International Journal of Engineering

Journal Homepage: www.ije.ir

Predicting the Empirical Distribution of Video Quality Scores Using Recurrent Neural Networks

H. Otroshi Shahreza*, A. Amini, H. Behroozi

Electrical Engineering Department, Sharif University of Technology, Tehran, Iran

PAPER INFO

Paper history:
Received 10 November 2019
Received in revised form 23 December 2019
Accepted 04 January 2020

Keywords:
Distribution
Feature
No-Reference
Opinion Score
Recurrent Neural Network
Video Quality Assessment

ABSTRACT

Video quality assessment is a crucial routine in the broadcasting industry. Due to the duration and the excessive number of video files, a computer-based video quality assessment mechanism is the only solution. While it is common to measure the quality of a video file at the compression stage by comparing it against the raw data, at later stages, no reference video is available for comparison. Therefore, a no-reference (Blind) video quality assessment (NR-VQA) technique is essential. The current NR-VQA methods predict only the mean opinion score (MOS) and do not provide further information about the distribution of people score. However, this distribution is informative for the evaluation of QoE. In this paper, we propose a method for predicting the empirical distribution of human opinion scores in the assessment of video quality. To the best of our knowledge, this is the first paper to investigate prediction of the distribution of human opinion scores for video quality. To this end, we extract some frame-level features, and next, we feed these features to a recurrent neural network. Finally, the distribution of opinion score is predicted in the last layer of the RNN. The experiments show that averages of predicted distributions have comparable or better results with previous methods on the KonVid-1k dataset.

doi: *10.5829/ije.2020.33.05b.32*

1. INTRODUCTION

The recent growth in social networks shows an increasing interest in video content. According to [1], IP video traffic was 75% of all IP traffic in 2017, and it will grow four-fold by 2022 to be 82% of all IP traffic. This growth contains different video-based applications. For example, Video-on-Demand (VoD) traffic will nearly double, and internet video to TV will increase threefold between 2017 to 2022. Obviously, the perceived quality of video content will affect the QoE in the user end. However, not all of the video files have an acceptable quality level. Depending on the involved processes, various types of distortions might affect the video content, most notably, distortions at the time of recording due to the camera, distortions caused by the compression, and data loss within the transmission channel. In addition, the receiving end rarely has access to the original/reference video file to make a comparison. Therefore, the type and extent of various distortion sources are not available. A no-reference video quality assessment (NR-VQA) algorithm is any method that predicts the quality of a video file according to a quality measure solely based on the available data. However, current NR-VQA algorithms, e.g., [2-8], predict the mean opinion score or shortly MOS and do not provide further information about the distribution of people score. In this paper, we propose a new method to predict people's opinion scores in the assessment of video content.

Generally, predicting the quality of video content without prior information is impossible. In particular, one needs a database of rated video files for training a method. The LIVE video quality database [9] has been very popular in the past for training NR-VQA methods (see, for instance, [2-4]). This dataset includes 4 types of distortions that are artificially applied to source files. In practice, however, multiple distortion types affect a video file. This fact complicates the detection of the distortion

*Corresponding Author Institutional Email: hatef.otroshi@ee.sharif.edu
(H. Otroshi Shahreza)

Please cite this article as: H. Otroshi Shahreza, A. Amini, H. Behroozi, "Predicting the Empirical Distribution of Video Quality Scores using Recurrent Neural Networks", International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications Vol. 33, No. 5, (May 2020) 984-991

type and, consequently, the overall quality level. Recently, the KonVid-1k database [10] was made available, which contains video files in their ordinary status (authentically distorted). In consequence, this database may provide a more realistic test-bed for NR-VQA methods. It is worth mentioning that some of the most successful methods, according to the LIVE database, perform poorly with the KonVid-1k database [11]. Moreover, the values of individual scores are also available for each video file in the KonVid-1k database. Therefore, we can extract the ground-truth empirical distribution of people's opinion scores for the video files in this database.

