No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics

Article *in* IEEE Transactions on Image Processing · April 2020

DOI: 10.1109/TIP.2020.2984879

2 authors:

Dendi Sathya Veera Reddy
Indian Institute of Technology Hyderabad
8 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Sumohana Channappayya
Indian Institute of Technology Hyderabad
88 PUBLICATIONS **880** CITATIONS
SEE PROFILE

No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics

Sathya Veera Reddy Dendi, Sumohana S. Channappayya, *Member*, *IEEE*

***Abstract*—Robust spatiotemporal representations of natural videos have several applications including quality assessment, action recognition, object tracking etc. In this paper, we propose a video representation that is based on a parameterized statistical model for the spatiotemporal statistics of mean subtracted and contrast normalized (MSCN) coefficients of natural videos. Specifically, we propose an asymmetric generalized Gaussian distribution (AGGD) to model the statistics of MSCN coefficients of natural videos and their spatiotemporal Gabor bandpass filtered outputs. We then demonstrate that the AGGD model parameters serve as good representative features for distortion discrimination. Based on this observation, we propose a supervised learning approach using support vector regression (SVR) to address the no-reference video quality assessment (NRVQA) problem. The performance of the proposed algorithm is evaluated on publicly available video quality assessment (VQA) datasets with both traditional and in-capture/authentic distortions. We show that the proposed algorithm delivers competitive performance on traditional (synthetic) distortions and acceptable performance on authentic distortions. The code for our algorithm will be released at https://www.iith.ac.in/~lfovia/downloads.html.**

***Index Terms*—Natural scene statistics of videos, spatiotemporal Gabor filters, human visual system (HVS), SVR and 3D-MSCN.**

I. Introduction

We are in the midst of a video-driven data consumption revolution. The scale of video consumption has been rapidly increasing due to the availability of high speed internet connectivity on handheld devices, technological advancement in acquisition systems and improved storage capacities. This in turn has led to a massive stress on existing network infrastructure and resources. The role of quality assessment in optimal resource allocation cannot be over emphasized. While subjective VQA is more appropriate, reliable and accurate, it is not always feasible because of its time consuming and expensive nature. An alternative approach is to rely on robust objective VQA method for the quality assessment (QA) task.

Based on Video Quality Expert Group (VQEG) [1] recommendations, objective video quality assessment techniques are classified into three types: (i) full-reference video quality assessment (FRVQA) where access to both reference and test video is available, (ii) reduced-reference video quality assessment (RRVQA) where only partial access to the reference video is available along with test video and (iii) no-reference video quality assessment (NRVQA) where no information about the reference video is available for the QA task. In this work, we focus on NRVQA algorithm design given its practical utility in a wide range of applications.

The authors are with the Lab for Video and Image Analysis (LFOVIA), Department of Electrical Engineering, Indian Institute of Technology Hyderabad, Kandi 502285, India (e-mail: {ee16resch01003, sumohana}@iith.ac.in)

Advancements in vision science research have led to important discoveries about the human visual system (HVS), specifically about the HVS architecture and it's functional mechanisms. Hubel and Wiesel's [2] seminal work classified the HVS into multiple regions and named the primary visual cortex as V1 which is located in the occipital lobe. Sekuler et al. [3] state that most of the neurons in the V1 area respond to contour motion in specific directions and that different neurons have different preferred directions. The output of these V1 neurons is fed to subsequent stages like V2, V3, and area middle temporal (MT). The second stage region like area MT processes the motion information, based on the input it receives directly from V1 and indirectly from other regions like V2 and V3 as well. Almost all the neurons in the area MT are selective for the direction and motion of the visual signals.

Videos are spatiotemporal signals which have both spatial and motion information. Since most of the spatial information is processed in the V1 region and motion information is processed in the MT area, it is essential to model both V1 and area MT for a better analysis of video signals. Adelson et al. [4] suggest that the responses of both V1 and area MT can be approximated well with spatiotemporal Gabor filters. Given these observations, we rely on spatiotemporal Gabor filters in our analysis.

