ORIGINAL PAPER

No-reference video quality assessment via pretrained CNN and LSTM networks

Domonkos Varga[1] · **Tamás Szirányi[2]**

Abstract

A general-purpose no-reference video quality assessment algorithm based on a long short-term memory (LSTM) network and a pretrained convolutional neural network (CNN) is introduced. Considering video sequences as a time series of deep features extracted with the help of a CNN, an LSTM network is trained to predict subjective quality scores. In contrast to previous methods, the resulting algorithm was trained on the recently published Konstanz Natural Video Quality Database (KoNViD-1k), which is the only publicly available database that contains sequences with authentic distortions. The results of experiments on KoNViD-1k demonstrate that the proposed method outperforms other state-of-the-art algorithms. Furthermore, these results are also confirmed using tests on the LIVE Video Quality Assessment Database, which consists of artificially distorted videos.

Keywords No-reference video quality assessment · Long short-term memory · Convolutional neural network

1 Introduction

In recent years, we have witnessed an explosive growth in the spread of multimedia technologies and digital visual content. With the increasing popularity of smart phones, social media, and video-sharing applications, digital videos are increasingly captured, transmitted, stored, shared, compressed, or edited. However, these transformations usually affect the perceived visual quality of videos. Furthermore, humans are the end consumers of digital video content whose quality requirements have to be satisfied. This has motivated video service providers and the research community to devise quality assessment methods for digital videos.

Apparently, perceived video quality relates to the visual stimuli received by the human visual system (HVS). Although huge amounts of research have been conducted to reveal the psychological and physiological mechanisms of the HVS,

it is not yet understood from all aspects. Thus, machine learning techniques have been employed extensively in this field. The most accurate and reliable method of assessing the quality of digital videos is through subjective evaluation [1]. Several international standards such as ITU P913 [22] for performing subjective video quality assessment (VQA) have been published. The main objective of subjective VQA is to collect subjective quality scores from users for each digital video from a given set. Finally, the mean opinion score (MOS) of each video is determined by averaging the individual quality ratings. However, subjective VQA has apparent drawbacks that restrict its application in real-world services. First, they are time-consuming and expensive because subjective results are obtained through experiments with many observers. Consequently, they cannot be part of real-time applications such as video transmission systems. Second, their results depend on the observers' physical conditions, personality, and emotional state [26]. Therefore, the development of objective VQA methods that are able to predict the perceptual quality of visual signals is essential.

The goal of objective VQA is to design mathematical models that are able to predict the quality of a video assessed by humans. According to the availability of reference videos, VQA methods can be divided into three groups: full-reference (FR-VQA), reduced-reference (RR-VQA), and no-reference (NR-VQA) algorithms.

✉ Domonkos Varga
   varga.domonkos7@upcmail.hu

   Tamás Szirányi
   sziranyi.tamas@sztaki.mta.hu

[1] Department of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, Hungary

[2] Machine Perception Laboratory, MTA SZTAKI, Institute for Computer Science and Control, Budapest, Hungary

Artificial intelligence and machine learning methods and algorithms are widely used in NR-VQA methods. Recently, deep learning techniques have become standard tools for many image processing and computer vision tasks. Furthermore, features extracted from pretrained deep convolutional neural networks (CNNs) have proved very effective in a broad range of applications, ranging from content-based image retrieval [2] to medical image analysis [13]. In this paper, we make the following contributions. In our proposed NR-VQA framework, we model a digital video sequence as a sequence of data of frame-level deep features extracted via pretrained CNNs. These sequence data are fed into a long short-term memory (LSTM) network containing LSTM layers and a fully connected (FC) layer to perform sequence-to-one regression. In other words, the main novelty of the presented architecture is that video sequences are considered as time series of deep features that are utilized by an LSTM network [8] to learn long-term dependencies for perceptual quality prediction. Owing to the memory cells applied in an LSTM, long-range temporal relationships that may also be useful in NR-VQA can be discovered effectively. LSTM networks are widely used to classify [7], process [31], or make predictions [34] using time series or sequential data. Unlike other LSTM applying NR-VQA methods [3,21], we model video sequences as sequential data of frame-level deep features and not employing image quality-related metrics at all. Consequently, the dimension of the sequence data used to train the LSTM network is many times larger, allowing us to exploit the effectiveness of CNN-extracted features. In contrast to previous deep-learning-based architectures [15,36,40], we rely only on features extracted from a pretrained CNN. Furthermore, to the best of the authors' knowledge, this is the first deep architecture that was trained on a natural video quality database. Previous works were trained on databases containing artificially distorted video sequences derived from 6–45 pristine videos, which limited their applications in authentic environments. On the other hand, our approach was trained on the recently published Konstanz Natural Video Quality Database (KoNViD-1k) [9], which contains 1200 unique video sequences with authentic distortions.