In recent years, machine learning approaches such as neural networks, and in particular, deep neural networks (DNN) are dominantly used for the purpose of image quality assessment. With the availability of large datasets and computational power needed for image data, these methods (e.g., the CNN structure) have recorded remarkable performances in practice [12-17]. The landscape is, however, different for video data; the available datasets are rather limited in size and the implementations via DNN are computationally very costly (if not out of reach). Currently, most NR-VQA techniques rely on extracting a few features from the video data and training a standard regressor(such as SVR). In this paper, we show that the special structure of recurrent neural networks (RNN) allows for realizing an NR-VQA technique with high performance and manageable computational cost. The main advantage of RNNs is that they admit sequential inputs, which is a perfect fit for video data (a sequence of images). Our proposed method is based on training an RNN with simple features extracted from video frames and frame differences. We include features that are independent of the frame size, which makes the overall method applicable to all video frame sizes. Besides, the use of RNN automatically comes with the advantage of applicability to arbitrary time duration (number of frames). Moreover, in contrast to the existing methods that predict the quality by estimating the average of subjective scores (i.e., mean opinion score or shortly MOS), our technique is capable of predicting the full distribution of the scores. The experiments indicate that the proposed method achieves a comparable performance with the state-of-the-art methods on the KonVid-1k database with a less computational cost. Indeed, this result confirms the feasibility of using tailored DNNs for video quality assessment. Indeed, the method could also be improved by incorporating more informative features.

1. 1. Related Works Although the study of image and video quality enhancement is an interesting topic [18-25], the study of image or video quality assessment is rather a new research topic that is gaining attention in recent years. In particular, NR-VQA is further a challenging problem. Until now, most developed NR-VQA methods focus on extracting video features and combining them via a regression technique (commonly SVR) to predict MOS as the overall quality. The features are mainly extracted from video frames, which could be considered as image features. The inclusion of video-specific features has been the main challenge. In [3], *Saad et. al.* used spatial Natural Scene Statistics (NSS) as frame-level features, and motion-related and Natural Video Statistics (NVS) as video specific features. An SVR is then used to combine these features. This method is one of the top performers on the LIVE database [9] and has a fair performance on the KonVid-1k database. We should highlight that the extraction of the required features for this method is rather time-consuming. The latter issue is improved in [2]; simpler features based on statistic changes of video frame differences are used without a regression. Indeed, the features are once calculated for the frame differences and once for a blurred version of the frame differences, a correlation between the two determines the overall quality metric. Although it achieves a considerable performance on the LIVE database, a disappointing performance is reported on the KonVid-1k database [11].

The CORNIA method developed in [26] is among the successful image quality assessment techniques (NR-IQA). The extension to the video is considered in [27]; this method (called V-CORNIA) first estimates the frame-level image scores, and then, applies a specific aggregation technique (a form of weighted averaging) to predict the overall video quality. It should be emphasized that no regression or training is applied at the video-level aggregation phase.

The FC method proposed by *Men et. al.* in [11] uses four sets of frame-level features and one set of temporal features to train an SVR. The same authors proposed an extension in [28] called STFC, which employs an additional set of spatiotemporal features (as well as the exposure time). The latter features are obtained by looking at the video as a 3D cube (stack of 2D images) and creating 2D slices that intersect all the frames. Both the FC and STFC methods are trained and tested on the KonVid-1k database with acceptable results; as one can predict, the STFC outperforms the FC method. By introducing a different set of spatiotemporal features, two methods are introduced in [5] that are trained on the KonVid-1k database using the multi-layer perceptron regressor (STSMLP) and the support vector regressor (SVR).

As described above, the main difference between these methods is in the used set of features, rather than the aggregation technique. In this paper, instead of introducing a new set of features, we use some of the simple and existing features to train a recurrent neural network. The good performance of the trained network reveals that the aggregation technique has a great impact

on the overall performance. Obviously, one could expect even better performances with more sophisticated sets of features. In [6], the feature map extracted by a convolutional neural network was fed to a simple recurrent neural network. This method achieved a high performance while the CNN part imposed a high computational cost.

We should highlight that in the last layer of our proposed neural network, the distribution of people's opinion score is predicted. In contrast, in previous methods, only the average of this distribution (i.e., MOS) is predicted. However, the complete distribution has further information about the QoE, which is of high interest in the broadcasting industry.

In the following sections, we first introduce the proposed method. We explain the frame-level features extracted from all video frames and the architecture of our neural network. Next, we describe the implementation details and the training procedure. Afterwards, we discuss the results of the proposed method and compare it with the state-of-the-art methods on the KonVid-1K dataset. Finally, we conclude the paper in the last section.

