https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-168

No Reference Video Quality Assessment with authentic distortions using 3-D Deep Convolutional Neural Network

Roger Gomez Nieto, Hernan Dario Benitez Restrepo, Roger Figueroa Quintero
Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali, Colombia.
Alan Bovik; Department Of Electrical And Computer Engineering, The University of Texas at Austin, USA.

Abstract

Video Quality Assessment (VQA) is an essential topic in several industries ranging from video streaming to camera manufacturing. In this paper, we present a novel method for No-Reference VQA. This framework is fast and does not require the extraction of hand-crafted features. We extracted convolutional features of 3-D C3D Convolutional Neural Network and feed one trained Support Vector Regressor to obtain a VQA score. We did certain transformations to different color spaces to generate better discriminant deep features. We extracted features from several layers, with and without overlap, finding the best configuration to improve the VQA score. We tested the proposed approach in LIVE-Qualcomm dataset. We extensively evaluated the perceptual quality prediction model, obtaining one final Pearson correlation of 0.7749 ± 0.0884 *with Mean Opinion Scores, and showed that it can achieve good video quality prediction, outperforming other state-of-the-art VQA leading models.*

Introduction

Every day, millions of videos are being shared and spread on platforms such as Youtube, Netflix, and Hulu. Cisco estimates that video traffic will be 82 percent of all Internet traffic (both business and consumer) by 2022, up from 75 percent in 2017. Because of the high availability of smartphones, many of these videos are recorded by regular users who distort these videos with impairments such as artifacts, color, exposure, focus, sharpness, stabilization, caused by hardware limitations. Users do this because of lack of knowledge about the generation of professional-quality video. Natural videos often contain in-capture distortions that affect the video quality perceived by humans. Video streaming and camera manufacturers companies are keen to understand the influence and presence of these distortions in natural videos. This quality prediction can be carried out automatically by using VQA algorithms. Nonetheless, one of the main challenges in VQA is video content dependency, which makes it difficult to generalize from a unique dataset.

Most of the related work in No Reference (NR) VQA models has focused on compression and transmission artifacts [1–14], and the most used applications of these NR-VQA models is quality monitoring in video storage, streaming, gaming and broadcasting [15]. Mittal *et al*. [16] created a No Reference Image Quality Assessment (IQA) model, called NIQE, that makes use of statistical regularities observed in natural images, without training on MOS. They base this model on the construction of a set of statistical features based on a Natural Scene Statistics (NSS) model. The IQA is given by the distance between NSS features extracted and a multivariate Gaussian model of the quality aware features extracted from the collection of images. In [17] Bampis *et al*. applied NIQE to VQA, and tested it on the LIVE-NFLX Database [18], with little success, probably because it is a frame-based NR model (intended initially to IQA).

In [19], the authors proposed a NR IQA method that uses spatial features. The model, called Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) uses scene statistics of locally normalized luminance coefficients to quantify losses of "naturalness" in an image due to distortions. BRISQUE is not limited by the type of distortions, obtaining an advantage compared to other approaches to NR IQA that are distortion-specific [20]. Sogaard *et al.* [14] proposed a NR VQA method; they built features by the estimation of selected video codec parameters along with BRISQUE features, applied to each frame. They exploited a changed version of the IQA BRISQUE and applied it to videos, along with features from the video codec analysis, and used these as input for a SVR machine learning method.

Saad *et al.* proposed in [21–23] a NR VQA method dubbed video BLind Image Integrity Notator using DCT Statistics (V-BLIINDS) that is built on a spatio-temporal model of videos converted to the discrete cosine transform domain, and on a model that characterizes the motion occurring in the scenes to predict video quality. V-BLIINDS uses features extracted under the spatio-temporal Natural Video Scenes model to feed a support vector regressor (SVR) trained to predict the visual quality of videos. In this model, they assessed the spatial and temporal dimensions of the videos collectively. They tested V-BLIINDS on the LIVE VQA database [24], demonstrating good performance on compression and packet loss [25].

Mittal *et al*. [10] proposed an NR VQA method called the video intrinsic integrity and distortion evaluation oracle (VIIDEO). They quantify distortions using statistical regularities of natural videos. VIIDEO is based upon a set of perceptually relevant temporal video statistic models of video frame difference signals to obtain a VQA score. To select appropriate features for the Asymmetric Generalized Gaussian Distribution model, they used three criteria: (i) regularity over pristine videos, (ii) regularity should be destroyed on distorted videos, (iii) the loss of regularity should vary with the degree of perceived distortion. However, they do not test VIIDEO on a dataset with naturally distorted videos, such as LIVE-Qualcomm [26].

