No-Reference Video Quality Assessment Using Space-Time Chips

Joshua P. Ebenezer,[§*] Zaixi Shang,[§*] Yongjun Wu,[†] Hai Wei,[†] Alan C. Bovik[*]
[*] Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin
[†] Amazon Prime Video
joshuaebenezer@utexas.edu, zxshang@utexas.edu, yongjuw@amazon.com, haiwei@amazon.com, bovik@ece.utexas.edu

***Abstract*—We propose a new prototype model for no-reference video quality assessment (VQA) based on the natural statistics of *space-time chips* of videos. Space-time chips (ST-chips) are a new, quality-aware feature space which we define as space-time localized cuts of video data in directions that are determined by the local motion flow. We use parametrized distribution fits to the bandpass histograms of space-time chips to characterize quality, and show that the parameters from these models are affected by distortion and can hence be used to objectively predict the quality of videos. Our prototype method, which we call ChipQA-0, is agnostic to the types of distortion affecting the video, and is based on identifying and quantifying deviations from the expected statistics of natural, undistorted ST-chips in order to predict video quality. We train and test our resulting model on several large VQA databases and show that our model achieves high correlation against human judgments of video quality and is competitive with state-of-the-art models.**

***Index Terms*—Video quality assessment, Space-time chips, Natural Video Statistics**

I. Introduction

Video content accounts for a very large portion of traffic on the Internet and continues to surge in volume, as more people stream content on smartphones, tablets, and high-definition screens. Being able to predict perceived video quality is important to content providers for monitoring and controlling streaming quality and thereby enhance customer satisfaction. Video quality is affected when the video is distorted, which occurs due to a number of reasons as the video is being captured, transmitted, received, and displayed. The task of predicting the quality of a distorted video without a pristine video of the same content to compare it with, which is called no-reference (NR) VQA, is difficult. NR VQA is a hard problem because less information is available than models that use a reference, and conventional notions of signal fidelity are not applicable. Here we describe a new NR VQA algorithm based on the statistics of local windows oriented in space and time.

We briefly review state-of-the-art NR VQA algorithms. V-BLIINDS [1] is a distortion agnostic NR VQA algorithm that models the statistics of DCTs of frame differences to predict video quality. These features are based on models of the human visual system (HVS). HVS-based algorithms posit that natural video statistics have a regularity from which the statistics of distorted videos deviate, and which the HVS is attuned to. Human perceptual judgments of quality are influenced by the degree to which the statistics of distorted videos deviate from those of natural videos. Such models thus attempt to understand and mimic operations of the HVS in order to identify the regularity in natural video statistics. VIIDEO [2] is another algorithm that is belongs to this category. VIIDEO is completely "blind," in the sense that it does not require any training and can be directly deployed. VIIDEO makes use of observed high inter subband correlations of natural videos to predict quality. TLVQM [3] takes a different approach to the problem, where the goal is to define features that capture distortion, and not to characterize naturalness per se. A number of spatiotemporal features are defined at two computational levels, collectively designed to capture a wide range of distortions. TLVQM has about 30 different user-defined parameters, which may affect its performance on databases it was not exposed to when it was designed.

It has also been observed [4] that NR image quality assessment algorithms work reasonably well when applied frame-by-frame to distorted videos of user-generated content because of a lack of temporal variation in such videos. FRIQUEE [5] is a state-of-the-art algorithm for NR IQA that uses a bag of perceptually motivated features. BRISQUE [6] is another NR IQA algorithm that models the statistics of spatially band-passed coefficients of images, motivated by the fact that the early stages of the HVS perform spatial bandpassing. NIQE [7] also models statistics of spatially bandpassed coefficients, but generates an opinion score by quantifying the deviation of a distorted image from the statistical fit to a corpus of natural images and does not require training. CORNIA [8] approaches the problem differently from HVS-based models, and builds a dictionary to represent images effectively for quality assessment.