2 Related and previous work

NR-VQA methods can be classified into two groups. *Distortion-specific* NR-VQA algorithms employ specific distortion models to predict the subjective quality; however, they can measure only a few distortions such as blurriness [5], H.264 compression [41], MPEG-2 compression [27], and jerkiness [4], whereas *general-purpose* (or *non-distortion-specific*) methods perform across various types of distortions. The performance of NR-VQA methods is rapidly advancing,

and there is a proliferation of NR-VQA metrics. Soundararajan and Bovik [30] gave a systematic review of visual quality metrics, whereas Shadid et al. [29] presented an overview of NR visual quality assessment algorithms. Xu et al. [16] covered the role of machine learning in visual quality assessment.

A popular feature extraction method originates from natural scene statistics (NSS), which relies on the premise that HVS has evolved via natural selection and, as a result, it inherently contains knowledge regarding the regularities of the physical reality surrounding us. Consequently, the statistical regularities of visual signals are apparently influenced by quality degradation. Saad et al. [23] devised a spatiotemporal model that combined the discrete cosine transform (DCT) model with a motion model. As a result, it was possible to quantify motion coherency to predict perceptual video quality. Later, this approach was extended to the 3D DCT domain by Xuelong et al. [14] using spatial and temporal information. Similarly, Konuk et al. [11] presented a spatiotemporal model, but they utilized bit rate and packet loss as features.

Motivated by the success of CORNIA [39] NR image quality assessment method, Xu et al. [37] presented an opinion-unaware architecture for NR-VQA, the so-called *Video CORNIA*. In particular, frame-level features are extracted via unsupervised feature learning and applied a support vector regressor (SVR) to map these onto subjective quality scores. Similarly, Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) [20] does not require human ratings on video quality. Namely, it has been assumed that pristine video sequences possess intrinsic statistical regularities and the deviation from them can be used to predict perceptual quality scores. The main idea was that local statistics of frame differences derived using mean removal and divisive contrast normalization should follow a generalized Gaussian distribution in the case of good video quality.

In contrast to previous work, Men et al. [17] introduced an NR-VQA method that was trained using a natural video quality database, KoNViD-1k [9], which consists of 1200 unique video sequences with authentic distortions. In particular, a video-level feature vector was compiled by combining multiple features, such as blurriness, colorfulness, contrast, and spatial and temporal information. The video-level feature vectors were mapped to subjective quality scores with an SVR. Later, this model was developed significantly [18] by combining spatial and temporal information more intensively.

Another line of methods focuses on the use of deep learning techniques. The method of Li et al. [15] divided the input video sequence into blocks and with the help of 3D shearlet transform features were extracted. Based on these feature vectors, CNN and logistic regression were applied to predict video quality. Similarly, the algorithm of Zhang et al. [40] also divided the input video into blocks, but the corre-
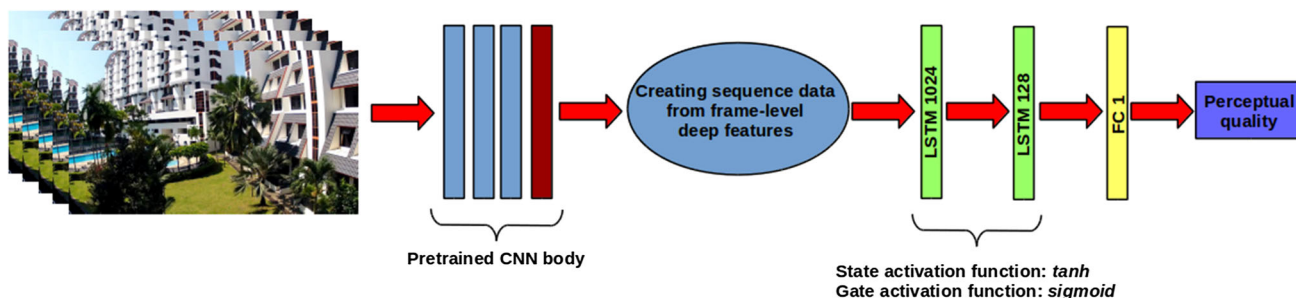
State activation function: *tanh*
Gate activation function: *sigmoid*

Fig. 1 High-level overview of the proposed NR-VQA algorithm. A pretrained CNN is run through all consecutive video frames to create $d \times N$ sequence data where d stands for the length of the video sequence and N is the length of the frame-level deep feature vector. Subsequently, an LSTM network is utilized to predict subjective quality scores

sponding weak labels were derived by an FR-VQA metric. Subsequently, a CNN was trained with the weak labeled data. Furthermore, a resampling strategy was applied to generate a regression function that mapped deep features onto quality scores. In contrast, Torres Vega et al. [36] trained a restricted Boltzmann machine (RBM) with lightweight NR metrics, such as the noise ratio and motion intensity. The experimental results were presented on live video streams.