Ghadiyaram *et al.* in [27] proposed a No Reference IQA method, called Feature Maps–Based Referenceless Image Quality Evaluation Engine (FRIQUEE), based on a large set of perceptually relevant feature maps, along with NSS models. These features feed an SVR learning model with radial basis kernel functions. They deployed these models in several color

spaces (HSI, LMS, LAB, CIELAB [28]), and tested the proposed method, mainly on the LIVE In the Wild Image Quality Challenge Database (1162 natural images, obtained using mobile camera devices, and evaluated by 8100 human observers) [29, 30], achieving good quality prediction on authentically distorted images.

In [31] the authors proposed a VQA method that takes into account video content. The algorithm calculates four measures: 1. picture resolution, 2. bitrates, 3. spatial Information (SI), and 4. temporal information (TI) to represent the visual quality in four separate dimensions. Thus, they created a dataset with ten videos represented by MOS provided by 20 subjects. However, they did do not test VQA performance on other datasets. Furthermore, the proposed dataset was small and did not allow for adequate generalization.

Ghadiyaram *et al.* [26] proposed a dataset dubbed LIVE-Qualcomm Mobile In-Capture Video Quality Database, comprising 208 videos obtained from 8 different smartphones, modeling six common in-capture distortions. They conducted a subjective quality assessment study on this dataset, where 39 subjects assessed each video. Likewise, they tested several state-of-the-art No-Reference IQA and VQA algorithms [10, 16, 19, 21, 27] on this dataset and reported FRIQUEE [27] to be the best performing VQA algorithm on LIVE Qualcomm in terms of PLCC and SROCC, which were 0.7349 and 0.6795 respectively, with all distortions commingled.

In [32] Zhang *et al.* showed the effectiveness of deep features as a perceptual metric. They found that deep features outperform previous metrics by large margins in perceptual tasks. They remarked that this performing is independent of CNN architecture and levels of supervision. They demonstrated that even CNN trained for common computer vision tasks achieve good performance on semantic and perceptual tasks [33]. However, this performance can be improved by training a simple linear scaling of layers activation using perceptual VQA or IQA datasets. Similarly, other studies determined that a CNN trained for object and video recognition can be useful to determine human perceptual characteristics [34, 35].

Göring *et al.* [8] trained two no-reference models for classical video quality up to 4K resolution. The first one is a BRISQUE+NIQE baseline model trained on per-frame VMAF scores, using one random forest regression model without feature selection. The second approach uses a pre-trained classification Deep Neural Network (DNN) in combination with hierarchical sub-image creation. They introduce a hierarchical patching approach to feed the DNN, dividing a frame into sub-images of equal size (1/2, 1/4, and 1/8 of each dimension). To generate deep features, the authors used the inception-v3 network [36], implemented in Keras, with re-scaling to 299$\times$ 299 pixels. However, they tested the DeViQ method on one dataset containing only 12 videos, which led to low generalization capabilities.

We have organized the remainder of the paper as follows: in the Section Materials and Methods, we describe the overall framework of the model. We show the process to train a Support Vector Regressor (SVR), and the dataset used. We explain the correlation measures used, and the color spaces used to transform the original videos in the dataset. In the Results and Discussion section, we report and analyze the experimental results, and section 5 concludes the paper.

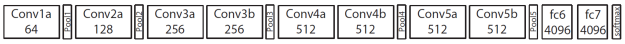


Figure 1. *C3D Architecture [37].*

Materials and Methods

Convolutional Neural Network

In all the experiments, we use the C3D Network [37] to extract spatio-temporal features. In [37] the authors showed that these spatio-temporal features, along with a simple linear classifier, can yield good performance on some video analysis tasks, such as action recognition [38], action similarity labeling, scene classification and object recognition. C3D uses a $3 \times 3 \times 3$ convolution kernel for all layers, and takes full video frames as input, with no pre-processing stage. C3D performs 3D convolutions and 3D pooling to propagate temporal information across all the layers, allowing access to the model temporal information.

C3D has five convolution layers and five pooling layers (a pooling layer immediately follows each convolution layer), two fully-connected layers and a softmax loss function to predict action labels. The number of filters for convolution layers from 1 to 5 are 64, 128, 256, 256, 256, respectively, as shown in Fig. 1. C3D resizes all video frames into 128×171 pixels [37]. In one experiment, we modified the C3D original architecture to take as input to the convolutional layers videos split into 16 frames clips with an 8-frame overlap between two consecutive clips, obtaining more feature vectors per video and thereby improving the features/data ratio. C3D has 17.5 million parameters in the fully connected layers. The authors tested four distinct architectures, varying the size and depth of the kernel in each one. The number of parameters in the convolutional layers are different for each architecture, but the variation is minimal compared with the 17.5 million parameters in the fully connected layers, which is the same for all architectures [37]. We extracted feature vectors with dimensions of 50175 from the fifth convolutional layer (conv5b) and 4096 of the fully connected layer (fc6).