Videos are spatiotemporal signals and distortions affecting videos can be spatial or temporal or a combination of both. An effective NR VQA algorithm must be able to build a robust representation of the spatiotemporal information in videos. Primary visual cortex (area V1) is implicated in decomposing visual signals into orientation and scale-tuned spatial and temporal channels. V1 neurons are sensitive to specific local orientations of motion. This decomposition is passed on to other areas of the brain, including area middle temporal (MT), where further motion processing occurs.

[§]Equal contribution

978-1-4799-7492-4/15/$31.00 ©2020 IEEE

Extra-striate cortical area MT is known to contain neurons that are sensitive to motion over larger spatial fields [9], [10]. We propose a perceptually motivated NR VQA model, ChipQA, that captures both spatial and temporal distortions, by building a representation of local spatiotemporal data that is attuned to local orientations of motion but is studied over large spatial fields. We show how observed statistical regularities of spatially bandpassed coefficients can be extended temporally and introduce the notion of space-time chips, which follow natural video statistics (NVS). We evaluate the new model on a number of databases and show that we are able to achieve state-of-the-art performance at reasonable computational cost.

II. Proposed Algorithm

A. Space-time Chips

If we consider videos as space-time volumes of data, Space-time (ST) chips can be defined as chips of this volume in any direction or orientation. ST-Chips are similar to space-time slices [11]–[13], but are highly localized in space and time. We are interested in finding the specific directions and orientations of these chips that capture the regularity of natural videos. Consider a frame I_T at time instant T, and its preceding frames $I_{T-T'+1}..I_{T-1}$ from time instant $T-T'+1$ onward, all of size $M \times N$. There are T' frames in this space-time volume. Assume that we have coherent motion vectors available to us at time instant T. To the best of our knowledge, localized groups of pixels are in motion along the directions of these vectors. Assuming T' is small enough, a local spatiotemporal chip of this volume that is oriented perpendicular to the motion vector at each spatial location would capture the local areas that are in motion at each location and across time.

Taking into account the smoothness of motion across time for natural videos, we expect ST-chips to follow similar regularities as we would expect to find for stationary frames. Spatially bandpassed coefficients are known to follow a generalized Gaussian distribution (GGD) in the first order, and their second order statistics can be approximated as following an asymmetric generalized Gaussian distribution (AGGD) [6], [7]. Accordingly, we define the mean subtracted contrast normalized (MSCN) coefficients

$$\hat{I}_T(i,j) = \frac{I_T(i,j) - \mu_T(i,j)}{\sigma_T(i,j) + C} \tag{1}$$

where (i,j) are spatial coordinates and we define the local mean and local variance as

$$\mu_T(i,j) = \sum_{k=-K}^{k=K} \sum_{k=-L}^{k=L} w(k,l) I_T(i+k, j+l) \tag{2}$$

$$\sigma_T(i,j) = \sqrt{\sum_{k=-K}^{k=K} \sum_{k=-L}^{k=L} w(k,l) (I_T(i+k, j+l) - \mu_T(i,j))^2} \tag{3}$$

respectively. $w = \{w(k,l), k \in -K,..,K, l \in -L,..,L\}$ is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and rescaled to unit volume.

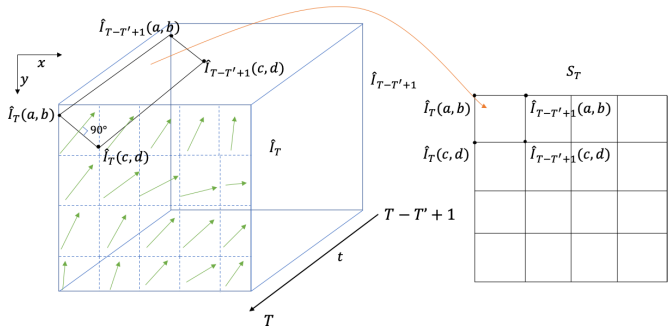


Fig. 1. Extracting ST-chips. On the left is a spatiotemporal volume of frames from time $T-T'+1$ to T. The green arrows represent motion vectors at each spatial patch at time T. ST-chips are cut perpendicular to these across time and aggregated across spatial patches to form the frame on the right at each time instance.

