD-1k), which contains 1200 video sequences with authentic distortion unlike other publicly available databases. The experimental results obtained based on KoNViD-1k demonstrated that the proposed method performed better than other state-of-the-art algorithms. Furthermore, these results were confirmed by tests using the LIVE VQA database, which contains artificially distorted videos.

**Keywords** No-reference video quality assessment · Convolutional neural network

## 1 Introduction

Multimedia technology and digital visual signal processing have developed rapidly during recent decades. Digital images and videos are very easy to create, transmit, store, and share. Owing to these developments, the design of reliable video quality assessment (VQA) algorithms has attracted considerable attention. Consequently, VQA has been the focus of many research studies and patents. Furthermore, the vast volume of user-created digital video content has led to the development of numerous VQA applications, which require reliable and effective quality monitoring [39].

---

✉ Domonkos Varga  
varga.domonkos7@upcmail.hu

<sup>1</sup> Department of Networked Systems and Services, Budapest University of Technology and Economics, Magyar tudósok krt. 2, Budapest 1117, Hungary

Visual signals can undergo a wide variety of distortions after their capture during compression, transmission, and storage. Human observers are the end users of visual content; thus, the quality of visual signals should ideally be evaluated in subjective user studies in a laboratory environment involving specialists. During these user studies, subjective quality scores are collected from each participant. Subsequently, the quality of a visual signal is given a mean opinion score (MOS), which is calculated as the arithmetic mean of all the individual quality ratings. In most cases, an absolute category rating is applied, which ranges from 1.0 (bad quality) to 5.0 (excellent quality). Other standardized quality ratings also exist, such as a continuous scale ranging from 1.0 to 100.0, but Huynh-Thu et al. [11] noted that there are no statistical differences between the different scales used for the same visual stimuli.

However, subjective VQA is expensive, time consuming, and labor intensive, thereby preventing its application to real-time systems. Moreover, the results obtained by subjective VQA depend on the physical condition, emotional state, personality, and culture of the observers [27]. As a consequence, there is an increasing need for objective VQA. The classification of VQA algorithms is based on the availability of the original (reference) signal. If a reference signal is not available, a VQA algorithm is regarded as a no-reference VQA (NR-VQA). NR algorithms can be classified into two further groups, where the so-called *distortion-specific* NR algorithms assume that specific distortion is present in the visual signal, whereas *general purpose* (or *non-distortion specific*) algorithms operate on various distortion types. Reduced-reference methods retain only part of the information from the reference signal, whereas full-reference algorithms have full access to the complete reference medium to predict the quality scores.

Deep learning is now applied widely in industry and research, and with great success in the fields of image processing and computer vision [7,8,44]. Thus, recently developed NR-VQA algorithms have employed deep learning techniques, such as neural networks [47], convolutional neural networks (CNNs) [2], and deep belief networks [5].

It has been shown that the features extracted using a pretrained CNN are rich and effective for a wide range of computer vision and image processing tasks, such as content-based image retrieval [35], NR image quality assessment [2], and medical image classification [10]. The main contribution and novel aspect of the present study is that we obtain possible solutions for NR-VQA using only the deep features extracted from pretrained CNNs (*Inception-V3* [32] and *Inception-ResNet-V2* [31]) without depending on manually selected features. In particular, for a given video sequence that needs to be assessed, frame-level deep features are extracted from each video frame with a pretrained CNN. Subsequently, these frame-level features are temporally pooled to compile a video-level feature vector that characterizes the video sequence. Finally, the temporally pooled video-level feature vectors are mapped onto perceptual quality scores with a support vector regressor (SVR). Furthermore, our architecture was trained based on the recently published Konstanz natural video quality database (KoNViD-1k) [9], which in contrast to other publicly available databases, contains video sequences with authentic distortion rather than artificial distortion. Moreover, KoNViD-1k contains more videos (1200 sequences) than any other publicly available databases, which allowed us to create a deep, temporally pooled model.

The remainder of this paper is organized as follows. In Sect. 2, we review related research, particularly into NR-VQA algorithms. In Sect. 3, we describe our proposed NR-VQA algorithm. In Sect. 4, we present the experimental results. We give our conclusions in Sect. 5.

## 2 Related Work

As mentioned earlier, NR methods only require an input signal and no information about the reference signal. Early NR algorithms largely focused on distortion-specific approaches. Thus, Borer [3] developed an algorithm for measuring jerkiness based on the mean squared difference between frames. By contrast, Xue et al. [42] trained a neural network to model the quality impact of jerkiness. An H.264-specific algorithm was introduced by [4], where the error estimate depended on the discrete cosine transform (DCT) coefficient for data. Subsequently, perceptual quality scores were derived from the error estimates, and the motion vectors were obtained from the bit stream. Similarly, Zhu et al. [47] proposed a H.264-specific method; however, they extracted the first frame-level features using DCT coefficients. In addition, video-level features were created by averaging the frame-level features (temporal pooling), and a trained neural network predicted the subjective quality scores. Algorithms were also developed to assess blocking artifacts in distorted videos in studies by [21,33,38].