Dataset

Several VQA databases have been created in recent years [39, 40]. We used the database LIVE-Qualcomm Mobile In-Capture Video Quality proposed by [26], because it contains videos with authentic distortions. The original videos from LIVE-Qualcomm are in YUV420 format. We converted all videos to AVI Uncompressed. This resulted in videos with an average size of 2.8 Gigabytes, duplicating its original size. This kind of conversion allows for minimizing the information lost by compression, thus avoiding adding other different post-capture distortions to the videos. However, a drawback is the enormous size of these videos, increasing the processing time to extract the features from C3D layers. The videos in the LIVE-Qualcomm dataset have an average duration of 15 seconds and have a rate of 30 Frames per Second (FPS). Each video has approximately 450 frames, but not all videos have the same duration; some videos have less than 400 frames. We discarded one video as an outlier (only 360 frames). Therefore, we used 207 out of the 208 videos from the LIVE-Qualcomm dataset.

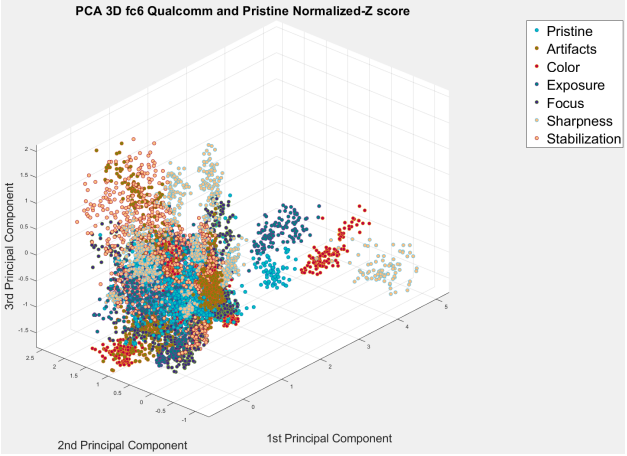


Figure 2. *3-D PCA of pristine and distorted videos from LIVE-Qualcomm dataset, with No Average Pooling NA (25 points represents one single video).*

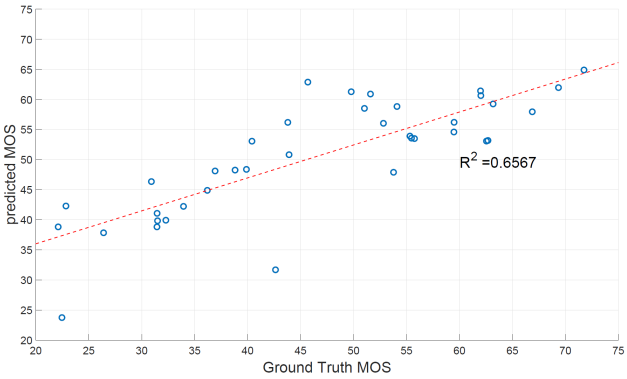


Figure 3. *Scatter plot between Ground Truth MOS and Predicted MOS obtained with Fc6_YCbCr_8Frames_AP VQA proposed method. There are 39 videos in the test set. We reported the linear correlation coefficient* $R^2 = 0.6567$*, and plotted the least-squares line (red).*

Pre-processing

We pre-processed each video using a transformation to the YCbCr color space on account of the rough correspondence between the YCbCr components and visual attributes. YCbCr is one of two primary color spaces used to represent digital videos (along with RGB). The distinction between YCbCr and RGB is that YCbCr represents color as brightness (Y) and two color difference signals (CB, Cr), while RGB represents color as red, green and blue [41, 42]. Likewise, YCbCr is less redundant than RGB, supporting the CNN encoding capabilities. In YCbCr color space, lightness changes that affect image contrast (but not color) are easily accessed. Besides, YCbCr is intended to take advantage of human colour-response characteristics. We believe that this can facilitate the detection of certain distortions. We also pre-processed each video by using a statistical image model based on Mean Subtracted Contrast Normalized (MSCN) coefficients. MSCN coefficients have characteristic statistical properties that are changed by distortions and it is known that quantifying these changes makes it possible to predict the distortion affecting an image and its perceptual quality [19].