2. PROPOSED METHOD

The feature extraction is the first step in our method. The features are evaluated either from isolated frames or a number of adjacent frames; in the latter case, the features are assigned to one of the involved frames (central frame, if applicable). In this way, we can evaluate the same set of features for all frames (the sliding window approach for features derived from adjacent frames). We describe the used set of features in Section 2. 1. Next, we train a recurrent neural network with the extracted features. It is well-known that the RNN structure is a perfect match for sequential data types. The specific architecture of our RNN is explained in Section 2. 2. The distribution of opinion score for the video quality predicted in our method is defined as the output of the last layer in the RNN structure (shown in Figure 1).

2. 1. Features While the value of pixels or certain transforms of the pixels convey very useful information about the frames and ultimately, the video, our features in this paper are defined by the statistics of a number of functions of the pixels (e.g., mean and variance). The benefit of this strategy is that the method becomes independent of the frame size. In other words, we do not need to retrain the neural network each time the frame size changes. The features used in our method are:

2. 1. 1. Luminance MSCN Statistics Let $F(i,j)$ represent the luminance values of a given frame in the video. Here, $i \in \{1,2,\dots,M\}$ and $j \in \{1,2,\dots,N\}$ point at the vertical and horizontal locations of a pixel within the frame, where M and N stand for the frame height and width, respectively. The Mean Subtracted Contrast Normalized (MSCN) luminance values are defined as:

$$\hat{F}(i,j) = \frac{F(i,j)-\mu(i,j)}{\sigma(i,j)+C}, \quad (1)$$

where

$$\mu(i,j) = \sum_{k=-K}^{K}\sum_{l=-L}^{L} \omega_{k,l} F(i+k,j+l) \quad (2)$$

is the local mean and

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L} \omega_{k,l}[F(i+k,j+l)-\mu(i,j)]^2} \quad (3)$$

is the local standard deviation. The constant C is conventionally set as 1 to avoid instability when the $\sigma(i,j)$ is small. The weights $\omega = \{\omega_{k,j} \mid k = -K,\dots,K, l = -L,\dots,L\}$ are obtained from a 2D circularly-symmetric Gaussian function:

$$\omega_{k,l} = \frac{e^{-18\left(\frac{k^2}{(2K+1)^2}+\frac{l^2}{(2L+1)^2}\right)}}{\sum_{k=-K}^{K}\sum_{l=-L}^{L} e^{-18\left(\frac{k^2}{(2K+1)^2}+\frac{l^2}{(2L+1)^2}\right)}} \quad (4)$$

The variance of the Gaussian distribution is set such that the main part of the pdf ($\pm 3\sigma$) fits within the considered window. We set $K = L = 3$ in this paper. Previous studies in [29], [30] show that the histogram of $\hat{F}(i,j)$ can be fairly described by a Generalized Gaussian Distribution (GGD):

$$f(x;\alpha,\beta) = \frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)} \exp\left(-\frac{|x|^{\alpha}}{\beta^{\alpha}}\right) \quad (5)$$

where $\Gamma(.)$ is the gamma function. Moreover, the pair-wise product of the adjacent pixels in $\hat{F}(i,j)$ fairly follow an Asymmetric Generalized Gaussian Distribution (AGGD):

$$f(x;\gamma,\beta_l,\beta_r) = \begin{cases} \frac{\gamma}{(\beta_l+\beta_r)\Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\frac{|x|^{\gamma}}{\beta_l^{\gamma}}\right), & \forall x \le 0 \\ \frac{\gamma}{(\beta_l+\beta_r)\Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\frac{x^{\gamma}}{\beta_r^{\gamma}}\right), & \forall x \ge 0 \end{cases} \quad (6)$$

The mean of this distribution is given by below:

$$\eta = (\beta_r - \beta_l)\frac{\Gamma\left(\frac{2}{\gamma}\right)}{\Gamma\left(\frac{1}{\gamma}\right)} \quad (7)$$

The parameters (α,β), and $(\gamma,\beta_l,\beta_r,\eta)$ for four orientations of pairing form a total of 18 statistical parameters. We further consider the image in two scales, the original one and a factor two down-sampled. We estimated the involved parameters (except for η) using a moment-matching approach [31]. We evaluate these features for every frame of the video file (36 values for each frame).

2. 1. 2. Luma Information The brightness histogram of video frame also plays a role in the perception of the quality. For this purpose, we compute the mean and

standard deviation of the luminance channel of each video frame. These two values are showing the frame's luminance information (2 values for each frame).