3 Methods

In this paper, we propose a CNN- and LSTM-network-based NR-VQA algorithm. The high-level overview of the algorithm is depicted in Fig. 1. For a given video sequence to be evaluated, frame-level deep features are extracted from all consecutive resized and center-cropped video frames with the help of a pretrained CNN. In this study, we report on the results of three different pretrained CNNs, *i.e.,* AlexNet [12], Inception-V3 [33], and Inception-ResNet-V2 [32]. Owing to the fixed input size, the consecutive video frames were resized to 338×338 and 299×299 center patches were cropped, when Inception-V3 or Inception-ResNet-V2 was applied. On the other hand, the frames were resized to 256×256 and 227×227 center patches were cropped, when AlexNet was applied. As an LSTM network accepts sequence data as input, the chosen pretrained CNN is run through each resized and center-cropped video frame. The corresponding frame-level feature vector is obtained by removing the last softmax and the last fully connected layer. The length of the feature vector is 4096 for AlexNet, 2048 for Inception-V3, and 1536 for Inception-ResNet-V2. Consequently, this process results in a $d \times N$ matrix of features where d is the length of the video sequence and N is the length of the corresponding deep feature vector. Subsequently, this feature matrix is transferred to an LSTM network to predict perceptual quality.

The remainder of this section is organized as follows. Section 3.1 presents the compilation of the training and test database. Section 3.2 deals with transfer learning, which is conducted on the pretrained CNN. Section 3.3 presents the training of the LSTM network.

3.1 Database compilation

In our work, we chose KoNViD-1k [9] from publicly available video quality databases to train and test our architecture. In contrast to other publicly available datasets, KoNViD-1k consists of 1200 video sequences—more than any other. The large number of video sequences allowed us to train an LSTM network directly with deep features. Furthermore, the sequences have authentic distortions and were sampled from Yahoo Flickr Creative Commons 100 Million [35] (YFCC100m) and the quality scores were collected online [24] using CrowdFlower platform. The spatial resolution is 960×540 in this database, while the length of the sequences is approximately 9 s with 30 fps.

A total of 960 sequences were selected randomly for training purposes, while the remaining videos were kept only for testing. The training videos were split into frames, and then 20% of them were taken randomly. In order to fit to the Inception-V3's [33] and Inception-ResNet-V2's [32] input size, the randomly selected video frames were resized to 338×338 and 299×299 center patches were cropped. As already mentioned, for AlexNet [12] base architecture these values were 256×256 and 227×227. The resulting training images inherited the MOS values of their source videos. Consequently, we made the assumption that the visual quality perception of individual frames is somehow related to those of the complete video sequence. On the whole, the resulting image database consists of 43,320 images which were used to carry out transfer learning on the chosen pretrained CNN. To this end, as already mentioned different pretrained CNNs were applied in this paper.

For the sake of completeness, we selected LIVE VQA database [28] as an additional test set in order to analyze the generalization capability of the proposed algorithm. LIVE
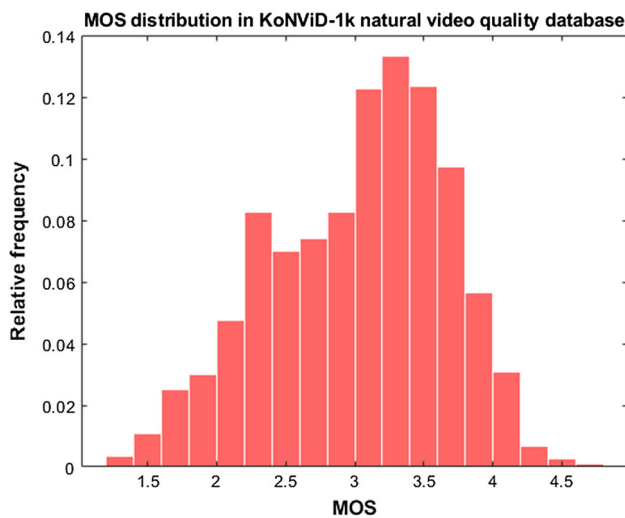
Fig. 2 MOS distribution in KoNViD-1k [9]

VQA contains 15 reference videos and 150 artificially distorted video sequences derived from the reference videos using four different distortion types: 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 videos' spatial resolution in LIVE VQA is 768×432.