SVR training

SVR has been applied successfully to VQA problems [12, 21, 43–48]. In some VQA frameworks, one usual way to obtain a quality score is to train the SVR with the proposed features. SVR is able to handle high-dimensional data, comparable to the length of features vector in the output of C3D convolutional and fully connected layers. We utilized the MATLAB Machine Learning Toolbox to implement the SVR with a radial basis function (RBF) and Gaussian kernel. We found the optimal model parameters of the SVR via 10-fold cross-validation. The aim of minimizing the error to the validation data guided our selection of the model. We used a random, nonoverlapping train and test split with 80% of the sequences for training and 20% for testing in each test case. To avoid any bias due to the division of data, we randomly split the data set 100 times. The median PLCC, SRCC results, and their standard deviations for each test case are reported in Tables 1-2. A higher value of each of these metrics shows better performance in terms of correlation between MOS and proposed VQA method scores.

To evaluate the performance of VQA methods, we adopted two criteria, namely the Pearson Linear Correlation Coefficient (PLCC), and the Spearman Rank Ordered Correlation Coefficient (SROCC) between ground-truth MOS and predicted MOS. The PLCC is a measure of the strength of linear dependence between two variables. The SROCC measures prediction monotonicity, since it operates only on the ranking of the data points and ignores the relative distance between them. The absolute value of SROCC describes the intensity of the monotonic relationship [49]. The feature vectors have a size of 4096×1 for Fully Connected layer fc6 and 50176×1 for the fifth convolutional layer conv5b. These vectors compose the matrix F_{input}, which is the input matrix to train and test the SVR machine. One of the deployed approaches uses Average Pooling (AP) by averaging all columns from matrix F_{input} (m is the number of features in the output of CNN layer), where the matrix of size $m \times n$ is converted in $m \times 1$, to represent each video as a single feature vector.

Results and Discussion

We obtained 73 pristine videos from several sources [24, 50, 51] and extracted the C3D deep features vectors from fully connected layer fc6. Thereafter, we studied these feature vectors with the equivalent feature vectors of distorted videos from the LIVE-Qualcomm dataset. Figure 2 shows the results of Principal Component Analysis projection (PCA). Diverse clusters occur because of the pristine and the various types of distorted videos. The videos were converted to YCbCr color space and processed by the C3D CNN. The values in Table 1 and 2 justify our choice of YCbCr color space, which plays a significant role in enhancing video quality predictions. Therefore, we take the feature vector from the output of the sixth fully connected layer fc6, using the advancement of eight frames in each block of frames to feed the CNN.

The LIVE-Qualcomm dataset contains videos acquired from four different smartphones per unique scene. We evaluated the impact of this information redundancy in the SVR performance. To do this, we took the method having higher overall performance (Fc6_YCbCr_8Frames_AP) and deleted duplicated videos for the same scene. We organized the dataset in such a way that each

TABLE 1: Median PLCC ± Standard Deviation of proposed VQA method. AP indicates Average Pooling. NA indicates No Average (we use one feature vector each XX frames), and AP* suggests that only one video per unique scene is used. The results of the methods of six upper rows was taken from [26], since they were evaluated on the same dataset

VQA Method	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	All distortions
FRIQUEE [27]	0.7638	0.3543	0.6808	**0.8107**	0.2203	0.7034	0.7349
BRISQUE [19]	0.6402	0.3392	0.6042	0.4550	0.5371	0.6940	0.5788
Average-pooled NIQE [16]	0.6078	0.2904	0.4625	0.5371	0.5595	0.6015	0.6802
Temporally-pooled NIQE	0.6766	0.3141	0.5213	0.5782	0.5508	0.6510	0.6749
V-BLIINDS [21]	**0.8386**	**0.6645**	**0.6900**	0.8077	**0.6845**	**0.7138**	0.6653
VIIDEO [10]	0.2888	0.3312	0.2073	0.2515	0.3012	0.3697	0.0982
Conv5b_RGB_8Frames_AP	0.2619 ± 0.2933	0.3714 ± 0.3994	0.5429 ± 0.3391	0.4000 ± 0.5107	0.6000 ± 0.3309	0.1429 ± 0.3529	0.6100 ± 0.1446
Conv5b_RGB_16Frames_AP	0.4524 ± 0.3486	0.3714 ± 0.3141	0.6000 ± 0.3042	0.4000 ± 0.5407	0.6000 ± 0.3589	0.2857 ± 0.3182	0.5906 ± 0.1092
Fc6_YCbCr_8Frames_AP	0.5714 ± 0.2635	0.6000 ± 0.3452	0.6000 ± 0.3349	0.5000 ± 0.4169	0.6000 ± 0.3407	0.4643 ± 0.3339	**0.7749±0.0884**
Fc6_YCbCr_8Frames_AP*	X	X	X	X	X	X	0.7648 ± 0.1487
Fc6_YCbCr_8Frames_NA	0.5494 ± 0.2380	0.6303 ± 0.3086	0.6183 ± 0.365	0.4828 ± 0.2854	0.5954 ± 0.3447	0.3767 ± 0.3175	0.7146 ± 0.0849
Fc6_MSCN_8Frames_AP	0.5476 ± 0.2399	0.4857 ± 0.3419	0.4286 ± 0.3015	0.3536 ± 0.5015	0.6559 ± 0.3929	0.1429 ± 0.4062	0.5824 ± 0.148
Fc6_MSCN_8Frames_NA	0.5000 ± 0.1992	0.5076 ± 0.2951	0.4146 ± 0.2588	0.4688 ± 0.353	0.5714 ± 0.2442	0.2284 ± 0.2947	0.5428 ± 0.1176
Fc6_MSCN_16Frames_NA	0.4730 ± 0.2061	0.4341 ± 0.2905	0.4635 ± 0.2622	0.5215 ± 0.3040	0.5082 ± 0.263	0.2215 ± 0.3029	0.5453 ± 0.1767
Fc6_MSCN_16Frames_AP	0.5000 ± 0.2887	0.4286 ± 0.3786	0.4857 ± 0.3020	0.3000 ± 0.4920	0.5798 ± 0.3255	0.1786 ± 0.3187	0.6129 ± 0.1439
Fc6_YCbCr_16Frames_AP	0.5476 ± 0.2593	0.4286 ± 0.3244	0.6000 ± 0.3566	0.3929 ± 0.2857	0.6571 ± 0.3381	0.4643 ± 0.3457	0.6380 ± 0.1194
Fc6_YCbCr_16Frames_NA	0.5495 ± 0.2581	0.4211 ± 0.2840	0.4698 ± 0.2685	0.4544 ± 0.3421	0.6192 ± 0.288	0.3757 ± 0.2754	0.5217 ± 0.1272