2. 1. 3. Chroma Information One possible way to include the color content of the video in the quality metric is to extract features from the two chrominance channels of the frames. Similar to the Luma features, we find the mean and standard deviation of each chrominance channel of each frame (4 values for each frame).

2. 1. 4. Colorfulness The introduced Chroma information somehow presents the distribution of color intensities within the video. Another metric for measuring the colorfulness of images is presented in [32]. Using the RGB color space, one can define rg and yb components as $rg = R - G$ and $yb = 0.5(R + G) - B$. Now, the colorfulness metric $M^{(3)}$ in [32] is defined as:

$$M^{(3)} = \sigma_{rgyb} + 0.3\mu_{rgyb}, \quad (8)$$

where

$$\sigma_{rgyb} = \sqrt{\mathrm{Var}(rg) + \mathrm{Var}(yb)}, \quad (9)$$

$$\mu_{rgyb} = \sqrt{\mathrm{mean}(rg)^2 + \mathrm{mean}(yb)^2}. \quad (10)$$

For the colorfulness features of the video, we find the value of $M^{(3)}$ for every frame of the video (1 value for each frame).

2. 1. 5. Spatial Gradient Each video frame represents an image; the horizontal and vertical gradients (derivatives) of this image reveal the edges. For this feature, we apply two 5 × 5 filter kernels (horizontal and vertical directions) on the luminance channel of the frame. We get the mean and standard deviation for each frame (4 values for each frame).

2. 1. 6. Spatial Laplacian The Laplacian of a frame is a particular 2nd order derivative of the data and is known to be rotation and scale-invariant. We implement the Laplacian operator via a 5 × 5 filter kernel applied to the luminance channel of the frame. We find the mean and standard deviation of the result over each frame (2 values for each frame).

2. 1. 7. Temporal Information So far, the features treat a video as a stack of images (frame). To include the temporal information of the video (changes of the frames), we simply generate difference frames by subtracting the luminance channel of each frame from the luminance channel of the preceding frame. The result is a real-valued frame (a mixture of positive and negative values). Next, we convert difference frames into features by evaluating the mean and standard deviation of each difference frame (2 values for each frame difference).

2. 2. Architecture of the Neural Network After the feature extraction step, we obtain 51 features per frame within 7 categories of features. In the next step, we normalize these features and feed them to a recurrent neural network in a sequential manner. As shown in Figure 1, our proposed structure consists of three Bidirectional-LSTM units [33] followed by two fully connected layers. In each of the fully connected layers, we apply Batch Normalization [34] before the tanh activation function. Eventually, the output layer containing 5 neurons with softmax function predict the distribution of scores in assessment of the video quality as follows:

$$\mathrm{softmax}(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{5} e^{s_j}}, \quad (11)$$

where s_i is the *i-th* neuron in the last layer indicating the probability of the score-value *i*.

3. IMPLEMENTATION

3. 1. Database For training and testing the proposed RNN we use the KonVid-1k database [10] in this paper. With 1200 humanrated video files, KonVid-1k is currently the largest database available for NR-VQA.

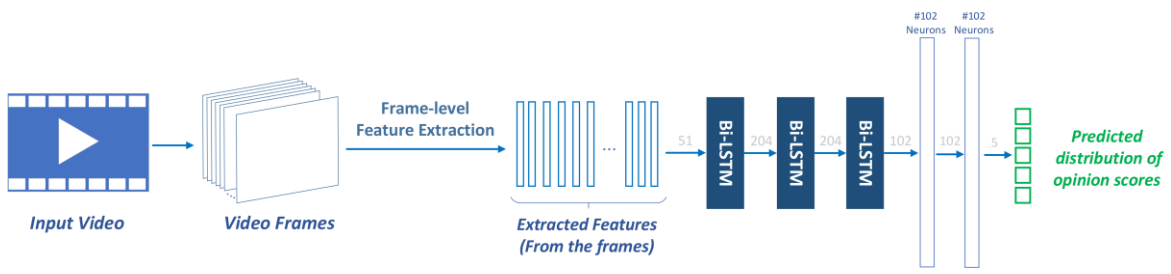

Figure 1. Overview of the proposed method: after the feature extraction stage, a recurrent neural network transforms the features into a distribution of quality scores at its output layer. The gray numbers indicate the output dimension of each layer. The inputs to the Bi-LSTM units are sequences of vectors each with the specified dimensions

Each video in this database is accompanied with around 50 human subjective scores using the ITU 5-point absolute category rating (ACR) quality scale (1 to 5 representing worst to best quality). Since the raw scores are published for this dataset, we can calculate the ground-truth empirical distribution of people's opinion score for the video files.