Motion has been used to guide video quality prediction in several full-reference models [9], [14], [15], but few NR models [1], [16]. Given a dense motion flow field at time T, we first find the median of the flow field across a spatial window of size $R \times R$. This gives us a more robust estimate of the flow in each spatial window. We cut the video volume of MSCN coefficients across time from $\hat{I}_{T-T'+1}$ to $\hat{I}_T$ at each window in the direction perpendicular to the motion vectors at each spatially localized window, as shown in Fig. 1, and then aggregate these chips across the spatial windows to form a single "frame" S_T of ST-Chips at each time instance. Each chip is constrained to pass through the center of the spatial $R \times R$ patch, and to be perpendicular to the median flow of that patch. R pairs of x and y coordinates obtained from the relevant line equations are rounded to integers for sampling from the video.In this way, the chips at each patch are uniquely defined by the spatial location of the patch and the direction of the median motion vector. These frames do not have well defined axes because each ST chip is oriented differently in space-time, but they contain important spatiotemporal data. For simplicity, we define $R = T'$. S_T then has dimensions $M' \times N'$, where $M' = T' \lfloor \frac{M}{T'} \rfloor$ and $N' = T' \lfloor \frac{N}{T'} \rfloor$.

Ideally, the directions we have defined are the ones most likely to capture objects in motion. To see this, consider an object in motion in a fixed direction. An ST chip that is perpendicular to the direction of motion of the object is a plane that cuts through the cross-section of the object as it moves along time. A concrete example of this is given in Fig. 2 for a video before its MSCN coefficients are computed.

ST-Chips are collected over all spatial patches of MSCN coefficients for every group of preceding T' frames at each time instant T and are then aggregated. We use the Farneback [17] optical flow algorithm in all our experiments. Due to the smoothness of motion over time and the fact that we expect these chips to have captured natural objects in motion, we hypothesize that this aggregated spatiotemporal data will follow a regular form of natural video statistics. Spatiotemporal distortions are expected to affect the optical flow as well as

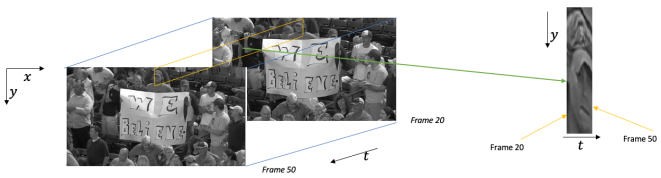


Fig. 2. ST-Chips capture views of objects in motion. In this video, the person in the center moves to their right over time. Consequently, a chip taken in the proximity of their face over this duration and perpendicular to their motion captures their face itself.

spatial data. The motion estimates from optical flow can be distorted for distorted videos, and hence we would expect to see a deviation from the statistics of chips computed from pristine videos. We would also expect to see a deviation in the statistics of the MSCN coefficients collected across time, as distortions affect their regularity across time and across space. Both the distortion of MSCN coefficients and the distortion of optical flow are expected to contribute to a deviation from the statistics of a pristine video, and we confirm this empirically, as shown in Figs. 3, 4, and 5. ST-Chips of MSCN coefficients are found to follow a generalized Gaussian distribution (GGD) of the form:

$$f(x;\alpha;\beta) = \frac{\alpha}{2\beta\Gamma(\frac{1}{\alpha})}\exp(-(\frac{|x|}{\beta})^{\alpha}) \tag{4}$$

where $\Gamma(.)$ is the gamma function:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)dt. \tag{5}$$

The shape parameter α of the GGD and the variance of the distribution are estimated using the moment-matching method described in [18].