Subsequent studies focused on general-purpose algorithms. A successful and widely applied feature extraction method was developed based on natural scene statistics [22], where it was assumed that natural visual signals contain statistical regularities that are changed by distortion. Saad et al. [24] implemented this feature of their NR-IQA method called BLind Image Integrity Notator with DCT Statistics (BLIINDS) [23] for NR-VQA to produce the *Video* BLIINDS method. *Video* BLIINDS employs a spatiotemporal model derived from the natural scene statistics of the DCT coefficients, and the extracted features are then employed to train an SVR. This method was later extended to the three-dimensional (3D)-DCT domain [14].

In contrast to other methods, the video intrinsic integrity and distortion evaluation oracle (VIIDEO) [19] requires no information regarding the distortion types or human ratings of the video quality. Instead, it is assumed that pristine video sequences contain intrinsic statistical regularities, and deviations from them can be used to predict perceptual quality scores. The main feature of this method is that local statistics related to the frame differences derived using mean removal and divisive contrast normalization should follow a generalized Gaussian distribution if the video is of a good quality. Based on the NR-IQA CORNIA method [45], Xu et al. [40] also proposed an opinion-unaware NR-VQA method called *Video* CORNIA, where the frame-level features are first extracted via unsupervised feature learning and these features are then used to train an SVR. Finally, the video's perceptual quality score is derived by temporal pooling of the frame-level features. Similarly, Anekekuh et al. [1] presented an opinion-unaware architecture for HEVC encoded videos, where the quality is predicted based on motion vector extraction and spatial information derived from the video content type.

In contrast to previous studies where training was conducted using artificially distorted videos, the algorithm proposed by Men et al. [18] was trained using the KoNViD-1k database [9], which comprises numerous video sequences with authentic distortion. They combined six spatial features and three temporal features to characterize a video sequence. Subsequently, a trained SVR was used to map these features onto perceptual quality scores.

Another area of research is based on deep learning techniques. Recently, deep learning-based NR-IQA algorithms have increased in popularity [2,12,13], although very few NR-VQA methods utilize deep learning. Zhang et al. [46] trained a CNN by weakly supervised learning where the corresponding labels were obtained for the video blocks according to a full reference-VQA metric. Subsequently, the feature vectors were extracted using the trained CNN and mapped onto subjective quality scores. By contrast, Li et al. [15] applied a

3D shearlet transform to video blocks and compiled spatiotemporal feature vectors for each video sequence. CNN and logistic regression were then utilized to map the features onto perceptual quality scores. Torres Vega et al. [36] proposed a restricted Boltzmann machine-based solution, which was trained with lightweight NR metrics, such as the noise ratio, motion intensity, and blockiness. This method was developed for assessing the quality of live video streams.

For further reviews of NR-VQA, we refer the reader to the studies by [29,37,41].

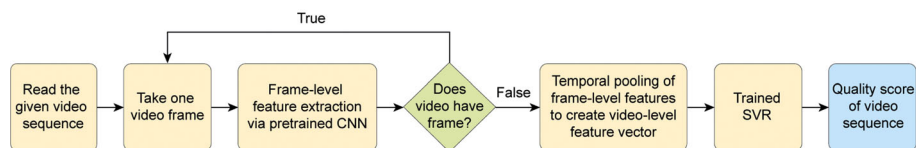
### 3 Methodology

The architecture of our proposed deep feature pooling algorithm is shown in Fig. 1. For a given video sequence that needs to be evaluated, the frame-level deep features are first extracted with the pretrained CNNs. Subsequently, these frame-level feature vectors are temporally pooled to create a video-level feature vector that characterizes the whole video. Finally, the temporally pooled video-level features are mapped onto subjective quality scores with a trained SVR.

The remainder of this section is organized as follows. In Sect. 3.1, we describe the training and test database compilation processes. Frame-level feature extraction is conducted by pretrained CNNs and the detailed process is presented in Sect. 3.2. Finally, we explain the video-level feature extraction method in Sect. 3.3.

#### 3.1 Database Compilation

Several video quality databases are publicly available, such as LIVE VQA [28], LIVE mobile video quality database [20], and MCL-V [16]. In this study, we selected the KoNViD-1k [9] natural video quality database to train and test our system. In contrast to most previously published data sets and similar to LIVE-VQC [30], KoNViD-1k [9] contains natural videos with authentic distortions. Furthermore, the videos were sampled from the Yahoo Flickr Creative Commons 100 Million (YFCC100m) [34] data set. The subjective quality scores were collected online [25] using the CrowdFlower platform. The spatial resolution is  $960 \times 540$  in this data set and the frame rate is 25, 27 or 30 fps. Furthermore, the length of video sequences varies between 7 and 8 s. The MOS values for the video sequences are on a scale from 1.0 (worst) to 5.0 (best). Furthermore, KoNViD-1k contains more quality labeled video sequences than any other publicly available data sets. The large number of video sequences in KoNViD-1k allowed us to directly train a temporal feature pooling model using the features extracted from pretrained CNNs.