3. 2. Optimization The training stage of a neural network is accomplished by tuning the parameters so as to minimize a cost. The implemented cost (loss function) in our case is the Kullback Leibler (KL) divergence between the ground-truth and the estimated distribution of opinion scores:

$$L(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{5} p_i \log \frac{p_i}{q_i}, \quad (12)$$

where **p** and **q** are the ground-truth and the predicted distribution of opinion score, respectively. To solve the optimization problem, the Adam optimizer [35] with initial learning rate of 1×10^{-2} is used. For better optimizing the neural network, we decreased the learning rate by factor 0.1 if the validation loss did not decreased after 10 epochs. To implement the training procedure, we used the Keras library with the TensorFlow back-end.

3. 3. Training To achieve a fast training stage, we use the batch mode in our optimization (batch size of 64). On one hand, the video files in each batch are constrained to have the same length (number of frames). On the other hand, the length of the video files in the KonVid-1k database are not the same. However, this database contains a subset of 810 videos with 240 frames. For the purpose of training and testing our network, we only consider this subset of size 810 (i.e., we ignore the remaining 390 files in the database). For the training of the RNN, we randomly select 75% of this subset; 80% of the latter files are used for training, and the remaining 20% are used for validation. We apply the dropout technique [36] with probability p = 0.3 in the Bi-LSTM and fully connected layers during the training procedure. In addition, we stop the optimization when the loss function stops improving (the early stopping technique). Our experiments are conducted on a Linux desktop computer with an Intel Core-i7 3.6GHz CPU, 32GB RAM, and one NVIDIA GeForce 1080 Ti. Using this computer, the training procedure with 50 epochs took about 5 minutes. Finally, we use the rest of the subset for testing the trained RNN. Figure 2 shows the scatter plot regarding the average of distribution for training and test subset. In this plot, the horizontal axis indicates the ground-truth MOS and the vertical axis indicates the predicted MOS.

4. DISCUSSION

The trained neural network acts as a predictor of the distribution of people's opinion scores in the assessment of the video quality. Figure 3 shows the results for two sample videos in the test subset. Indeed, the predicted distribution provides further information about the users' quality of experience than its average value.

To compare our model with previous methods, we calculated the average of distributions to have MOS as a similar quality score. Next, we find the Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC) of the calculated average values with the ground-truth MOS values. To have a basis for comparison, we also included a number of existing NR-VQA methods[1] in Table 1. The last row of Table 1 shows these values in the evaluation of the proposed method, dubbed as DNet, which stands for distribution network. This Table indicates that the proposed method achieves comparable or better results in

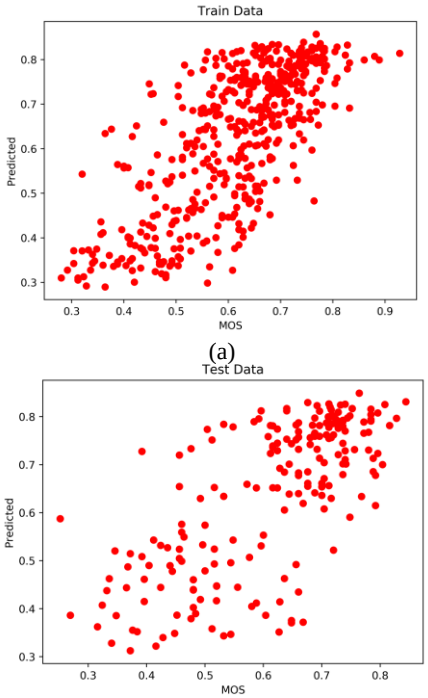

Figure 2. The scatter plot of quality scores of (a) train videos and (b) test videos. The horizontal axis indicates the ground-

[1] The PLCC and SROCC values of VIIDEO, V.BLINDS, FC model, and STFC model are reported from [28]. Similar values for STS-MLP, STSSVR, CNN+LSTM, TL-CNN+LSTM, and F-RNN are extracted from their papers [5],[6],[16]. We have implemented the V.CORNIA method for the purpose of this comparison.