3.2 Transfer learning

In general, transfer learning is applied to transfer stored knowledge gained by a model trained on a previous task to a new task. It is typically used if the amount of labeled training data is insufficient to train a CNN from scratch or a pretrained CNN exists for a similar task. In our work, the common practice was applied to transfer learning. First, the last 1000-way softmax layer was cut and it was replaced by a 5-way softmax layer relevant to our problem. Five classes in our training set were defined: class A for excellent image quality (MOS $\in [4.2, 5.0]$), class B for good image quality (MOS $\in [3.4, 4.2[$), class C for fair image quality (MOS $\in [2.6, 3.4[$), class D for poor image quality (MOS $\in [1.8, 2.6[$), and class E for very poor image quality (MOS $\in [1.0, 1.8[$). The initial learning rate was 0.0001, and it was 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 is trained from scratch utilizing Xavier initialization [6], while the initial weights of the other layers come from the corresponding layers of the pretrained networks and all layers are updated using the back-propagation algorithm [25]. As shown in Fig. 2, the MOS distribution in KoNViD-1k [9] is imbalanced. This could cause problems in transfer learning. That is why each instance is sampled in the batch by

the inverse frequency of the class. In consequence, instances in larger classes have smaller probability to be selected. Due to population differences of the classes, the final batch will be equally distributed. Figure 3 plots the training process of transfer learning on training database described above.

3.3 Training of LSTM layers and quality regression

As already mentioned, an LSTM network accepts sequence data as input and the dimension of a feature matrix is $d \times N$ where d is the length of the corresponding video sequence and N stands for the length of the frame-level deep feature vector (4096 for AlexNet, 2048 for Inception-V3, and 1536 for Inception-ResNet-V2). During training, the training data are split into mini-batches and we pad the sequences in order to have the same length. However, too much padding deteriorates the performance of an LSTM network. To reduce the amount of padding, the training data are sorted by the video sequence length and the mini-batch size was set to 27. In consequence, sequences in a mini-batch have similar length. Furthermore, the LSTM network consists of two LSTM layers with 1024 and 128 hidden units, respectively. Finally, a fully connected layer of size one terminates the structure to predict MOS values. Furthermore, ADAM [10] solver was applied and the gradient threshold was set to 0.5 during training. Mean square error was utilized as regression loss function. Figure 4 depicts the training progress of the LSTM network.

4 Experimental results and analysis

The evaluation of objective video quality assessment is based on the correlation between the predicted and the ground-truth quality scores [16]. Pearson's linear correlation coefficient (PLCC) and Spearman's rank order correlation coefficient (SROCC) are widely applied to this end. The PLCC between data set A and B is defined as

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

where $\overline{A}$ and $\overline{B}$ stand for the average of set A and B, A_i and B_i denote the ith element of set A and B, respectively. For two ranked sets A and B, SROCC is defined as

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

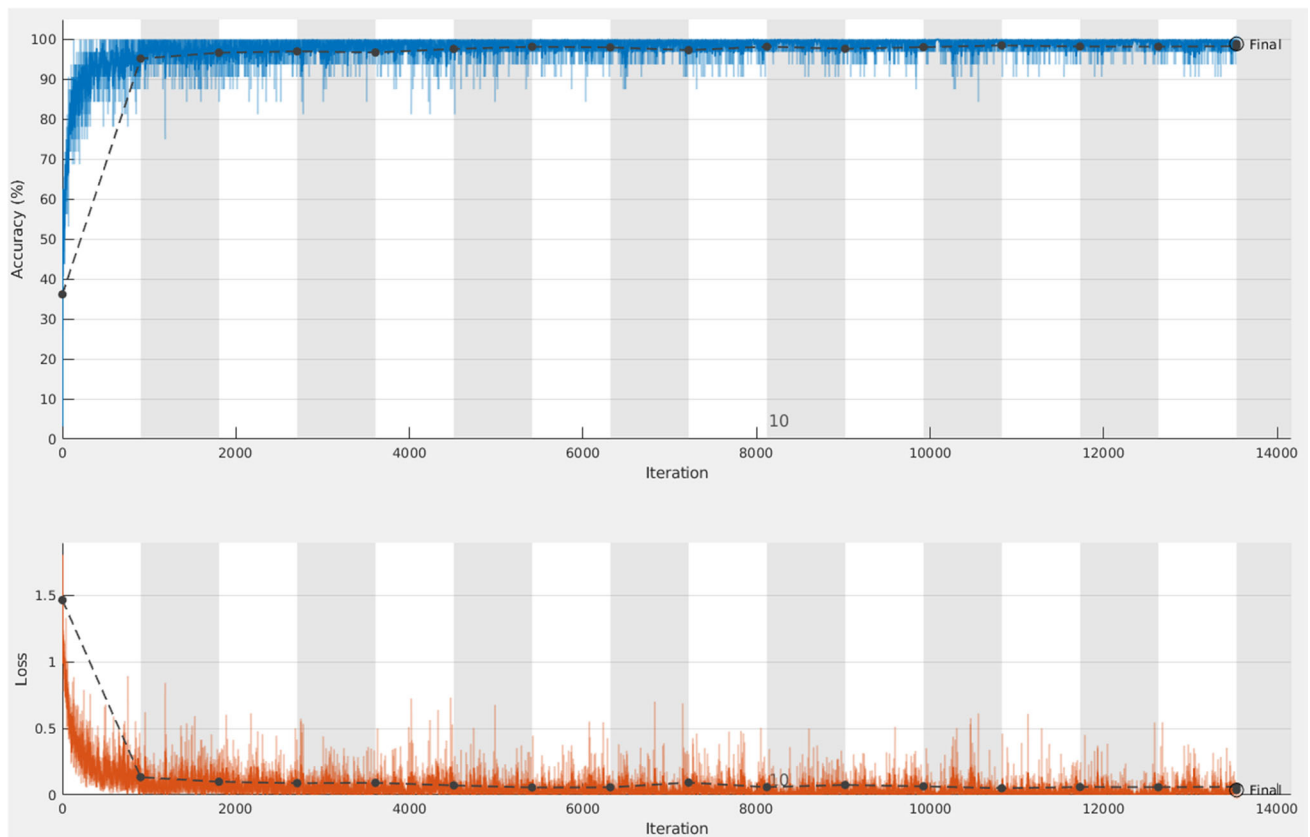where $\hat{A}$ and $\hat{B}$ are the middle ranks of set A and B.