**Fig. 1** General structure of the proposed NR-VQA algorithm. The algorithm reads a given video sequence and processes each of the video frames in turn to extract the frame-level feature vectors with the pretrained CNN. Finally, the extracted frame-level feature vectors are temporally pooled to form a video-level feature vector, which is mapped onto a quality score with a trained SVR

KoNViD-1k contains 1200 video sequences. We randomly selected 960 sequences for training purposes, whereas the remaining 240 sequences were retained only for testing and they were not utilized in the training process. The videos selected for training were split into frames and 20% of the frames were then selected randomly. We employed Inception-V3 [32] and Inception-ResNet-V2 [31] as feature extractors because their input receptive fields are significantly larger than those in other pretrained networks ( $299 \times 299$  vs.  $224 \times 224$  or  $227 \times 227$ ) and in case of input image's resizing, the visual clues of perceptual quality deteriorates to a lesser extent than with other pretrained CNNs. As a consequence of fixed input size, the selected video frames were resized to  $338 \times 338$  and  $299 \times 299$  center patches were cropped from the resized video frames. The resulting training images retained the MOS values of their source videos. Consequently, we assumed that the perceived visual quality of the individual frames was related to that of the complete video sequence. The final image database contained 43,320 images, which were used for transfer learning with the selected pretrained CNNs.

For completeness, we selected the LIVE VQA [28] database as an additional test set in order to analyze the generalizability of the proposed algorithm. LIVE VQA contains 15 reference videos and 150 artificially distorted video sequences with length of 8, 10 or 20 s obtained using four different types of distortion: simulated transmission of H.264 compressed videos through error-prone wireless networks and through error-prone IP networks, H.264 compression, and MPEG-2 compression. The spatial resolution of the videos in LIVE VQA is  $768 \times 432$ .

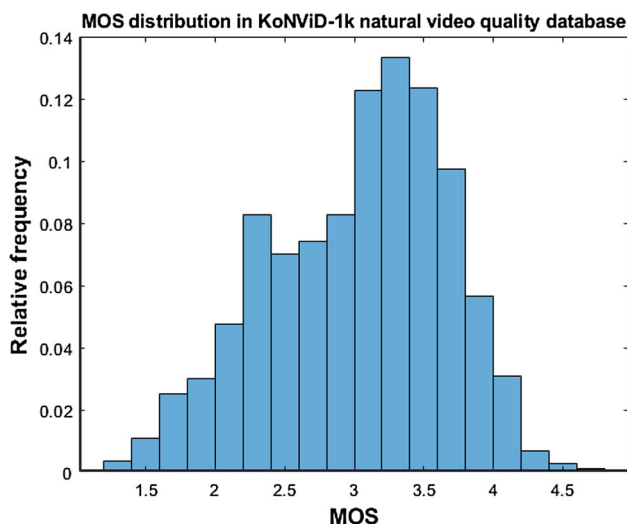
### 3.2 Frame-Level Feature Extraction

The features were extracted by providing the CNN with the whole image, which had to fit with the CNN's input size. As mentioned above, both Inception-V3 and Inception-ResNet-V2 accept images measuring  $299 \times 299$ , which is why the input video frames were resized to  $338 \times 338$  and the  $299 \times 299$  center patches were then cropped. The CNNs employed were fine-tuned (so-called transfer learning) based on the image database described above.

#### 3.2.1 Transfer Learning

The usual method was employed for transfer learning, where we truncated the last 1000-way softmax layer of the Inception-v3 and Inception-ResNet-v2 network. Furthermore, this layer was replaced by a 5-way softmax layer, which was relevant to the problem addressed. Five classes were defined in our training image database: class *A* for excellent image quality ( $5.0 \geq MOS \geq 4.2$ ), class *B* for good image quality ( $4.2 > MOS \geq 3.4$ ), class *C* for fair image quality ( $3.4 > MOS \geq 2.6$ ), class *D* for poor image quality ( $2.6 > MOS \geq 1.8$ ), and class *E* for very poor image quality ( $1.8 > MOS \geq 1.0$ ). During transfer learning, the initial learning rate was set to 0.0001 and divided by 10 when the validation error stopped improving. Moreover, the batch size was set to 32 and the momentum was adjusted to 0.9. During transfer learning, the last new layer was trained from scratch utilizing Xavier initialization [6], where the initial weights for the other layers came from the corresponding layers in the pretrained networks and all the layers were updated using the back-propagation algorithm [26].

As shown in Fig. 2, the MOS distribution is imbalanced in the KoNViD-1k natural video quality database, which could cause problems during transfer learning. Thus, we sampled each instance in the batch based on the inverse frequency of the class. Consequently, instances



**Fig. 2** MOS distribution in the KoNViD-1k [9] natural video quality database. KoNViD-1k is a video quality database that contains 1200 real-world video sequences with authentic distortion collected from the YFCC100m data set [34]. Furthermore, the database contains the corresponding MOS values on a scale from 1.0 (worst) to 5.0 (best)

were selected in larger classes with lower probabilities. The final batch was equally distributed because of differences in the populations of the classes.

Figure 3 depicts the training process with Inception-V3 [32] during transfer learning, where the training accuracy, training loss, validation accuracy, and validation loss are plotted.