TABLE 2: Median SROCC ± Standard Deviation of proposed VQA method. AP indicates Average Pooling. NA indicates No Average (we use one feature vector each XX frames), and AP* suggests that only one video per unique scene is used.

VQA Method	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	All distortions
FRIQUEE [27]	**0.75**	0.4107	0.6071	0.7879	0.0714	**0.6607**	0.6795
BRISQUE [19]	0.6071	0.3571	0.5536	0.3929	0.4821	0.6429	0.5585
Average-pooled NIQE [16]	0.5	0.3214	0.3929	0.3393	0.5	0.2143	0.5451
Temporally-pooled NIQE	0.5357	0.3214	0.4821	0.3750	0.5179	0.2143	0.5525
V-BLIINDS [21]	0.7321	0.6071	0.6429	**0.8036**	0.6786	**0.6607**	0.6177
VIIDEO [10]	-0.1786	0.1429	-0.0714	0	-0.1786	-0.1071	-0.1414
Conv5b_RGB_8Frames_AP	0.2631 ± 0.2793	0.2127 ± 0.3107	0.5786 ± 0.3288	0.5908±0.4828	**0.7493±0.3068**	0.0 ± 0.2906	0.5752 ± 0.2141
Conv5b_RGB_16Frames_AP	0.3557 ± 0.3308	0.1564 ± 0.28	0.6153 ± 0.3282	0.478 ± 0.4948	0.7627 ± 0.333	0.0 ± 0.2962	0.5868 ± 0.1536
Fc6_YCbCr_8Frames_AP	0.6076±0.2364	0.7077 ± 0.3603	**0.7271±0.3263**	0.4524 ± 0.3819	0.7236 ± 0.346	0.445±0.3156	**0.7517±0.0853**
Fc6_YCbCr_8Frames_AP*	X	X	X	X	X	X	0.7507 ± 0.1281
Fc6_YCbCr_8Frames_NA	0.5365 ± 0.2688	**0.7294±0.3487**	0.6562 ± 0.4141	0.407 ± 0.2864	0.66 ± 0.3744	0.3361 ± 0.3048	0.69 ± 0.2985
Fc6_MSCN_8Frames_AP	0.5776 ± 0.2814	0.4583 ± 0.3482	0.4484 ± 0.3055	0.3401 ± 0.431	0.6493 ± 0.3867	0.1465 ± 0.3774	0.5781 ± 0.171
Fc6_MSCN_8Frames_NA	0.4943 ± 0.219	0.4348 ± 0.2872	0.4029 ± 0.2711	0.3527 ± 0.3269	0.6114 ± 0.2818	0.226 ± 0.3087	0.4631 ± 0.263
Fc6_MSCN_16Frames_NA	0.4761 ± 0.2101	0.3725 ± 0.2798	0.4192 ± 0.2782	0.4534 ± 0.2974	0.5854 ± 0.243	0.281 ± 0.3122	0.4901 ± 0.2458
Fc6_MSCN_16Frames_AP	0.4869 ± 0.298	0.4512 ± 0.3203	0.469 ± 0.2971	0.3096 ± 0.4598	0.6562 ± 0.3403	0.0488 ± 0.291	0.5831 ± 0.1898
Fc6_YCbCr_16Frames_AP	0.5776 ± 0.2549	0.3734 ± 0.2988	0.5396 ± 0.346	0.4009 ± 0.2969	0.7477 ± 0.3627	0.4276 ± 0.3254	0.617 ± 0.1274
Fc6_YCbCr_16Frames_NA	0.5256 ± 0.274	0.4393 ± 0.2799	0.4888 ± 0.2945	0.4744 ± 0.3206	0.7043 ± 0.3372	0.3679 ± 0.2782	0.5238 ± 0.2349