Fig. 3 Training progress of Inception-V3 [33] during transfer learning. This figure plots the smoothed training accuracy with dark blue line, the training accuracy with light blue line, the smoothed training loss with orange line, and the training loss with light orange line. Furthermore, the validation accuracy and validation loss are also depicted with dashed lines (color figure online)

4.1 Parameter study

First, we evaluated the design choices of our proposed method on KoNViD-1k [9], before comparing it with other state-of-the-art NR-VQA techniques. We evaluated our algorithm using fivefold cross-validation and report on median PLCC and SROCC values like Men et al. [18] and Yan et al. [38]. First of all, the effects of the applied pretrained CNNs and transfer learning were evaluated. Figure 5 summarizes the results of the parameter study. Specifically, the results showed that Inception-V3's [33] features gave slightly better results than Inception-ResNet-V2's [32] features. Furthermore, AlexNet's [12] features performed significantly poorer than the previous two CNNs. Our analysis also demonstrated that fine-tuning on the target database enormously improves the prediction's quality. In the following, we denote by CNN+LSTM our best model.

4.2 Comparison with the state of the art

Eight state-of-the-art NR-VQA methods are compared with our proposed algorithm. All methods were evaluated using fivefold cross-validation with 10 random train–validation–test split, and median PLCC and SROCC values are reported as proposed in [17] and [18]. The median PLCC and SROCC values of five baseline methods (Video BLIINDS [23], VIIDEO [20], Video CORNIA [37], FC Model [17], and STFC Model [18]) were measured by Men et al. in [17] and [18]. On the other hand, the results of STS-MLP [38] and STS-SVR [38] were taken from their original publication because their authors also report on median PLCC and SROCC values using fivefold cross-validation with 10 random train–validation–test split. Furthermore, we retrained the NVIE method [19] on KoNViD-1k (80% of videos for training and 20% for testing) and evaluated it using the above-mentioned methodology. As a consequence, the fairness of comparison is assured because the evaluation methodology is exactly the same. The proposed architecture was also assessed on all videos of LIVE VQA [28] without any cross-validation because it was trained on KoNViD-1k [9]. State-of-the-art methods' PLCC and SROCC values for LIVE VQA were taken from their original publications.