Specifically, we study the spatiotemporal representations of natural videos at multiple spatiotemporal resolutions. The spatiotemporal representations are extracted by modeling the 3D mean subtract contrast normalized (MSCN) coefficients and spatiotemporal Gabor filter responses of 3D-MSCN natural videos. We observe that the empirical distributions of 3D-MSCN coefficients and spatiotemporal Gabor filter responses are unimodal and sensitive to distortions in the natural videos. We model these empirical distributions using an Asymmetric Generalized Gaussian Distribution (AGGD) and the parameters of the AGGD serve as key features to estimate the perceptual quality of the video in the no-reference (NR) setting.

Our contributions in this paper are:

- Modeling of 3D-MSCN coefficients of natural videos using AGGD;
- Modeling of spatiotemporal Gabor bandpass filter responses of 3D-MSCN natural videos using AGGD;
- Design of an NRVQA algorithm using spatiotemporal scene statistics of the videos.

The rest of the paper is organized as follows: related work is presented in Section II, followed by modeling of natural videos in Section III. The proposed NRVQA technique is discussed in Section IV and results are presented and discussed in Section V, followed by concluding remarks in Section VI.

II. Related Work

In this section, we review related work on objective VQA techniques. A straightforward way of solving the VQA problem is to consider the frames of the video as images and apply image quality assessment (IQA) metrics to each frame and pool the frame level quality scores. The IQA literature is rich with excellent algorithms like SSIM [5], MS-SSIM [6] and FSIM [7] in the full reference setting, and with methods like NIQE [8], DIIVINE [9], C-DIIVINE [10], BRISQUE [11], FRIQUEE [12], QAC [13] in the no-reference setting. While this approach is somewhat effective, performance gains resulting from the use of motion information for VQA has been clearly demonstrated in the literature [14]–[16].

We now briefly review VQA techniques which incorporate both spatial and motion/temporal information. Popular FRVQA algorithms like the MOVIE index [14] and FLOSIM [15] are based on the optical flow of videos. The MOVIE index [14] quantifies the error in the optical flow planes of the distorted and reference video over several spatiotemporal frequency bands. It then pools these errors to form the perceptual quality estimate of the distorted video. Optical flow in MOVIE is computed using spatiotemporal Gabor filters [17]. FLOSIM [15] is based on local optical flow statistics and these statistics are shown to be sensitive to distortions in videos. The deviation of test video optical flow statistics from pristine video optical flow statistics is quantified as the perceptual quality of test video. Optical flow in FLOSIM is estimated using two popular techniques: Black and Anandan [18] and Farneback [19].

Ortiz et al. [20] proposed an FRVQA technique using optical flow based motion information to extract temporal distortions. These temporal distortions are used to estimate the perceptual quality of a test video with respect to it's reference video. Kim et al. [21] proposed a FRVQA technique using deep convolutional neural networks called DeepVQA. DeepVQA takes a distorted frame, the spatial error map, the frame difference map and the temporal error map as input that is regressed to subjective score using average pooling. VQM_VFD [16] is another popular FRVQA technique which quantifies the perceptual quality of videos in terms of temporal distortions due to frame delays. VQM_VFD features are extracted using simple edge detection filters and a neural network is trained on these features. Video VIF [22] and V-IFC [23] are information theoretic FRVQA techniques.

Gunawan et al. [24] proposed an RRVQA technique using the harmonic strength of edges present in video frames. Harmonic gain and loss information is computed from the edge detected frames and is used in the overall quality estimation. ST-RRED [25] is a popular RRVQA technique based on the spatial and temporal entropic differences. ST-RRED takes a hybrid approach to combine statistical models and perceptual principles to come up with a quality assessment technique. It models wavelet coefficients of frames and frame differences of both reference and distorted videos using the Gaussian Sale Mixture (GSM) model. It then applies entropic differences to quantify the perceptual quality of a distorted video. Zeng et al. [26] proposed an RRVQA technique based on temporal motion smoothness by examining the temporal variations of local phase structures in the complex wavelet transform domain.