smartphone device was represented by approximately same number of videos, with a total of 54 videos (Galaxy GS5=8, Galaxy GS6=8, HTC One VX=8, Iphone 5S=8, LG G2=8, Lumia 1020 = 3, Note 4 =4, Oppo Find 7 =7). The results of this test are summarized in method Fc6_YCbCr_8Frames_AP*, as shown in the Tables 1 and 2. It may be observed that the overall performance varies by only 0.0099. These results support the conclusion that the redundancy in information per scene is not a critical factor in the overall performance of the proposed VQA method when tested on the LIVE-Qualcomm dataset. By contrast, the results reported in [26] were tested using all the videos, including those involved in the same scene.

For further analysis, we also calculated the R^2 correlation coefficient (values > 0 show a linear correlation). Figure 3 shows a scatter plot of our results using the Fc6_YCbCr_8Frames_AP VQA method on one test sequence of 39 videos, compared with the Ground Truth MOS from LIVE-Qualcomm. Our proposed VQA method obtained the best overall performance (all distortions), achieving a PLCC correlation of 0.7749 ± 0.0884, outperforming the best performance reported in [26] by FRIQUEE [27] (PLCC = 0.7349). Furthermore, our VQA method Fc6_YCbCr_8Frames_AP obtained the best performance on videos with exposure distortion, achieving a SROCC of 0.7271 ± 0.3263. Similarly, another proposed method, Fc6_YCbCr_8Frames_NA, obtained the second best results on exposure distortion, with a SROCC of 0.6562 ± 0.4141. Respect to distortions in separative way, and comparing with PLCC score, the method Fc6_YCbCr_8Frames_NA outperform the others methods in Color distortion. This method used the features extracted from the first Fully Connected layer, with the input data

converted to YCbCr format, and batch size separation of 8 frames.

Thereby, both methods outperformed the best performance reported in [26] on videos affected by exposure distortion with V-BLIINDS [21], which obtained 0.64290 without reporting standard deviation. The results for other methods (i.e., V-BLIINDS [21], VIIDEO [10], NIQE [16], BRISQUE [19], and FRIQUE [27]), are reproduced here [26] because they also report on median PLCC and SROCC values using tenfold cross-validation with 100 random train–validation–test split, and used the same dataset. Those methods that use Average Pooling (AP) required lower computational times; by reducing the size of matrix F_{input}, allowing accelerated execution of SVR training and testing.

Conclusions and Future Work

In this paper, we have proposed a NR VQA method, explicitly aimed at videos with natural distortions, such as color, artifacts, exposure, focus, sharpness, and stabilization. Our method is based on a 3D convolutional neural network approach, using features extracted from several layers of the CNN to feed an SVR model to produce an NR VQA model providing a high level of video quality prediction power. Our VQA method outperforms several state-of-the-art VQA methods when applied to authentically distorted videos.

Acknowledgments

The authors acknowledge the funding provided by COLCIENCIAS and Pontificia Universidad Javeriana with the project *Vigilancia Inteligente para la red de cámaras de la Policía Metropolitana de Cali*. The authors would like to thank NVIDIA Corporation for the donation of a TITAN XP GPU used in these experiments. The authors would also like to acknowledge the grant provided by *Comision Fulbright Colombia* to fund the Visiting Scholar Scholarship granted to Hernan Benitez and the research project entitled: 404 "Formacion e innovación para el fortalecimiento de la competitividad del sector TIC de la región: 405 FormaTIC e InnovaTIC Valle del Cauca, Occidente", BPIN 2014000100051, sponsored within 406 the Colombian General Royalty System and executed by PacificTIC.

References

[1] Maria Torres Vega, Decebal Constantin Mocanu, and Jeroen Famaey et al. Deep learning for quality assessment in live video streaming. *IEEE Signal Processing Letters*, 24(6):736–740, jun 2017.

[2] Vladimir Frants, Viacheslav Voronin, and Alexander Zelenskiy. Blind visual quality assessment for smart cloud-based video storage. In *2018 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, sep 2018.