We also model the second-order statistics of ST-Chips. Define the collection of ST-Chips aggregated at each time instance T as S_T, and define the pairwise products

$$H_T(i,j) = S_T(i,j)S_T(i,j+1) \tag{6}$$

$$V_T(i,j) = S_T(i,j)S_T(i+1,j) \tag{7}$$

$$D1_T(i,j) = S_T(i,j)S_T(i+1,j+1) \tag{8}$$

$$D2_T(i,j) = S_T(i,j)S_T(i+1,j-1) \tag{9}$$

These pairwise products of neighboring ST chip values along four orientations are modeled as following an asymmetric generalized Gaussian distribution (AGGD), which is given by:

$$f(x;\nu,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp(-(-\frac{x}{\beta_l})^{\nu}) & x<0 \\ \frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp(-(\frac{x}{\beta_r})^{\nu}) & x>0 \end{cases} \tag{10}$$

where

$$\beta_l = \sigma_l\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \tag{11}$$

$$\beta_r = \sigma_r\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \tag{12}$$

ν controls the shape of the distribution and σ_l and σ_r control the spread on each side of the mode. The parameters $(\eta,\nu,\sigma_l^2,\sigma_r^2)$ are extracted from the best AGGD fit to each pairwise product, where

$$\eta = (\beta_r-\beta_l)\frac{\Gamma(\frac{2}{\nu})}{\Gamma(\frac{1}{\nu})}. \tag{13}$$

Videos are affected at multiple scales by distortions, and so all the features defined above are extracted at a reduced resolution as well. Each frame is low-pass filtered and downsampled by a factor of 2. Motion vectors are computed at the reduced scale and ST-Chips are extracted, as described previously, from volumes of MSCN coefficients at the reduced scale.

B. Spatiotemporal Gradient Chips

The spatial gradients of videos contain important information about edges and corners. Distortions are often more noticeable around edges and affect gradient fields. For this reason, we compute the magnitude of the gradient for each frame using the Sobel filter. We then extract ST-Chips from spatiotemporal volumes of the MSCN coefficients of the gradient magnitude at two scales. The first-order and second-order statistics are modeled as GGD and AGGD, respectively, as described earlier for MSCN coefficients of pixel data.

C. Spatial features

Spatial features and a naturalness score are computed frame-by-frame with the image naturalness index NIQE. These capture purely spatial aspects of distortion that may not be completely contained in ST-Chips, and hence boost performance.

D. Quality assessment

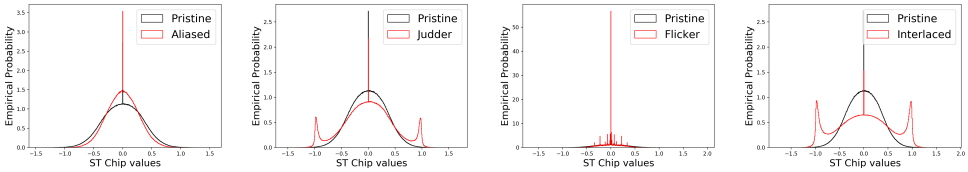
Table I gives a summary of all the features used in the prototype algorithm, ChipQA-0. A total of 109 features are extracted at each time instance, starting from the $T' = 5^{\text{th}}$ frame. Note that the way in which ST-Chips are extracted could vary, and the use of optical flow in this prototype algorithm is just one of many ways in which these chips could be defined for the task of quality assessment.

III. Experiments and Results

A. Databases

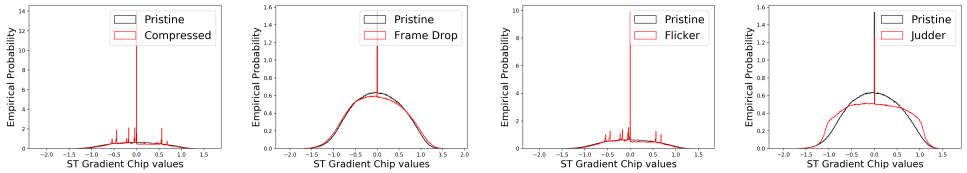
We evaluated our algorithm on four databases, as described in the following section.

1) LIVE-APV Livestream VQA Database: This new database, which we call the LIVE-APV Livestream VQA database, and which will soon be released, was created for the purpose of developing tools for the quality assessment of live streamed videos. The database contains 367 videos, including both synthetically and authentically distorted videos. Of these, 52 videos are authentically distorted videos of different events. The remaining 315 videos were created by applying 6 different distortions on 45 unique contents. The 6 distortions are aliasing, compression, flicker, judder, interlacing, and frame drop. All the videos are of 4K resolution and were shown on a 4K TV in a human study in which 37 subjects participated.