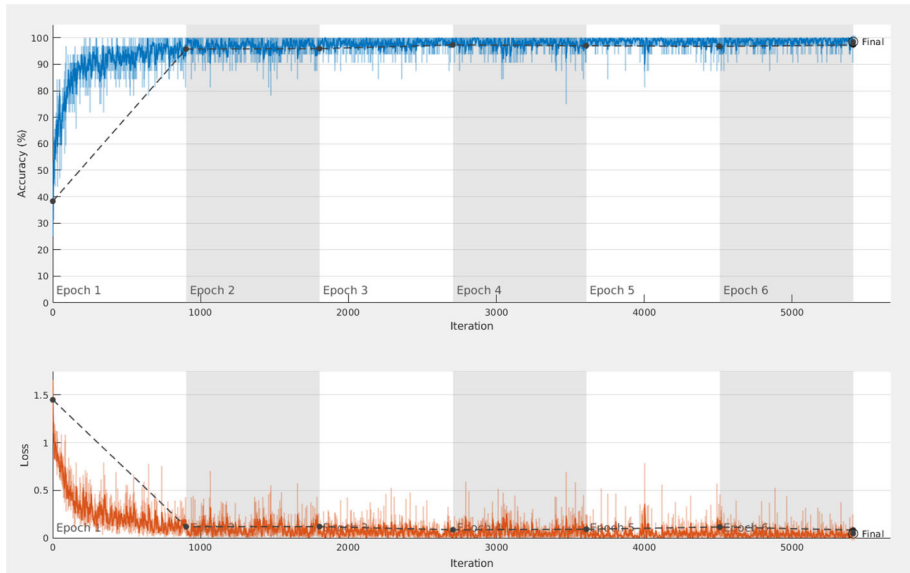
### 3.2.2 Feature Extraction

Frame-level feature vectors were extracted by providing the CNN with video frames that fitted with the CNN's input receptive field. As mentioned above, both Inception-V3 [32] and Inception-ResNet-V2 [31] accept  $299 \times 299$ -sized images, which is why each video frame was resized to  $338 \times 338$  and the  $299 \times 299$  center patch was cropped.

The CNN performed all its defined operations for an input image. Therefore, it is run through an each resized and center-cropped video frame saving the output from the final pooling layer which is named '*avg\_pool*' both in Inception-V3 and Inception-ResNet-V2. As a consequence, the length of the frame-level feature vectors was 2048 using Inception-V3 and 1536 when we employed Inception-ResNet-V2.

### 3.3 Video-Level Feature Extraction

Information fusion was conducted element by element based on each frame's feature vectors to create a single, video-level feature vector for each video sequence. Average, median, minimum, and maximum pooling were considered. Let  $f_i^{(j)}$  denote the  $i$ th entry of the  $j$ th video frame's feature vector. Furthermore, let  $N_f$  be the number of frames in a video sequence and  $M$  is the length of the frame-level feature vector. The four different pooling strategies can be formally expressed as:



**Fig. 3** Training process for Inception-V3 [32] during transfer learning. The smoothed training accuracy is shown by the dark blue line, the training accuracy by the light blue line, the smoothed training loss by the orange line, and the training loss by the light orange line. Furthermore, the validation accuracy and validation loss are depicted with dashed lines. The final checkpoint is denoted by a double round which is determined by early stopping. (Color figure online)

$$F_i^{avg} = \frac{1}{N_f} \sum_{j=1, \dots, N_f} f_i^{(j)}, i = 1, \dots, M, \quad (1)$$

$$F_i^{median} = \text{median}_{j=1, \dots, N_f} f_i^{(j)}, i = 1, \dots, M, \quad (2)$$

$$F_i^{min} = \min_{j=1, \dots, N_f} f_i^{(j)}, i = 1, \dots, M, \quad (3)$$

$$F_i^{max} = \max_{j=1, \dots, N_f} f_i^{(j)}, i = 1, \dots, M, \quad (4)$$

where  $F_i$  denotes the  $i$ th entry of the video-level feature vector. Consequently, the length of the video-level feature vector is equal to the length of the frame-level feature vector.

## 4 Experimental Results and Analysis

The proposed NR-VQA algorithms were evaluated based on their performance with the benchmark VQA databases, which were labeled with the subjective scores and MOS values representing the overall image quality. The Pearsons linear correlation coefficient (PLCC) and Spearman's rank ordered correlation coefficient (SROCC) were computed between the predicted and ground-truth scores, which are widely accepted performance metrics. The PLCC between two data sets, A and B, is defined as:

$$PLCC(A, B) = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}}, \quad (5)$$

where  $\bar{A}$  and  $\bar{B}$  denote the average of sets  $A$  and  $B$ , and  $A_i$  and  $B_i$  denote the  $i$ th elements of sets  $A$  and  $B$ , respectively. For two ranked sets  $A$  and  $B$ , SROCC is defined as:

$$SROCC(A, B) = \frac{\sum_{i=1}^n (A_i - \hat{A})(B_i - \hat{B})}{\sqrt{\sum_{i=1}^n (A_i - \hat{A})^2} \sqrt{\sum_{i=1}^n (B_i - \hat{B})^2}}, \quad (6)$$

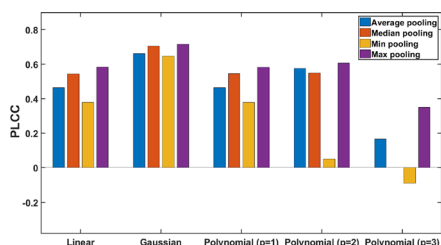
where  $\hat{A}$  and  $\hat{B}$  are the middle ranks.