[3] M. Alizadeh, A. Mohammadi, and M. Sharifkhani. No-reference deep compressed-based video quality assessment. In *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, oct 2018.

[4] Yasamin Fazliani, Ernesto Andrade, and Shahram Shirani. Learning based hybrid no-reference video quality assessment of compressed videos. In *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, may 2019.

[5] Nabajeet Barman, Emmanuel Jammeh, and Seyed Ali Ghorashiet al. No-reference video quality estimation based on machine learning for passive gaming video streaming applications. *IEEE Access*, 7:74511–74527, 2019.

[6] Domonkos Varga and Tamás Szirányi. No-reference video quality assessment via pretrained cnn and lstm networks. *Signal, Image and Video Processing*, Jun 2019.

[7] Zhiming Shi and Chengti Huang. Network video quality assessment method using fuzzy decision tree. *IET Communications*, jun 2019.

[8] Steve Göring, Janto Skowronek, and Alexander Raake. DeViQ – a deep no reference video quality model. *Electronic Imaging*, 2018(14):1–6, jan 2018.

[9] Jari Korhonen. Learning-based prediction of packet loss artifact visibility in networked video. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, may 2018.

[10] Anish Mittal, Michele A. Saad, and Alan C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, jan 2016.

[11] Nabajeet Barman, Steven Schmidt, and Saman Zadtootaghaj. An evaluation of video quality assessment metrics for passive gaming video streaming. In *Proceedings of the 23rd Packet Video Workshop on PV*. ACM Press, 2018.

[12] Saman Zadtootaghaj, Nabajeet Barman, and Steven Schmidt et al. NR-GVQM: A no reference gaming video quality metric. In *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, dec 2018.

[13] Steve Goring, Rakesh Rao Ramachandra Rao, and Alexander Raake. nofu — a lightweight no-reference pixel based video quality model for gaming content. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, jun 2019.

[14] Jacob Sogaard, Soren Forchhammer, and Jari Korhonen. No-reference video quality assessment using codec analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(10):1637–1650, oct 2015.

[15] Jari Korhonen. Two level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, pages 1–1, 2019.

[16] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[17] Christos G. Bampis, Zhi Li, and Alan C. Bovik. Continuous prediction of streaming video QoE using dynamic networks. *IEEE Signal Processing Letters*, 24(7):1083–1087, jul 2017.

[18] Christos George Bampis, Zhi Li, and Alan Conrad Bovik et al. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, nov 2017.

[19] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, dec 2012.

[20] Shan Suthaharan. No-reference visually significant blocking artifact metric for natural scene images. *Signal Processing*, 89(8):1647–1652, aug 2009.

[21] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, mar 2014.

[22] Michele A Saad, Alan C Bovik, and Christophe Charrier. A DCT statistics-based blind image quality index. *IEEE*

Signal Processing Letters, 17(6):583–586, jun 2010.

[23] Michele A. Saad and Alan C. Bovik. Blind quality assessment of videos using a model of natural scene statistics and motion coherency. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, nov 2012.

[24] Kalpana Seshadrinathan, Rajiv Soundararajan, and Alan Conrad Bovik et al. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, jun 2010.

[25] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference approaches to image and video quality assessment. In *Multimedia Quality of Experience (QoE)*, pages 99–121. John Wiley & Sons, Ltd, nov 2015.

[26] Deepti Ghadiyaram, Janice Pan, and Alan C. Bovik et al. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, sep 2018.

[27] Deepti Ghadiyaram and Alan C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32, jan 2017.

[28] Umesh Rajashekar, Zhou Wang, and Eero P. Simoncelli. Perceptual quality assessment of color images using adaptive signal representation. In *Human Vision and Electronic Imaging XV*. SPIE, feb 2010.

[29] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, jan 2016.

[30] Deepti Ghadiyaram and Alan C. Bovik. Crowdsourced study of subjective image quality. In *48th Asilomar Conference on Signals, Systems and Computers*. IEEE, nov 2014.

[31] Jinling Chen, Denghui Huang, and Huiwen Huang. Content-aware video quality modeling based on neural network. In *2018 IEEE International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, pages 173–176. IEEE, 2018.

[32] Richard Zhang, Phillip Isola, and Alexei A et al. Efros. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[33] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[34] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems 29*, pages 658–666. Curran Associates, Inc., 2016.

[35] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[36] Christian Szegedy, Vincent Vanhoucke, and Sergey et al. Ioffe. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] Du Tran, Lubomir Bourdev, and Rob et al. Fergus. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[38] Joe Yue-Hei Ng and Matthew et al. Hausknecht. Beyond short snippets: Deep networks for video classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[39] Mikko Nuutinen and Toni Virtanen et al. CVD2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, jul 2016.