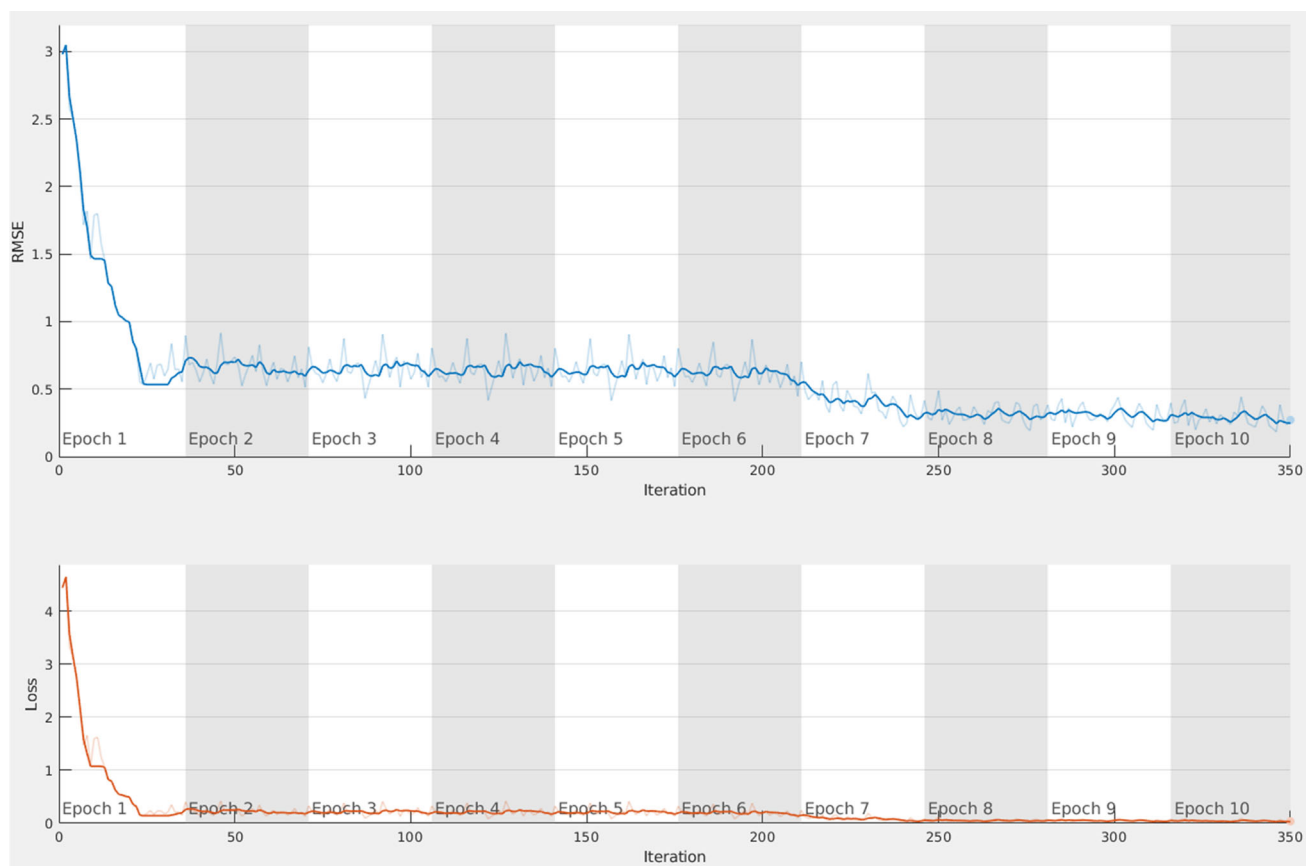The results are summarized in Table 1. From these results, it can be concluded that the proposed method is able to

Fig. 4 Training progress of the LSTM network. This figure plots the smoothed root-mean-square error (RMSE) with dark blue line, the RMSE with light blue line, the smoothed training loss with orange line, and the training loss with light orange line (color figure online)
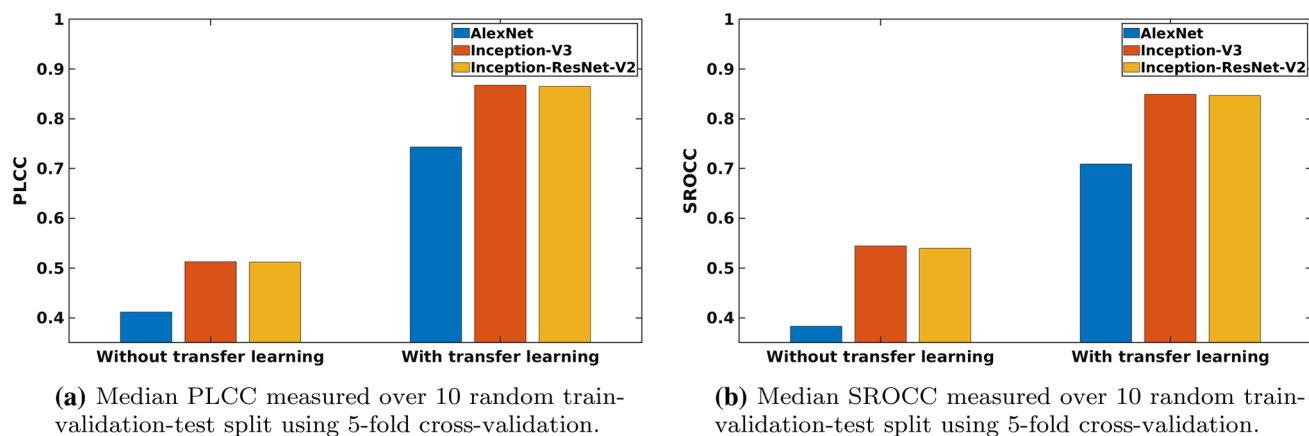

(a) Median PLCC measured over 10 random train-validation-test split using 5-fold cross-validation.