#### 4.1 Parameter Study

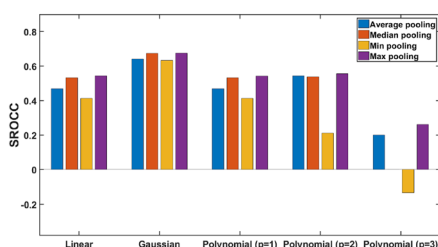
First, we evaluated the design choices for our proposed method, before comparing it with other state-of-the-art NR-VQA techniques. As mentioned above, two different publicly available databases were used for training and testing purposes, i.e., KoNViD-1k [9] for training and testing, and the LIVE VQA database only for testing. To evaluate the performance of our proposed architecture and the effects of the parameters in the algorithm, we used four different pooling strategies (average, median, minimum, and maximum) and SVRs with different kernel functions (linear, Gaussian, 1st-order polynomial, 2nd-order polynomial, and 3rd-order polynomial). The different versions of our algorithm were assessed based on KoNViD-1k [9] by fivefold cross-validation with ten replicates in the same manner as the study by [14].

Figures 4, 5, 6 and 7 summarize the results obtained with different design choices. The results showed that the architectures based on SVRs with Gaussian kernel functions obtained significantly better results than the architectures with other kernel functions. Furthermore, SVRs with third order polynomial kernel functions apparently overfit the training data because they produce 0 or negative values PLCC and SROCC values on the test. The difference between linear and 1st order kernel function is marginal. Compared to these, 2nd order polynomial kernel function performs slightly worse results.

Further, we evaluated the architectures with and without transfer learning, where the results demonstrated that transfer learning significantly improved the performance. Specifically, it can be clearly seen that transfer learning significantly improved the performance because it was able to improve PLCC and SROCC by at least 0.1 in all cases except those architectures



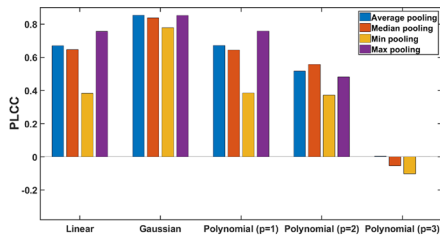
(a) PLCC 5-fold cross-validation with 10 repetitions. Transfer learning was not carried out on Inception-V3 [32].



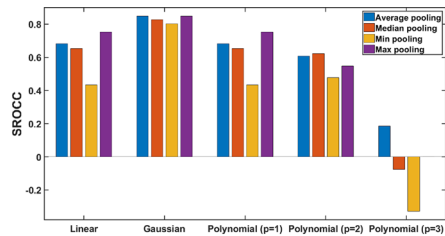
(b) SROCC 5-fold cross-validation with 10 repetitions. Transfer learning was not carried out on Inception-V3 [32].

**Fig. 4** Pooling technique and SVR comparison trained and tested on KoNViD-1k [9] using Inception-V3 [32] base architecture



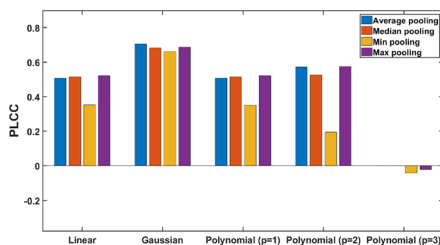


(a) PLCC 5-fold cross-validation with 10 repetitions. Transfer learning was carried out on Inception-V3 [32].

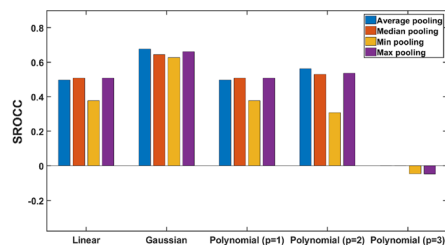


(b) SROCC 5-fold cross-validation with 10 repetitions. Transfer learning was carried out on Inception-V3 [32].

**Fig. 5** Pooling technique and SVR comparison trained and tested on KoNViD-1k [9] using Inception-V3 [32] base architecture

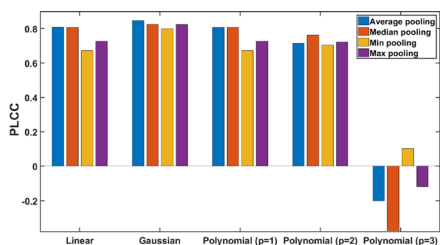


(a) PLCC 5-fold cross-validation with 10 repetitions. Transfer learning was not carried out on Inception-ResNet-V2 [31].