truth MOS and the vertical axis indicates the average of predicted distribution

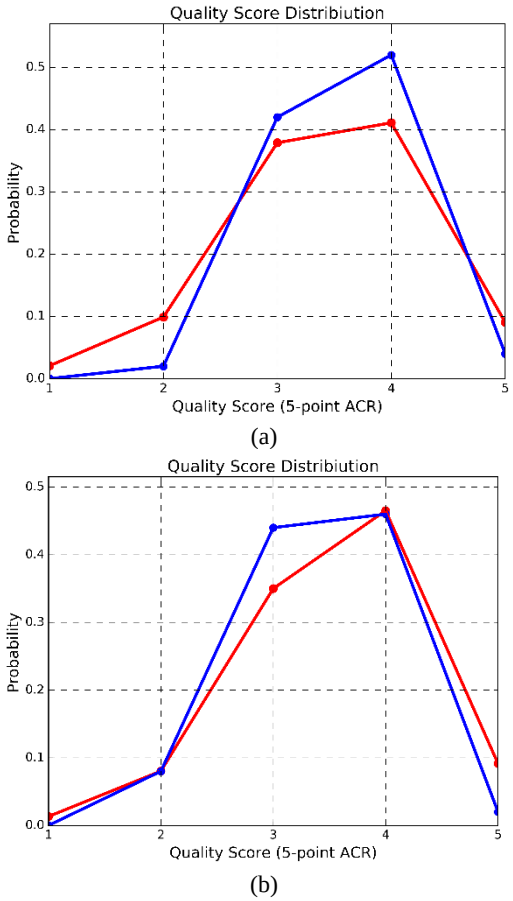

Figure 3. Two sample of people opinion score distribution. Note that the red lines indicate the predicted distribution and the blue lines indicate the ground-truth distribution

TABLE 1. comparison of NR-VQA methods on the KonVid-1k database. The method names shown in italic are based on deep learning approaches

Method	PLCC	SROCC
VIIDEO [2]	-0.015	0.031
V.BLINDS [3]	0.565	0.526
V.CORNIA [27]	-0.535	-0.544
FC Model [11]	0.492	0.472
STFC Model [28]	0.639	0.606
STS-MLP [5]	0.407	0.420
STS-SVR [5]	0.680	0.673
CNN+LSTM [6]	0.513	0.545
TL-CNN+LSTM [6]	0.867	0.849
F-RNN [16]	0.683	0.710
DNet [Proposed]	0.596	0.591

predicting MOS, while it even provides further information about the video quality. The PLCC and SROCC values of our method, DNet, are reported through 5-folded cross-validation.

We emphasize again that the features used in this paper are simple and fast to extract. We interpret the success of the proposed method as the usefulness of the RNN structure for the NR-VQA application. Since the extracted features are simple and have limited information on the frame and video quality, we believe that more detailed features can lead to better performance using the same RNN structure. Another possible direction for improvement is to combine the convolutional neural networks (CNNs) and RNNs; in particular, CNN provides richer feature sets while the RNN provides a smart aggregation technique. A similar approach is proposed in [6], but we should consider that using CNN impose a high computational cost. There are two models proposed in [6], one using a pre-trained CNN which is frozen during the training process and in the other one the CNN part is optimized as well. The results show that the model with fine-tuned CNN has superior performance, while the model frozen CNN achieves poor performance. This observation indicates that the feature map of the pre-trained CNN does not provide rich information. However, we should notice that fine-tunning the CNN part needs more computation resource in the training stage.

Additionally, we should highlight that since our method is a learning-based approach, it needs adequate training data. Like any other machine learning and deep learning models, one can achieve a better performance if more training samples with a vast diversity are available. The largest publically available dataset for NR-VQA is KonVid-1k which is used in this paper.

5. CONCLUSION

In this paper, we proposed a new method assessing the quality of in-the-wild video data without having any information about the source file (no-reference video quality assessment or briefly, NR-VQA). Our method uses the recurrent neural network structure to better cope with the sequential nature of the video. Instead of directly using the video frames (which is computationally heavy), we extract simple frame-based statistical features that are independent of the frame size. In contrast with previous studies which predicted only mean opinion score, shortly MOS, our model can predict the empirical distribution of people's opinion score. To the best of our knowledge, this is the first paper to investigate prediction of the empirical distribution of human opinion scores for video quality instead of MOS. Our numerical experiments using the

KonVid-1k database reveals that the proposed method has a suitable performance, which is comparable or even better than the existing methods.