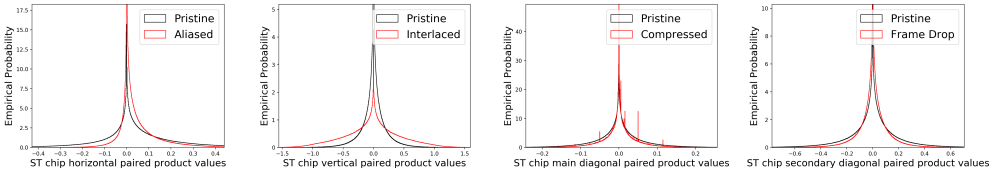
(a) Aliased and pristine (b) Judder and pristine (c) Flicker and pristine (d) Interlaced and pristine

Fig. 3. Empirical distributions of ST-Chips. Pristine (original) distributions are in black and distorted distributions are in red.



(a) Compressed and pristine(b) Frame Drop and pristine (c) Flicker and pristine (d) Judder and pristine.

Fig. 4. Empirical distributions of ST Gradient chips. Pristine (original) distributions are in black and distorted distributions are in red.



(a) Aliased and pristine (b) Interlaced and pristine (c) Compressed and pristine (d) Frame drop and pristine

Fig. 5. Empirical distributions of paired products of ST-Chips. Pristine (original) distributions are in black and distorted distributions are in red.

TABLE I
DESCRIPTIONS OF FEATURES IN CHIPQA.

Domain	Description	Feature index
ST-Chip	Shape and scale parameters from GGD fits at two scales.	$f_1 - f_4$
ST-Chip	Four parameters from AGGD fitted to pairwise products at two scales.	$f_5 - f_{36}$
ST Gradient Chips	Shape and scale parameters from GGD fits at two scales.	$f_{37} - f_{40}$
ST Gradient Chips	Four parameters from AGGD fitted to pairwise products at two scales.	$f_{41} - f_{72}$
Spatial	Features and scores of spatial naturalness index NIQE.	$f_{73} - f_{109}$

2) Konvid-1k: Konvid-1k [19] is a database of 1200 videos of user-generated content with authentic distortions. Many videos in this database do not have significant temporal variation, and it has been found that NR IQA algorithms applied frame-by-frame often achieve high performance on Konvid-1k without making use of temporal information. All videos are of resolution 960×540.

3) LIVE Video Quality Challenge (VQC): The LIVE VQC database [20] has 585 videos of authentically distorted videos, each labeled by an average of 240 human opinion scores. The videos are of user-generated content.

4) LIVE Mobile Database: The LIVE Mobile [21] database has 200 distorted videos created from 10 reference videos. The synthetically applied distortions are compression, wireless-packet loss, temporally varying compression levels, and frame-freezes. This study was conducted on mobile and tablet devices.

B. Training Protocol

The LIVE-APV database contains a mix of synthetically and authentically distorted content. Each algorithm was tested on 1000 random train-test splits, where 80% of the data was used for training, and 20% for testing. 5-fold cross validation was performed while ensuring content separation between training and validation data for each fold. Videos of the same content were not allowed to mix between folds. On the other databases, we implemented a 80-20 train-test split. A support vector regressor (SVR) was used to learn mappings from features to mean opinion scores. Cross validation was used to find the best parameters for the SVR for each algorithm. NIQE and VIIDEO are completely blind algorithms and were not trained, but were evaluated against the test set.