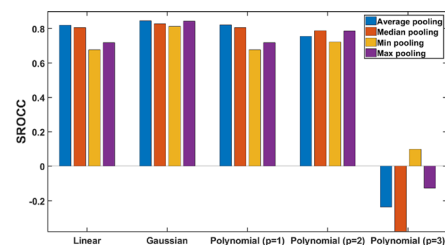


(b) SROCC 5-fold cross-validation with 10 repetitions. Transfer learning was not carried out on Inception-ResNet-V2 [31].

**Fig. 6** Pooling technique and SVR comparison trained and tested on KoNViD-1k [9] using Inception-ResNet-V2 [31] base architecture



(a) PLCC 5-fold cross-validation with 10 repetitions. Transfer learning was carried out on Inception-ResNet-V2 [31].



(b) SROCC 5-fold cross-validation with 10 repetitions. Transfer learning was carried out on Inception-ResNet-V2 [31].

**Fig. 7** Pooling technique and SVR comparison trained and tested on KoNViD-1k [9] using Inception-ResNet-V2 [31] base architecture

with 3rd order polynomial kernel function. Moreover, in most cases average pooling was the best choice, except in one case where max pooling was the best option. Subsequently, we compared the four best methods with state-of-the-art NR-VQA techniques.

## 4.2 Comparison to the State-of-the-Art

We compared seven state-of-the-art NR-VQA methods with our architectures. First, the algorithms were assessed based on KoNViD-1k [9] by fivefold cross-validation in a similar manner to the study by [18]. The PLCC and SROCC values for five baseline methods (Video BLIINDS [24], VIIDEO [19], Video CORNIA [40], FC Model [17], and STFC Model [18]) were those measured by Men et al. [9] and [18], while the results of STS-MLP [43] and STS-SVR [43] were taken from their original publication. The proposed architectures were also assessed based on all the videos in LIVE VQA [28] but without cross-validation because they were trained based on KoNViD-1k [9] and we wanted to demonstrate the generalizability of the proposed method. The PLCC and SROCC values for the baseline methods were those reported in the original studies.

Table 1 shows the comparisons with other state-of-the-art algorithms, which demonstrate that our architecture could also achieve state-of-the-art results without transfer learning. In addition, our fine-tuned CNN-based architectures performed significantly better than the state-of-the-art algorithms. In particular, the PLCC and SROCC values both improved by approximately 0.1. The scatter plots showing the ground-truth MOS values versus the predicted MOS values are depicted in Fig. 8.

For completeness, we also performed a comparison based on the widely used LIVE VQA database [28], which unlike KoNViD-1k [9] contains artificially distorted video sequences. Furthermore, LIVE VQA contains several videos with length of 20 s. On the other hand, KoNViD-1k typically consists of videos with length of 8 s. This difference between the two databases was essentially a serious limiting factor considering our temporally pooled video-level feature vectors. In spite of this, the results demonstrated that our architecture could obtain state-of-the-art results on the LIVE VQA database [28], although it was not employed as the training set. As shown in Table 1, Video CORNIA obtained the best performance with LIVE VQA, and it performed better than our best proposed method by 0.06 in terms of PLCC and 0.045 in terms of SROCC. It should be noted that except for the FC model [17] and the STFC model [18], the previous methods were trained with or optimized for artificially distorted sequences, which explains why the ranking of the methods was different based on KoNViD-1k [9] and LIVE VQA [28]. However, our method still obtained state-of-the-art results on LIVE VQA [28]. Therefore, the experimental results confirmed the effectiveness and generalizability of the proposed approach for NR-VQA.

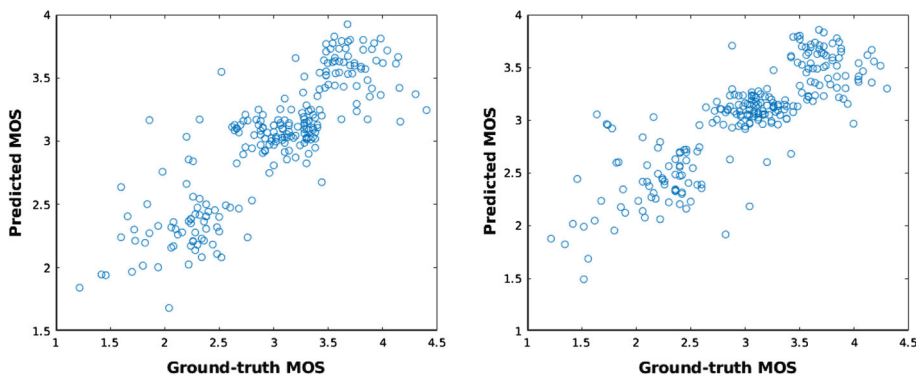
## 4.3 Implementation Details

The introduced algorithm was implemented in MATLAB R2018b mainly relying on the functions of Deep Learning Toolbox (formerly Neural Network Toolbox) and Statistics and Machine Learning Toolbox. Furthermore, it was trained and tested on a personal computer containing 8-core i7-7700K CPU and an NVidia Titan X GPU. In this environment, the evaluation of a video from KoNViD-1k (length is 7 or 8 s) lasts for on average 13.105–13.323 s from which the loading of network the trained network to the GPU lasts for 1.8 s, 11.3 s is the frame-level feature extraction, video-level feature vector compilation takes 0.003 s, and SVR regression takes 0.002–0.22 s depending on the applied kernel function.