6. REFERENCES

1. Cisco White Paper, "Cisco visual networking index: Forecast and methodology" (2019), 2017-2022.
2. Mittal A., Saad M. A., and Bovik A. C., "A completely blind video integrity oracle," ***IEEE Transactions on Image Processing***, Vol. 25, No. 1, (2016), 289-300.
3. Saad M. A., Bovik A. C., and Charrier C., "Blind rediction of natural video quality," ***IEEE Transactions on Image Processing***, Vol. 23, No. 3, (2014), 1352-1365.
4. X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," ***IEEE Transactions on Image Processing,*** Vol. 25, No. 7, (2016), 3329-3342.
5. P. Yan and X. Mou, "No-reference video quality assessment based on perceptual features extracted from multi-directional video spatiotemporal slices images," in Proceedings of SPIE Optoelectronic Imaging and Multimedia Technology V, Beijing, China, (2018).
6. Varga D. and Sziranyi T., "No-reference video quality assessment via ´ pretrained cnn and lstm networks," ***Signal, Image and Video Processing***, (2019), 1-8.
7. Varga D., "No-reference video quality assessment based on the temporal pooling of deep features," ***Neural Processing Letters***, (2019), 1–14.
8. Otroshi Shahreza H., Amini A., and Behroozi H., "No-reference video quality assessment using recurrent neural networks," in Proceedings of the 5th Conference on Signal Processing and Intelligent Systems (ICSPIS), Sharood, Iran, (2019).
9. Seshadrinathan K., Soundararajan R., Bovik A. C., and Cormack L. K., "Study of subjective and objective quality assessment of video," ***IEEE Transactions on Image Processing***, Vol. 19, No. 6, (2010), 1427-1441.
10. Hosu V., Hahn F., Jenadeleh M., Lin H., Men H., Sziranyi T., Li S., and Saupe D., "The konstanz natural video database (konvid-1k)," in Proceedings of IEEE International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, (2017).
11. Men H., Lin H., and Saupe D., "Empirical evaluation of no-reference vqa methods on a natural video quality database," in Proceedings of IEEE International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, (2017).
12. Kim J., Zeng H., Ghadiyaram D., Lee S., Zhang L., and Bovik A. C., "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," ***IEEE Signal Processing Magazine***, Vol. 34, No. 6, (2017), 130-141.
13. S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," ***IEEE Transactions on Image Processing***, Vol. 27, No. 1, (2017), 206-219.
14. S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," ***Signal, Image and Video Processing***, Vol. 12, No. 2, (2018), 355-362.
15. Otroshi-Shahreza H., Amini A., and Behroozi H., "No-reference image quality assessment using transfer learning," in Proceedings of the 9th International Symposium on Telecommunications (IST), Tehran, Iran, (2018).
16. Otroshi Shahreza H., Amini A., and Behroozi H., "In-the-wild noreference image quality assessment using deep convolutional neural networks," in Proceedings of the 5th Conference on Signal Processing and Intelligent Systems (ICSPIS), Shahrood, Iran, (2019).
17. Gu J., Meng G., Xiang S., and Pan C., "Blind image quality assessment via learnable attention-based pooling," ***Pattern Recognition***, Vol. 91, (2019), 332-344.
18. Yang R., Xu M., Wang Z., and Li T., "Multi-frame quality enhancement for compressed video," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, (2018).
19. Vu T., Van Nguyen C., Pham T. X., Luu T. M., and Yoo C. D., "Fast and efficient image quality enhancement via desubpixel convolutional neural networks," in Proceedings of the European Conference on Computer Vision (ECCV), (2018).
20. Ignatov A., Kobyshev N., Timofte R., Vanhoey K., and Van Gool L., "Dslr-quality photos on mobile devices with deep convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision (CVPR), Honolulu, HI, USA, (2017).
21. Lore K. G., Akintayo A., and Sarkar S., "Llnet: A deep autoencoder approach to natural low-light image enhancement," ***Pattern Recognition***, Vol. 61, (2017), 650–662.
22. Hassanpour H. and Asadi A. S., "Image quality enhancement using pixel-wise gamma correction via svm classifier," ***International Journal of Engineering***, (2011).
23. Azari Nasrabad F., Hassanpour H., and Asadi Amiri S., "Adaptive image dehazing via improving dark channel prior," ***International Journal of Engineering***, Vol. 32, No. 2, (2019), 249–255.
24. Asadi S., Hassanpour H., and Mortezaei Z., "Image enhancement using an adaptive un-sharp masking method considering the gradient variation," ***International Journal of Engineering***, Vol. 30, No. 8, (2017), 1118-1125.
25. Iravani S. and Ezoji M., "A general framework for 1-d histogram-baesd image contrast enhancement," ***International Journal of Engineering***, Vol. 29, No. 10, (2016), 1384-1391.
26. Ye P., Kumar J., Kang L., and Doermann D., "Unsupervised feature learning framework for no-reference image quality assessment," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Rhode Island, USA, (2012).
27. Xu J., Ye P., Liu Y., and Doermann D., "No-reference video quality assessment via feature learning," in Proceedings od IEEE International Conference on Image Processing (ICIP), Paris, France, (2014).
28. Men H., Lin H., and Saupe D., "Spatiotemporal feature combination model for no-reference video quality assessment," in Proceedings of IEEE International Conference on Quality of Multimedia Experience (QoMEX), Sardinia, Italy, (2018).
29. Mittal A., Moorthy A. K., and Bovik A. C., "No-reference image quality assessment in the spatial domain," ***IEEE Transactions on Image Processing***, Vol. 21, No. 12, (2012), 4695-4708.
30. Mittal A., Soundararajan R., and Bovik A. C., "Making a "completely blind" image quality analyzer." ***IEEE Signal Processing Letters***, Vol. 20, No. 3, (2013), 209-212.
31. Lasmar N.-E., Stitou Y., and Berthoumieu Y., "Multiscale skewed heavy tailed model for texture analysis," in Proceedings of IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, (2009).
32. Hasler D. and Suesstrunk S. E., "Measuring colourfulness in natural images," in Proceedings of SPIE Human Vision and Electronic Imaging VIII, (2003).
33. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," ***IEEE Transactions on Signal Processing***, Vol. 45, No. 11, (1997), 2673-2681.