TABLE II
MEDIAN SROCC AND LCC FOR 1000 SPLITS ON THE LIVE-APV LIVESTREAM VQA DATABASE

METHOD	SROCC	LCC
NIQE [7]	0.3395	0.4962
BRISQUE [6] (1 fps)	0.6224	0.6843
HIGRADE [23] (1 fps)	0.7159	0.7388
CORNIA [8] (1 fps)	0.6778	0.7076
TLVQM [3]	0.7597	0.7743
VIIDEO [2]	-0.0039	0.2155
V-BLIINDS [1]	0.7264	0.7646
Spatial	0.6770	0.7370
ST-Chips	0.6742	0.7235
ST Gradient Chips	0.7450	0.7611
ChipQA-0	**0.7802**	**0.8054**

TABLE III
MEDIAN SROCC AND LCC FOR 100 SPLITS ON THE KONVID DATABASE

METHOD	SROCC/LCC
NIQE [7]	0.3559/0.3860
BRISQUE [6] (1 fps)	0.5876/0.5989
HIGRADE [23] (1 fps)	0.7310/0.7390
FRIQUEE [5] (1 fps)	0.7414/0.7486
CORNIA [8] (1 fps)	0.7685/0.7671
TLVQM [3]	**0.7749/0.7715**
VIIDEO [2]	0.3107/0.3269
V-BLIINDS [1]	0.7127/0.7085
ChipQA-0	0.6973/0.6943

We report the Spearman's rank ordered correlation coefficient (SROCC) and the Pearson's linear correlation coefficient (LCC) between the scores predicted by the different algorithms and the subjective mean opinion scores. The predicted score was passed through the non-linearity described in [22] before the LCC was computed. We report results for 1000 splits for all algorithms on the LIVE-APV database, and 100 splits on all other databases. Results are shown in Tables II, III, IV, and V. We did not compute NR IQA features for every frame of each video in the databases, but computed them for at least 1 frame every second for each video. We averaged the features obtained by the NR IQA algorithms across frames and trained them with an SVR to map to mean opinion scores.

TABLE IV
MEDIAN SROCC AND LCC FOR 100 SPLITS ON THE LIVE MOBILE DATABASE

METHOD	SROCC/LCC
BRISQUE [6] (1 fps)	0.4876/0.5215
VIIDEO [2]	0.2751/0.3439
VBLIINDS [1]	0.7960/0.8585
TLVQM [3]	**0.8247/0.8744**
ChipQA-0	0.7898/0.8435

C. Performance

ChipQA-0 performs better than competing algorithms on the new LIVE-APV database, in which there are a number of commonly occurring temporal distortions, such as interlacing, judder, frame drop, and temporal variation of compression levels. We found the individual performance of each feature space on the LIVE-APV database, and the results are shown in Table II. It is clear that ST-Chips and ST Gradient chips provide quality-aware information that boosts the performance of the algorithm on fast-moving, livestreamed videos. Studies have shown that UGC videos are dominated by spatial distortions [24]. Both Konvid-1k and LIVE VQC are known to not contain significant temporal variation and therefore NR IQA algorithms perform well on them without the need for temporal information or processing. Nevertheless, ChipQA-0 achieves competitive performance on these databases as well. It also performs competitively on the LIVE Mobile database, which has a mix of spatial and temporal distortions.

We perform a one-sided t test on the 1000 SROCCs of the various algorithms on the live streaming database with a 95% confidence level to evaluate the statistical signifiance of the results. The results show that ChipQA-0 is statistically superior to all other algorithms on the LIVE-APV Livestream VQA database.

TABLE V
MEDIAN SROCC AND LCC FOR 100 SPLITS ON THE LIVE VQC DATABASE

METHOD	SROCC/LCC
BRISQUE [6] (1 fps)	0.6192/0.6519
VIIDEO [2]	-0.0336/-0.0064
VBLIINDS [1]	0.7005/0.7251
TLVQM [3]	**0.8026/0.7999**
ChipQA-0	0.6692/0.6965

TABLE VI
COMPUTATION TIME FOR A SINGLE 3840X2160 VIDEO WITH 210 FRAMES FROM THE LIVE-APV LIVESTREAM VQA DATABASE

METHOD	Time (s)
BRISQUE [6]	273
HIGRADE [23]	14490
CORNIA [8]	1797
FRIQUEE [5]	924000
VIIDEO [2]	4950
VBLIINDS [1]	10774
TLVQM [3]	892
ChipQA-0	2284