**Table 1** Comparison to state-of-the-art NR-VQA algorithms applied on KoNViD-1k [9] and LIVE VQA [28] databases

	KoNViD-1k [9]		LIVE VQA [28]	
	PLCC	SROCC	PLCC	SROCC
Video BLIINDS [24]	0.565	0.572	0.752	0.737
VIIDEO [19]	−0.015	0.031	0.651	0.624
Video CORNIA [40]	0.747	0.765	<b>0.768</b>	<b>0.740</b>
FC Model [17]	0.492	0.472	—	—
STFC Model [18]	0.639	0.606	—	—
STS-MLP [43]	0.407	0.420	—	—
STS-SVR [43]	0.680	0.673	—	—
Inception-V3 + max pooling (*)	0.717	0.676	0.554	0.502
Inception-ResNet-V2 + average pooling (*)	0.707	0.678	0.548	0.499
Inception-V3 + average pooling	<b>0.853</b>	<b>0.849</b>	0.718	0.705
Inception-ResNet-V2 + average pooling	0.833	0.846	0.704	0.696

The correlation values for KoNViD-1k [9], except those of our method, were measured by Men et al. [18] or were taken from the literature. The correlation values for LIVE VQA [28] database, except those of our method, were taken from the literature. We indicated by ‘—’ if the data is not available. The (\*) marking denotes that transfer learning was not carried out on the CNN. The best value is typed bold



(a) Scatter plot showing the ground-truth MOS values against the predicted MOS values based on the KoNViD-1k [9] test set. Inception-V3 + average pooling using Gaussian SVR architecture.

(b) Scatter plot showing the ground-truth MOS values against the predicted MOS values based on the KoNViD-1k [9] test set. Inception-ResNet-V2 + average pooling using Gaussian SVR architecture.

**Fig. 8** Scatter plots showing the ground-truth MOS values against the predicted MOS values

## 5 Conclusions

In this study, we developed a novel framework for NR-VQA based on the features obtained from pretrained CNNs (Inception-V3 [32] and Inception-ResNet-V2 [31]), transfer learning, temporal pooling, and regression. The main novel aspect and contribution of this study is that we developed a possible architecture for NR-VQA that depends on temporally pooled frame-level deep feature vectors and it does not require manually derived features. Furthermore, we showed that the deep features extracted from a fine-tuned, pretrained CNN can

provide effective and rich representations for video quality tasks. Thus, our architecture can be considered as a proof of concept regarding the successful application of deep features extracted from pretrained CNNs in NR-VQA. Our approach was trained and tested based on KoNViD-1k, which is a natural video quality database containing 1200 sequences with quality scores, and it performed better than the best state-of-the-art solution by approximately 0.1 in terms of both the PLCC and SROCC. Our method was also tested with the LIVE VQA database and it achieved state-of-the-art results, although the best state-of-the-art technique performed slightly better.

**Acknowledgements** Open access funding provided by Budapest University of Technology and Economics (BME). The author would like to show his gratitude to Professor Dietmar Saupe, Hanhe Lin, Vlad Hosu, Franz Hahn, Hui Men, and Mohsen Jenadeleh for sharing their knowledge in visual quality assessment and helping the author to use KoNViD-1k. The author would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Anekekuh L, Sun L, Jammeh E, Mkwawa IH, Ifeakor E (2015) Content-based video quality prediction for HEVC encoded videos streamed over packet networks. *IEEE Trans Multimed* 17(8):1323–1334
2. Bianco S, Celona L, Napoletano P, Schettini R (2018) On the use of deep learning for blind image quality assessment. *Signal Image Video Process* 12(2):355–362
3. Borer S (2010) A model of jerkiness for temporal impairments in video transmission. In: 2010 second international workshop on quality of multimedia experience (QoMEX), pp 218–223
4. Brandao T, Queluz MP (2010) No-reference quality assessment of H. 264/AVC encoded video. *IEEE Trans Circuits Syst Video Technol* 20(11):1437–1447
5. Ghadiyaram D, Bovik AC (2014) Blind image quality assessment on real distorted images using deep belief nets. In: 2014 IEEE global conference on signal and information processing (GlobalSIP), pp 946–950
6. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 249–256
7. Hong C, Chen X, Wang X, Tang C (2016) Hypergraph regularized autoencoder for image-based 3d human pose recovery. *Signal Process* 124:132–140
8. Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoencoder for human pose recovery. *IEEE Trans Image Process* 24(12):5659–5670
9. Hosu V, Hahn F, Jenadeleh M, Lin H, Men H, Szirányi T, Li S, Saupe D (2017) The Konstanz natural video database (KoNViD-1k). In: 2017 Ninth international conference on quality of multimedia experience (QoMEX), pp 1–6
10. Huynh BQ, Li H, Giger ML (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging* 3(3):034501
11. Huynh-Thu Q, Garcia MN, Speranza F, Corriveau P, Raake A (2011) Study of rating scales for subjective quality assessment of high-definition video. *IEEE Trans Broadcast* 57(1):1–14
12. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1733–1740
13. 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 Video Process* 10(4):609–616
14. Li X, Guo Q, Lu X (2016) Spatiotemporal statistics for video quality assessment. *IEEE Trans Image Process* 25(7):3329–3342
15. Li Y, Po LM, Cheung CH, Xu X, Feng L, Yuan F, Cheung KW (2016) No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans Circuits Syst Video Technol* 26(6):1044–1057