[40] Deepti Ghadiyaram, Alan C. Bovik, and Hojatollah Yeganeh et al. Study of the effects of stalling events on the quality of experience of mobile streaming videos. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, dec 2014.

[41] Chiunhsiun Lin. Face detection in complicated backgrounds and different illumination conditions by using YCbCr color space and neural network. *Pattern Recognition Letters*, 28(16):2190–2200, dec 2007.

[42] Sooyeon Lee, Youngshin Kwak, and Youn Kim et al. Contrast-preserved chroma enhancement technique using YCbCr color space. *IEEE Transactions on Consumer Electronics*, 58(2):641–645, may 2012.

[43] Christos G. Bampis, Zhi Li, and Alan C. Bovik. Enhancing temporal quality measurements in a globally deployed streaming video quality predictor. In *25th IEEE International Conference on Image Processing (ICIP)*. IEEE, oct 2018.

[44] Hui Men, Hanhe Lin, and Dietmar Saupe. Spatiotemporal feature combination model for no-reference video quality assessment. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, may 2018.

[45] Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. Blind 3d mesh visual quality assessment using support vector regression. *Multimedia Tools and Applications*, 77(18):24365–24386, Sep 2018.

[46] Balasubramanyam Appina, Sathya Veera Reddy Dendi, and K. Manasa et al. Study of subjective quality and objective blind quality prediction of stereoscopic videos. *IEEE Transactions on Image Processing*, pages 1–1, 2019.

[47] Sathya Veera Reddy Dendi and Gokul Krishnappa et al. Full-reference video quality assessment using deep 3d convolutional neural networks. In *2019 National Conference on Communications (NCC)*. IEEE, feb 2019.

[48] Ahmed Aldahdooh, Enrico Masala, and Olivier Janssens et al. Improved performance measures for video quality assessment algorithms using training and validation sets. *IEEE Transactions on Multimedia*, 21(8):2026–2041, aug 2019.

[49] Mohsen Jenadeleh. *Blind Image and Video Quality Assessment*. PhD thesis, Universitat Konstanz, 2018.

[50] Phong V. Vu and Damon M. Chandler. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1):013016, feb 2014.

[51] F. De Simone et al. Subjective assessment of h.264/AVC video sequences transmitted over a noisy channel. In *International Workshop on Quality of Multimedia Experience*, jul 2009.

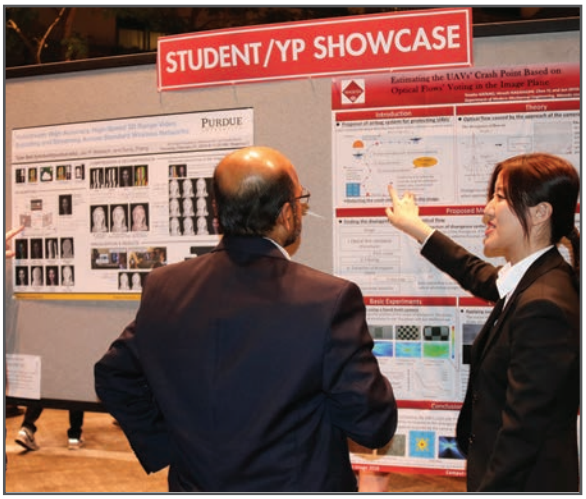
JOIN US AT THE NEXT EI!

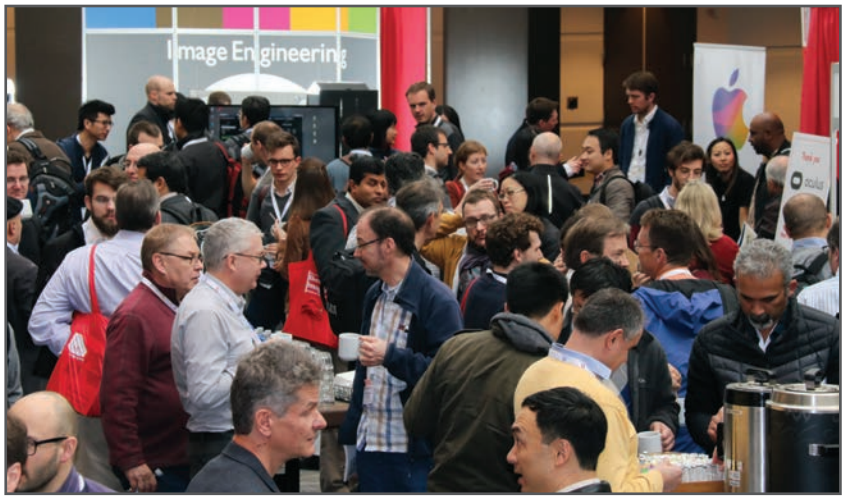
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org