The Discrete Cosine Transform (DCT) [27] is a popular transform used to analyze images and video signals, and its application to quality assessment is also well-studied. Brandao et al. [28] proposed a two step approach to measure H.264/advanced video coding distortions. The first step is error estimation and the next is perceptual weighting of this error. This method is proposed in the DCT domain based on the hypothesis that the quantization noise corrupts the DCT coefficients. The DCT coefficients are modeled using probability density functions like Cauchy or Laplace and the parameters of these density functions are estimated using maximum likelihood. In general, approaches that model the statistics of natural scenes are called natural scene statistics (NSS) models. The overall quality of the video is estimated using spatiotemporal contrast sensitivity function as the weighting function. Video BLIINDS [29] extends the DCT based NSS model of images to videos. It quantifies motion coherency in video sequences to design a blind VQA algorithm using support vector regression (SVR). In brief, 2D DCT coefficients of successive frame differences of a video are modeled using a statistical model. These statistics are mapped to perceptual quality scores using SVR. Li et al. [30] proposed a NRVQA metric using spatiotemporal scene statistics of natural videos in the 3D DCT domain. The scene statistics of the videos in the 3D DCT domain are modeled using the generalized Gaussian distribution (GGD) and the scene statistics based features are mapped to video quality scores using linear SVR.

Manasa et al. [31] proposed an NRVQA metric using local and global statistics of optical flow. The distortions in the video are captured in the form of flow statistics. The flow statistics are in turn used to train an SVR to map the flow statistics to the perceptual quality of the video. SACONVA [32] is an NRVQA technique designed based on the 3D shearlet transform and a convolutional neural network (CNN). The 3D shearlet transform is used to extract the natural scene statistics of the videos and a CNN with logistic regression is employed to pool the quality score. Liu et al. [33] proposed an NRVQA technique dubbed V-MEON by merging the traditional two-stage approach i.e., feature extraction and regression. V-MEON uses deep neural networks to jointly optimize the feature extraction and regression stages. Korhonen et al. [34] proposed an NRVQA technique by using a two-level approach. Low complexity features are extracted from the full video sequence and high complexity features are extracted from keyframes. The keyframes are identified using low complexity features. These features are mapped to perceptual quality scores using SVR and random forest.

Shabeer et al. [35] proposed an NRVQA technique based on the spatiotemporal statistics of sparse representations. In this approach, the authors have learned a 3D dictionary with spatiotemporal video volumes using the KSVD [36] algorithm.

The sparse representations of videos are modeled using a statistical model whose parameters used to train an SVR to map the statistics to perceptual quality levels of the videos. Xu et al. [37] proposed an NRVQA algorithm based on frame-level unsupervised feature learning and hysteresis temporal pooling. These features are used to train an SVR to predict the perceptual quality score.

Vega et al. [38] proposed a deep learning based video quality assessment technique for video streaming settings. A deep unsupervised learning based model is employed at the server end to extract features and a light weight and computationally effective no-reference metric is employed at the client side. Zhang et al. [39] proposed a NRVQA technique using a convolutional neural network (CNN) and score mapping function. This technique transforms the natural videos using 3D-DCT and the transforms coefficients are used to train a CNN with target labels obtained using FRVQA techniques. This pre-trained model is used to find block-wise scores and these scores are mapped to perceptual quality scores using a frequency histogram mapping function. You et al. [40] proposed an NRVQA technique using a 3D convolution network (3D-CNN) and long short-term memory (LSTM). A 3D-CNN used to extract local spatiotemporal features from small cubic clips in the video. These features are fed to an LSTM to predict the perceptual quality of the video.

Caviedes et al. [41] proposed an NRVQA algorithm based on sharpness, contrast, noise, clipping, ringing, and blocking artifacts. Farias et al. [42] measured video quality by analyzing the blockiness, blurriness, and noisiness in videos. Yang et al. [43] proposed an NRVQA method based on the temporal dependency between adjacent frames of the videos. An attempt to develop a completely blind VQA algorithm was made in VIIDEO [44] by exploiting the statistical regularities in natural videos. In VIIDEO, consecutive frame differences are used to understand the statistical behavior of the videos with and without distortions. These statistical observations are used to measure the perceptual quality of a natural video in the blind setting.

We also briefly reviewed literature about natural scene statistics of videos. Dong and Atick [45] measured the spatiotemporal correlations using the power spectrum of natural time-varying images and defined the relation between spatial and temporal frequencies. van Hateren and Ruderman [46] observed that independent component analysis on natural image sequences results in spatiotemporal filters similar to simple cells of the primary visual cortex qualitatively. Olshausen et al. [47] showed a way to adapt the over-complete dictionary of the space-time function to represent the natural videos with maximum sparsity. Wang and Li [48] studied the temporal variation of local phase structures in the complex wavelet transform domain and observed the strong prior of temporal motion smoothness of natural image sequences. These statistical regularities are exploited in designing a reduced reference video quality assessment technique. Varghese and Wang [49] proposed a video denoising technique in the wavelet transform domain by building a spatiotemporal Gaussian scale mixture model. These models capture the local correlations between the wavelet coefficients of natural videos across space and time. These works demonstrate ways of modeling the statistics of natural videos and how these models can be applied to video quality assessment and denoising problems.