(b) Median SROCC measured over 10 random train-validation-test split using 5-fold cross-validation.

Fig. 5 Parameter study. Trained and tested on KoNViD-1k [9] using AlexNet [12], Inception-V3 [33], and Inception-ResNet-V2 [32] as base architectures

achieve state-of-the-art results without transfer learning as well. On the other hand, with transfer learning our algorithm significantly outperforms the state of the art on KoNViD-1k [9]. Specifically, we could improve both PLCC and SROCC by approximately 0.1 compared to the best proposal in the literature. A scatter plot of the ground-truth MOS against the predicted MOS is depicted in Fig. 6. As regards LIVE

VQA, our method was outperformed by approximately 0.05 in PLCC and SROCC by the best algorithm. Please note that previous methods except for FC Model [17] and STFC Model [18] were trained on or optimized for artificially distorted video sequences. That is why the results on the two different databases can be radically different. In spite of this, the proposed method is able to achieve state-of-the-art results on

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
V. BLIINDS [23]	0.565	0.572	0.752	0.737
VIIDEO [20]	− 0.015	0.031	0.651	0.624
V. CORNIA [37]	0.747	0.765	**0.768**	**0.740**
NVIE [19]	0.403	0.332	0.693	0.629
FC Model [17]	0.492	0.472	–	–
STFC Model [18]	0.639	0.606	–	–
STS-MLP [38]	0.407	0.420	–	–
STS-SVR [38]	0.680	0.673	–	–
CNN(*) + LSTM	0.513	0.545	0.347	0.365
CNN + LSTM	**0.867**	**0.849**	0.691	0.703

We indicated by '–' if the data are not available. The (*) marking denotes that transfer learning was not carried out. The best value is typed bold
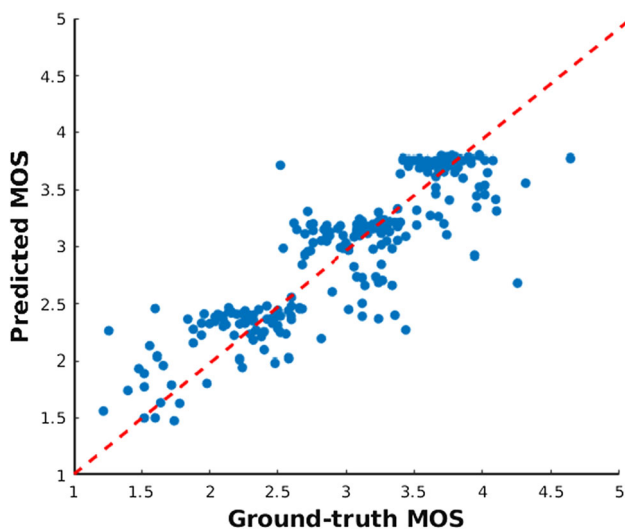
Fig. 6 Scatter plot of the ground-truth MOS against the predicted MOS on KoNViD-1k [9] test set

LIVE VQA [28] as well. Therefore, the experimental results confirmed the effectiveness and generalization capability of the proposed approach for NR-VQA.

5 Conclusions

In this paper, we have introduced a novel architecture for NR-VQA utilizing deep features extracted from a pretrained CNN and LSTM network for sequence-to-one regression. The main novelty was that video sequences were considered as time series of deep features and an LSTM network was applied to learn long-term dependencies for perceptual quality prediction. Unlike previous methods, our work relies only on deep features and does not use handcrafted features at all.

The large number of videos with authentic distortions found in KoNViD-1k [9] allowed us to build a purely data-driven model. The presented algorithm outperformed the best solution in the state of the art by approximately 0.1 in terms of both PLCC and SROCC on KoNViD-1k [9]. Our method was further tested on LIVE VQA [28] where it achieved the state-of-the-art results and was slightly outperformed by the best method in the state of the art.

References

1. Aflaki, P., Hannuksela, M.M., Gabbouj, M.: Subjective quality assessment of asymmetric stereoscopic 3D video. Signal Image Video Process. **9**(2), 331–345 (2015)
2. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: IEEE International Conference on Computer Vision, pp. 1269–1277 (2015)
3. Bampis, C.G., Li, Z., Katsavounidis, I., Bovik, A.C.: Recurrent and dynamic models for predicting streaming video quality of experience. IEEE Trans. Image Process. **27**(7), 3316–3331 (2018)
4. Borer, S.: A model of jerkiness for temporal impairments in video transmission. In: International Workshop on Quality of Multimedia Experience, pp. 218–223 (2010)
5. Dardi, F., Abate, L., Ramponi, G.: No-reference measurement of perceptually significant blurriness in video frames. Signal Image Video Process. **5**(3), 271–282 (2011)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The Konstanz natural video database (KoNViD-1k). In: International Conference on Quality of Multimedia Experience, pp. 1–6 (2017)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Konuk, B., Zerman, E., Nur, G., Akar, G.B.: A spatiotemporal no-reference video quality assessment model. In: International Conference on Image Processing, pp. 54–58 (2013)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

13. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: Conference on Computer and Robot Vision, pp. 133–138 (2015)

14. Li, X., Guo, Q., Lu, X.: Spatiotemporal statistics for video quality assessment. IEEE Trans. Image Process. **25**(7), 3329–3342 (2016)

15. Li, Y., Po, L.M., Cheung, C.H., Xu, X., Feng, L., Yuan, F., Cheung, K.W.: No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Trans. Circuits Syst. Video Technol. **26**(6), 1044–1057 (2016)

16. Long, X., Lin, W., Kuo, C.C.J.: Visual Quality Assessment by Machine Learning. SpringerBriefs in Signal Processing, vol. 2015. Springer, Berlin (2015)

17. Men, H., Lin, H., Saupe, D.: Empirical evaluation of no-reference VQA methods on a natural video quality database. In: International Conference on Quality of Multimedia Experience, pp. 1–3 (2017)

18. Men, H., Lin, H., Saupe, D.: Spatiotemporal feature combination model for no-reference video quality assessment. In: International Conference on Quality of Multimedia Experience, pp. 1–3 (2018)

19. Mittal, A.: Natural scene statistics-based blind visual quality assessment in the spatial domain. DISSERTATION Presented to the Faculty of the Graduate School of The University of Texas at Austin. http://hdl.handle.net/2152/22015 (2013)

20. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Trans. Image Process. **25**(1), 289–300 (2016)

21. Mohamed, S., Rubino, G.: A study of real-time packet video quality using random neural networks. IEEE Trans. Circuits Syst. Video Technol. **12**(12), 1071–1083 (2002)

22. RECOMMENDATION, P.I.T.: Methods for the subjective assessment of video quality, audio quality and audivisional quality of Internet video and distribution quality television in any environment. https://www.itu.int/rec/T-REC-P.913/en (2016)

23. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Trans. Image Process. **23**(3), 1352–1365 (2014)

24. Saupe, D., Hahn, F., Hosu, V., Zingman, I., Rana, M., Li, S.: Crowd workers proven useful: a comparative study of subjective video quality assessment. In: International Conference on Quality of Multimedia Experience (2016)

25. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)

26. Scott, M.J., Guntuku, S.C., Lin, W., Ghinea, G.: Do personality and culture influence perceived video quality and enjoyment? IEEE Trans. Multimed. **18**(9), 1796–1807 (2016)

27. Søgaard, J., Forchhammer, S., Korhonen, J.: No-reference video quality assessment using codec analysis. IEEE Trans. Circuits Syst. Video Technol. **25**(10), 1637–1650 (2015)

28. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Trans. Image Process. **19**(6), 1427–1441 (2010)

29. Shahid, M., Rossholm, A., Lövström, B., Zepernick, H.J.: No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP J. Image Video Process. **2014**(1), 40 (2014)

30. Soundararajan, R., Bovik, A.C.: Survey of information theory in visual quality assessment. Signal Image Video Process. **7**(3), 391–401 (2013)

31. Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: Advances in Neural Information Processing Systems, pp. 2998–3006 (2015)

32. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

34. Tax, N., Verenich, I., Rosa, M.L., Dumas, M.: Predictive business process monitoring with LSTM neural networks. In: International Conference on Advanced Information Systems Engineering, pp. 477–492 (2017)

35. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. arXiv preprint arXiv:1503.01817 (2015)

36. Vega, M.T., Mocanu, D.C., Famaey, J., Stavrou, S., Liotta, A.: Deep learning for quality assessment in live video streaming. IEEE Signal Process. Lett. **24**(6), 736–740 (2017)

37. Xu, J., Ye, P., Liu, Y., Doermann, D.: No-reference video quality assessment via feature learning. In: IEEE International Conference on Image Processing, pp. 491–495 (2014)

38. Yan, P., Mou, X.: 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, p. 108171D (2018)

39. Ye, P., Kumar, J., Kang, L., Doermann, D.: Unsupervised feature learning framework for no-reference image quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1098–1105 (2012)

40. Zhang, Y., Gao, X., He, L., Lu, W., He, R.: Blind video quality assessment with weakly supervised learning and resampling strategy. IEEE Transactions on Circuits and Systems for Video Technology. https://ieeexplore.ieee.org/document/8453041 (2018)

41. Zhu, K., Asari, V., Saupe, D.: No-reference quality assessment of H. 264/AVC encoded video based on natural scene features. In: Mobile Multimedia/Image Processing, Security, and Applications 2013, vol. 8755, p. 875505 (2013)