34. Ulyanov D., Vedaldi A., and Lempitsky V., "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022, 2016.

35. Kingma D. P. and Ba J., "Adam: A method for stochastic optimization," in Proceedings of International Conference on Learning Representations (ICLR), San Diego, California., USA, (2015).

36. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R., "Dropout: a simple way to prevent neural networks from overfitting," ***The Journal of Machine Learning Research***, Vol. 15, No. 1, (2014), 1929–1958.

Persian Abstract

چکیده

ارزیابی کیفیت ویدئو فرایندی پراهمیت در صنعت رسانه است. به دلیل حجم زیاد و همچنین مدت طولانی محتوای ویدئویی، استفاده از سامانه‌ای کامپیوتری تنها راه ممکن برای ارزیابی کیفیت ویدئو است. باوجود اینکه معمولا ارزیابی کیفیت ویدئو در مرحله فشرده‌سازی با مقابسه با ویدئو مرجع (بدون فشرده‌سازی) انجام میشود، اما در مراحل بعدی مانند انتقال، ویدئو مرجع وجود ندارد. بنابراین ارزیابی کیفیت ویدئو لازم است به صورت بدون مرجع (کور) صورت گیرد. روش‌های فعلی ارزیابی کیفیت ویدئو بدون مرجع، تنها میانگین نظرات افراد در ارزیابی کیفیت ویدئو را تخمین میزنند و اطلاعات بیشتری از توزیع این نظرات محاسبه نمی‌کنند. این درصورتی است که توزیع نظرات اطلاعات ارزشمندی درمورد کیفیت تجربه کاربران در اختیار می‌گذارد. در این مقاله، روشی برای پیش‌بینی توزیع نظرات افراد در ارزیابی کیفیت ویدئو پیشنهاد می‌شود. تا جایی که نویسندگان این مقاله اطلاع دارند، این اولین مقاله‌ای است که در آن توزیع نظر انسان در ارزیابی کیفیت ویدئو پیش‌بینی می‌شود. برای این مهم، ابتدا مشخصه‌هایی از فریم‌های ویدئویی استخراج کرده و سپس آن‌ها را به یک شبکه عصبی بازگشتی می‌دهیم. در پایان، توزیع نظرات در آخرین لایه شبکه بازگشتی پیش‌بینی می‌شود. آزمایش‌ها نشان می‌دهد که میانگین توزیع‌های پیش‌بینی شده دارای عملکرد بهتر یا قابل مقایسه با روش‌های پیشین بر روی پایگاه‌داده KonVid-1k است.