The proposed approach is similar in philosophy to BRISQUE [50]. While our motivation is indeed the success of natural scene statistics in no-reference image quality assessment (NRIQA), to the best of our knowledge, this is the first work to study the video level scene statistics using spatiotemporal Gabor filters. The approaches in Li et al. [30] and Shabeer et al. [35] are similar to the proposed approach in terms of the framework. While these methods have used 3D DCT and 3D dictionary based scene statistics modeling, in the proposed approach we use 3D-MSCN and spatiotemporal Gabor filter based scene statistics modeling. As discussed earlier, the reason for using 3D Gabor filters is that they model the V1 and MT areas of the HVS well compared to the other approaches. We present the proposed statistical model for the natural videos next.

III. Modeling of natural videos

A. 3D-MSCN Coefficients of Natural Videos

In the case of natural images, it is well known that the local mean subtracted contrast normalized (MSCN) coefficients of pristine natural images are modeled conveniently by a Gaussian distribution [50]. Further, it has been shown that the distribution of the MSCN coefficients of distorted images deviates from a Gaussian distribution as distortion strength increases. These distributions are modeled well either by a generalized Gaussian distribution (GGD) or an asymmetric generalized Gaussian distribution (AGGD). The model parameters of GGD and AGGD have been extensively used in designing NRIQA techniques and have shown state-of-the-art performance [8]–[12]. The primary reason to work with MSCN images is due to the decorrelation of local pixel dependency. Similarly, natural videos have high correlation among neighboring pixels both in space and time. To decorrelate such local dependency in natural videos, we propose to apply 3D-MSCN. The 3D-MSCN coefficient of a natural video V is defined as

$$\hat{V}(x,y,t) = \frac{V(x,y,t) - \mu(x,y,t)}{\sigma(x,y,t) + 1}, \tag{1}$$

where, $x \in \{1,2,\ldots,M\}$, $y \in \{1,2,\ldots,N\}$ are spatial indices and M, N are the height and width of the video frame respectively. $t \in \{1,2,\ldots,T\}$ is the temporal index and T is number of frames in a video. A constant value 1 is added to the denominator to avoid instabilities when $\sigma(x,y,t)$ becomes very small or zero,

$$\mu(x,y,t) = \sum_{j=-J}^{J}\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{j,k,l} V(x+j, y+k, t+l) \tag{2}$$

$$\sigma(x,y,t) = \sqrt{\sum_{j=-J}^{J}\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{j,k,l}[V(x+j,y+k,t+l) - \mu(x,y,t)]^2} \tag{3}$$

and $w = [w_{j,k,l} | j = -J,\ldots,J, k = -K,\ldots,K, l = -L,\ldots,L]$ is a symmetric normalized 3D Gaussian filter with

zero mean and standard deviation of 1.166. In our experiments we chose $J = K = L = 2$, since it provided a good performance complexity trade off.

empirical distribution of a natural pristine video bf_{org} from the LIVE Mobile [51] VQA dataset before and after 3D-MSCN is shown in Figure 1. It is clear that applying 3D-MSCN to the natural video results in its distribution changing from multimodal to unimodal. Importantly, in Figure 5(a), we show that the distributions of 3D-MSCN coefficients vary with respect to perceptual video quality levels. Again, videos from the LIVE Mobile VQA dataset are considered to demonstrate the above observation. As in Figure 1, bf_{org} is a pristine video, bf_{r1}, bf_{r2}, bf_{r3} and bf_{r4} are compressed version of bf_{org} with decreasing levels of H.264 compression. This results in the quality of bf_{r1} being lower than bf_{r4} and the corresponding DMOS scores of these videos are given in Figure 2.