D. Computational cost

Table VI shows the computation times required to extract features for each algorithm on a single 4K video from the LIVE-APV database. Costs for the IQA algorithms were estimated by multiplying the computation time for a single frame by the total number of frames in the video. VIIDEO, VBLIINDS, and ChipQA-0 were implemented with Python. The other algorithms were implemented with MATLAB®. It is not possible to directly compare these numbers because they were implemented on different platforms with different optimization strategies, but they can serve as a rough estimate, and ChipQA-0 is reasonably efficient. It was run on an Intel i9 9820X CPU with 10 cores and a maximum frequency of

TABLE VII
RESULTS OF ONE-SIDED T-TEST PERFORMED BETWEEN SROCC VALUES OF VARIOUS ALGORITHMS ON THE LIVE-APV DATABASE. '1' ('-1') INDICATES THAT THE ROW ALGORITHM IS STATISTICALLY SUPERIOR (INFERIOR) TO THE COLUMN ALGORITHM. THE MATRIX IS SYMMETRIC

METHOD	NIQE	BRISQUE	HIGRADE	CORNIA	TLVQM	VIIDEO	V-BLIINDS	ChipQA-0
NIQE	-	-1	-1	-1	-1	1	-1	-1
BRISQUE	1	-	-1	-1	-1	1	-1	-1
HIGRADE	1	1	-	1	-1	1	-1	-1
CORNIA	1	1	-1	-	-1	1	-1	-1
TLVQM	1	1	1	1	-	1	1	-1
VIIDEO	-1	-1	-1	-1	-1	-	-1	-1
V-BLIINDS	1	1	1	1	-1	1	-	-1
ChipQA-0	1	1	1	1	1	1	1	-

4.1 GHz. All other algorithms were run on a AMD Ryzen 5 3600 with a maximum frequency of 4.2 GHz.

IV. CONCLUSION

We have proposed the novel concept of ST-Chips, and defined how they are extracted and described why they are relevant to video quality. We used the statistics of these chips to model 'naturalness' and deviations from naturalness, and proposed parameterized statistical fits to their statistics. We further used the parameters from these statistical fits to map videos to subjective opinions of video quality without explicitly finding distortion-specific features and without reference videos. We showed that our prototype distortion-agnostic, no-reference video quality assessment algorithm, ChipQA-0, is highly competitive with other state-of-the-art models on a number of databases. We continue to refine the model, with one aim being to eliminate the need for an optical flow algorithm.

V. ACKNOWLEDGMENTS

The authors thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported in this paper. URL: http://www.tacc.utexas.edu.

REFERENCES

[1] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[2] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

[3] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, 2019.

[4] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *arXiv preprint arXiv:2005.14354*, 2020.

[5] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vision*, vol. 17, no. 1, pp. 32–32, 2017.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2012.

[8] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[9] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, 2009.

[10] M. Wainwright, O. Schwartz, and E. Simoncelli, *Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons*. MIT Press, 2002, pp. 203–222.

[11] P. Yan and X. Mou, "Video quality assessment based on motion structure partition similarity of spatiotemporal slice images," *J. Electron. Imag.*, vol. 27, no. 3, pp. 1 – 13, 2018.

[12] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, pp. 1 – 25, 2014.

[13] P. Yan, X. Mou, and W. Xue, "Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices," in *Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications*, vol. 9411. SPIE, 2015, pp. 182 – 191.

[14] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2007, pp. I–869.

[15] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, 2016.

[16] K. Manasa and S. Channappayya, "An optical flow-based no-reference video quality assessment algorithm," in *IEEE Int. Conf. Image Processing*, 2016, pp. 2400–2404.

[17] G. Farneback, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," in *IEEE Int. Conf. Computer Vision*, 2001, pp. 171–177.

[18] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, 1995.

[19] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (Konvid-1k)," in *Int. Conf. Quality of Multimedia Experience*, 2017, pp. 1–6.

[20] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2018.

[21] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, 2012.

[22] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[23] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, 2017.

[24] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," *arXiv preprint arXiv:2002.10651*, 2020.