16. Lin JY, Song R, Wu CH, Liu T, Wang H, Kuo CCJ (2015) MCL-V: a streaming video quality assessment database. *J Vis Commun Image Represent* 30:1–9
17. Men H, Lin H, Saupe D (2017) Empirical evaluation of no-reference VQA methods on a natural video quality database. In: 2017 Ninth international conference on quality of multimedia experience (QoMEX), pp 1–3
18. Men H, Lin H, Saupe D (2018) Spatiotemporal feature combination model for no-reference video quality assessment. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp 1–3
19. Mittal A, Saad MA, Bovik AC (2016) A completely blind video integrity oracle. *IEEE Trans Image Process* 25(1):289–300
20. Moorthy AK, Choi LK, Bovik AC, Veciana GD (2012) Video quality assessment on mobile devices: subjective, behavioral and objective studies. *IEEE J Sel Top Signal Process* 6(6):652–671
21. Muijs R, Kirenko I (2005) A no-reference blocking artifact measure for adaptive video processing. In: 2005 13th European signal processing conference, pp 1–4
22. Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. *Netw Comput Neural Syst* 10(4):341–350
23. Saad MA, Bovik AC, Charrier C (2011) DCT statistics model-based blind image quality assessment. In: 2011 18th IEEE international conference on image processing (ICIP), pp 3093–3096
24. Saad MA, Bovik AC, Charrier C (2014) Blind prediction of natural video quality. *IEEE Trans Image Process* 23(3):1352–1365
25. Saupe D, Hahn F, Hosu V, Zingman I, Rana M, Li S (2016) Crowd workers proven useful: a comparative study of subjective video quality assessment. In: QoMEX 2016: 8th international conference on quality of multimedia experience
26. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
27. Scott MJ, Guntuku SC, Lin W, Ghinea G (2016) Do personality and culture influence perceived video quality and enjoyment? *IEEE Trans Multimed* 18(9):1796–1807
28. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. *IEEE Trans Image Process* 19(6):1427–1441
29. Shahid M, Rossholm A, Löfström B, Zepernick HJ (2014) No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP J Image Video Process* 2014(1):40
30. Sinno Z, Bovik AC (2019) Large-scale study of perceptual video quality. *IEEE Trans Image Process* 28(2):612–627
31. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol 4, p 12
32. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
33. Tan KT, Ghanbari M (2000) Blockiness detection for MPEG2-coded video. *IEEE Signal Process Lett* 7(8):213–215
34. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2015) YFCC100M: the new data in multimedia research. arXiv preprint [arXiv:1503.01817](https://arxiv.org/abs/1503.01817)
35. Varga D, Szirányi T (2016) Fast content-based image retrieval using convolutional neural network and hash function. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC), pp 002636–002640
36. Vega MT, Mocanu DC, Famaey J, Stavrou S, Liotta A (2017) Deep learning for quality assessment in live video streaming. *IEEE Signal Process Lett* 24(6):736–740
37. Vega MT, Sguazzo V, Mocanu DC, Liotta A (2016) An experimental survey of no-reference video quality assessment methods. *Int J Pervasive Comput Commun* 12(1):66–86
38. Vlachos T (2000) Detection of blocking artifacts in compressed video. *Electron Lett* 36(13):1106–1108
39. Wu Y, Cao N, Gotz D, Tan YP, Keim DA (2016) A survey on visual analytics of social media data. *IEEE Trans Multimed* 18(11):2135–2148
40. Xu J, Ye P, Liu Y, Doermann D (2014) No-reference video quality assessment via feature learning. In: 2014 IEEE international conference on image processing (ICIP), pp 491–495
41. Xu L, Lin W, Kuo CCJ (2015) Visual quality assessment by machine learning. Springer, Berlin
42. Xue Y, Erkin B, Wang Y (2015) A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing. *IEEE Trans Multimed* 17(1):134–139
43. Yan P, Mou X (2018) No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images. In: Optoelectronic imaging and multimedia technology V, vol 10817. International Society for Optics and Photonics, p 108171D

44. Yang M, Liu Y, You Z (2017) The euclidean embedding learning based on convolutional neural network for stereo matching. *Neurocomputing* 267:195–200
45. Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: 2012 IEEE conference on computer vision and pattern recognition, pp 1098–1105
46. Zhang Y, Gao X, He L, Lu W, He R (2018) Blind Video Quality Assessment with Weakly Supervised Learning and Resampling Strategy. *IEEE Trans Circuits Syst Video Technol.* <https://doi.org/10.1109/TCSVT.2018.2868063>
47. Zhu K, Asari V, Saupe D (2013) No-reference quality assessment of H. 264/AVC encoded video based on natural scene features. In: *Mobile multimedia/image processing, security, and applications*, vol 8755, p 875505

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.