B. Bandpass Filtered Natural Videos

Natural videos are time-varying spatial signals and the information in these signals is attributed to the motion of the objects in successive frames along with the spatial variations in pixels. It has been hypothesised that the HVS employs spatiotemporal bandpass filters to analyse and process natural videos, and specifically so in modeling the V1 region and area middle temporal (MT) of the primary visual cortex of HVS. As discussed in Section I , the proposed approach is motivated by this HVS hypothesis that spatiotemporal Gabor filters are a good approximation of the bandpass behavior of the HVS. The design of the filter follows the work by Petkov and Subramanian [52], and is detailed in the following.

The spatiotemporal Gabor filter $g_{v,\theta,\varphi}(x,y,t)$ is defined as the product of a Gaussian envelope function that limit the spatial extent of $g_{v,\theta,\varphi}(x,y,t)$, a cosine wave moving with phase speed v (pixel per frame) in the θ direction, a Gaussian function that is dependent only on time t and determines the decay along time of $g_{v,\theta,\varphi}(x,y,t)$.

The impulse response of a Gabor filter is given by

$$g_{v,\theta,\varphi}(x,y,t) = \frac{\gamma}{2\pi\sigma^2}\exp\left(\frac{-((\overline{x}+v_c t)^2+\gamma^2\overline{y}^2)}{2\sigma^2}\right). \cos\left(\frac{2\pi}{\lambda}(\overline{x}+vt)+\varphi\right).\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(t-\mu_t)^2}{2\tau^2}\right) \quad (4)$$

where $\overline{x} = x\cos(\theta)+y\sin(\theta)$, $\overline{y} = -x\sin(\theta)+y\cos(\theta)$ and γ is the spatial aspect ratio which is the ellipticity of the spatial domain Gaussian envelope. σ indicates the standard deviation of the Gaussian factor which defines the size of the receptive field. λ is the wavelength of the cosine factor. v_c is the speed at which center of the spatial Gaussian envelope moves in the direction of $\overline{x}$-axis. The angle parameter $\theta \in [0, 2\pi)$ specifies the preferred orientation of the filter. The velocity factor v specifies the phase speed of the cosine term and its speed. The phase offset φ is used to generate quadrature pair of filters by setting $\varphi = 0$ and $\varphi = \pi/2$. The other Gaussian function with mean μ_t and standard deviation τ is used to model the intensity changes in time. The choice of the $g_{v,\theta,\varphi}(x,y,t)$ parameters are motivated from HVS [52] and we chose the value of aspect ratio $\gamma = 0.5$, $\sigma/\lambda = 0.56$, $\mu_t = 1.75$ and $\tau = 2.75$. We set $v_c = v$ to obtain filter with moving envelope of velocity v. Wavelength $\lambda = \lambda_0\sqrt{1+v^2}$, where we choose $\lambda_0 = 2$. The half-response spatial frequency bandwidth b (in octaves) and the ratio σ/λ are related as:

$$\frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\frac{2^b+1}{2^b-1}. \quad (5)$$

The spatiotemporal bandpass filter size is $(2n+1)\times(2n+1)\times(2n+1)$, where $n = \lceil 7*\sigma \rceil$. The sample spatiotemporal bandpass filters with velocity factor $v = 2$ (in pixels per frame), angle parameter $\theta = \pi/3$, moving envelope speed $v_c = v$, bandwidth $b = 1$ octave, and $\varphi = 0$ and $\varphi = \pi/2$ are shown in Figure 3. With the above set of parameters, the spatiotemporal filter resolution becomes $37 \times 37 \times 37$. In Figure 3, we show the sample frames from 19 to 24 to demonstrate how the spatiotemporal filters vary in space and time. These filters enable us to extract the bandpass spatiotemporal representations of the video signals.

In this work, the bandpass spatiotemporal Gabor filters are generated by varying the three key parameters (v, θ and φ) of the impulse response of the spatiotemporal Gabor function $g_{v,\theta,\varphi}(x,y,t)$. We choose 3 different speeds ($v \in \{0,1,2\}$), 4 different orientation ($\theta \in \{0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi\}$) and phase offset $\varphi = 0$ for symmetry and $\varphi = \pi/2$ for anti-symmetry. Due to computational complexity, we could not consider finer level of v and θ. Specifically, if envelope velocity v increases, the wavelength λ increases, which has direct relation with standard deviation σ of the Gaussian envelope. The higher the value of σ the higher the filter size which automatically increases the computational complexity of 3D-convolution.

The bandpass filter response of a 3D-MSCN normalized video is defined as

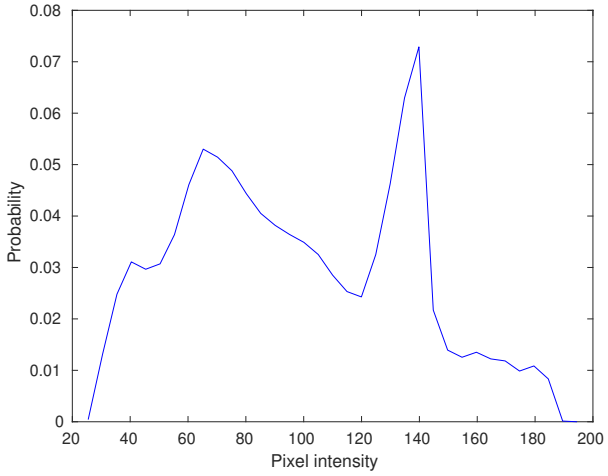
$$r_{v,\theta,\varphi}(x,y,t) = \hat{V}(x,y,t) * g_{v,\theta,\varphi}(x,y,t), \quad (6)$$

where $*$ is the 3D-convolution operator, $\hat{V}(x,y,t)$ is the 3D-MSCN normalized video, $g_{v,\theta,\varphi}(x,y,t)$ is a spatiotemporal Gabor filter and $r_{v,\theta,\varphi}(x,y,t)$ is the bandpass filtered output of the 3D-MSCN normalized video.

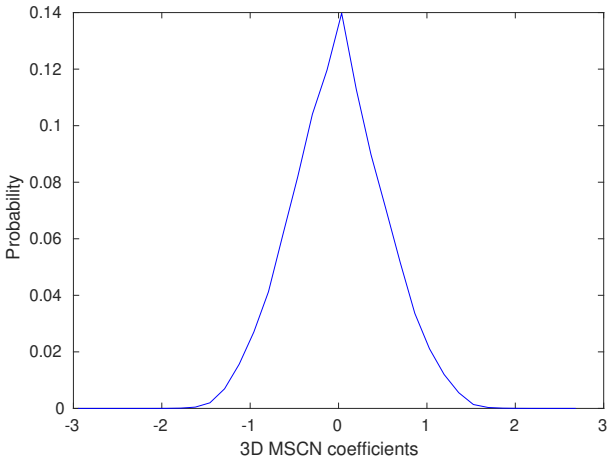
C. Proposed Statistical Model

In this section, we present a statistical model for 3D-MSCN and spatiotemporal bandpass filter responses of natural videos. In Figure 2, we show a pristine video with DMOS score of 0 and four distorted videos at different quality levels. The distorted videos are encoded using H.264 compression and are arranged in increasing order of perceptual quality (can be observed with their DMOS scores as well as the tree branches). In Figure 4, we plot the empirical distribution (normalized histogram) of these videos using 3D-MSCN coefficients and their spatiotemporal Gabor filter responses. Both 3D-MSCN coefficients distributions and spatiotemporal Gabor filter response distributions of natural videos clearly vary with respect to their perceptual quality levels. We note that these plots correspond to the entire video sequence. Similarly, Figure 5 shows the empirical distribution of 3D-MSCN coefficients and spatiotemporal Gabor filter response of natural videos across

(a) Frame from a pristine video (bf_{org}) in the LIVE Mobile VQA dataset [51].



(b) Density curve of a natural pristine video (bf_{org}) from the LIVE Mobile VQA dataset [51].



(c) Density curve of 3D-MSCN coefficients of a natural pristine video (bf_{org}) from the LIVE Mobile VQA dataset [51].

Fig. 1: Empirical distribution of a natural video before and after 3D-MSCN: Figure (a) is a sample frame of the natural video (bf_{org}), Figure (b) is density curve of pixel intensities of bf_{org} and Figure (c) is 3D-MSCN coefficient density curve of the video bf_{org}.

different types and levels of distortions. Specifically, H.264 compression, temporal dynamics, rate adaptation, and wireless packet loss distorted versions of bf_{org} are considered. These distributions also vary with respect to their perceptual quality levels, as can be observed in Figure 5 (by comparing with their DMOS scores).

We model the distributions of the $\hat{V}(x,y,t)$ and $r_{v,\theta,\varphi}(x,y,t)$ using AGGD. Our motivation to use a AGGD for modeling the distribution of the MSCN coefficients comes from the IQA literature where it has been used very successfully. A case in point is the NIQE index [8]. Further, the AGGD is a flexible model that allows for effectively representing a large variety of unimodal data (with different peaks and tails) using only three parameters.

$$f(x;\gamma,\beta_l,\beta_r) = \begin{cases} \frac{\gamma}{(\beta_l+\beta_r)\Gamma(\frac{1}{\gamma})} \exp\left(-\left(\frac{-x}{\beta_l}\right)^{\gamma}\right); \forall x \leq 0 \\ \frac{\gamma}{(\beta_l+\beta_r)\Gamma(\frac{1}{\gamma})} \exp\left(-\left(\frac{x}{\beta_r}\right)^{\gamma}\right); \forall x > 0, \end{cases} \tag{7}$$

where γ, β_l, β_r are the shape parameters. The parameters of the AGGD are estimated using the moment estimation method [53]. Here, $\Gamma(\cdot)$ is defined as,

$$\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}dt; a > 0. \tag{8}$$

The shape parameters γ, β_l, β_r are used in designing the proposed objective quality assessment technique, as discussed next.

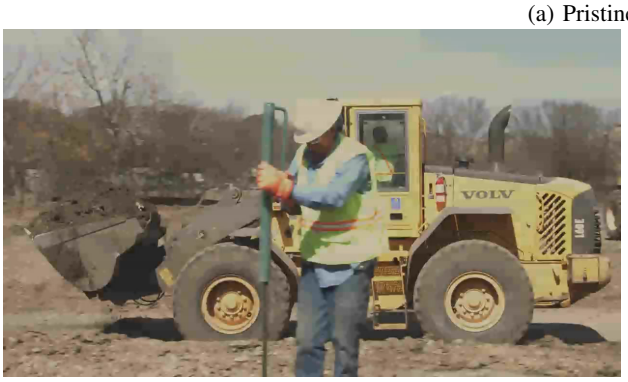
IV. No-reference video quality assessment

The first step in the proposed approach is feature extraction using 3D-MSCN coefficients and spatiotemporal Gabor filter responses of the natural videos. Let the scene statistics of the natural video corresponding to 3D-MSCN coefficients be denoted by the vector

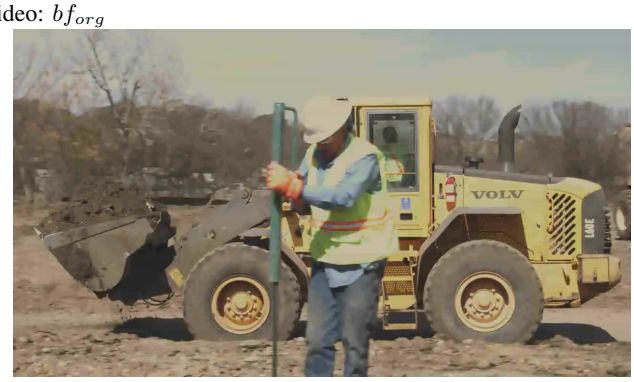
$$\hat{f}_{mscn} = [\gamma, \beta_l, \beta_r, \eta], \tag{9}$$

where $\eta = \frac{\gamma}{\beta_l+\beta_r}$. We would like to reiterate that these features correspond to an entire video sequence. The choice of the feature element η is empirical and it has shown a small

(a) Pristine video: bf_{org}



(b) bf_{r1}, DMOS = 3.24375



(c) bf_{r2}, DMOS = 2.09375

(d) bf_{r3}, DMOS = 1.04375

(e) bf_{r4}, DMOS = 0.35625

Fig. 2: Sample videos from LIVE Mobile [51] VQA dataset: (a) bf_{org} is a pristine video and (b)-(e) are H.264 compressed videos (bf_{r1} to bf_{r4}) arranged in increasing order of perceptual quality.

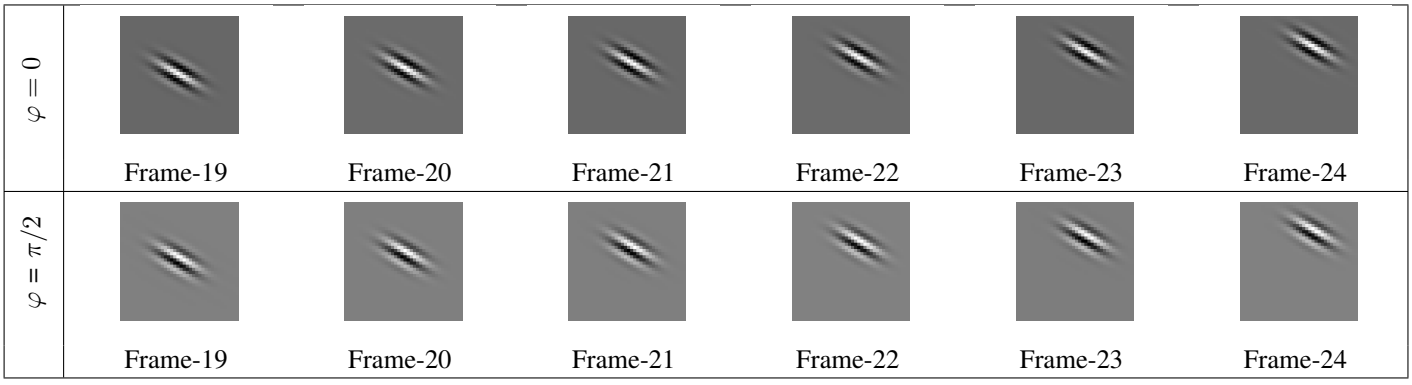
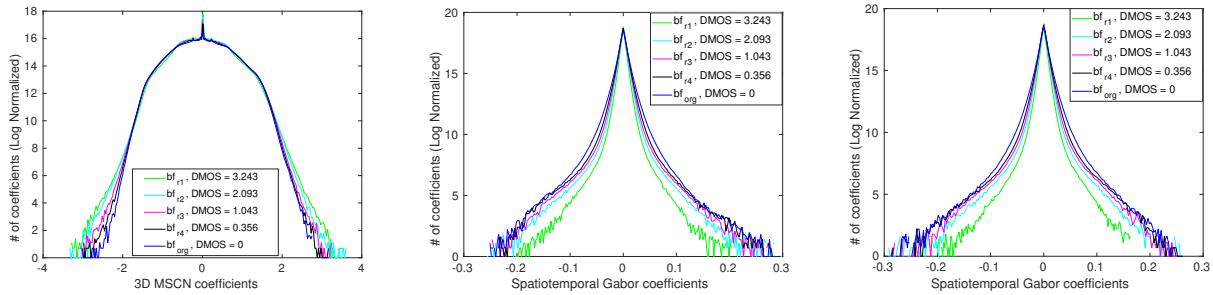
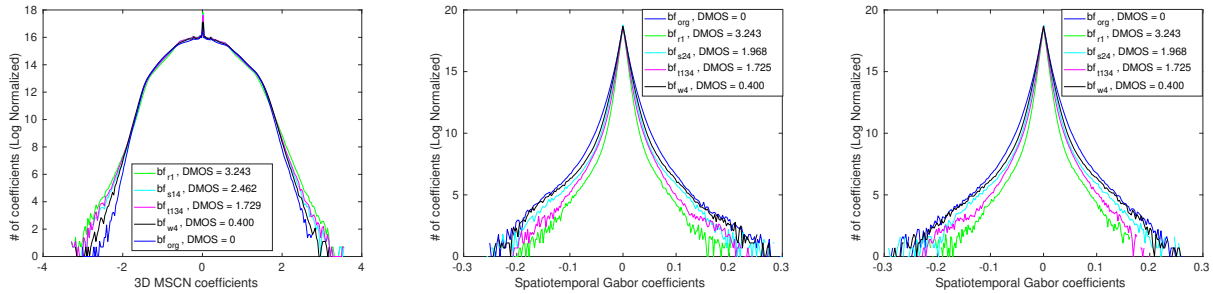


Fig. 3: The spatiotemporal Gabor filters visualization for $v = 1$ (in pixels per frame), $\theta = \pi/3$, moving envelope speed $v_c = v$ and bandwidth $b = 1$ octave. The top row shows the real component or in-phase ($\varphi = 0$) component of $g_{v,\theta,\varphi}(x, y, t)$ and the second row shows the imaginary component ($\varphi = \pi/2$) of $g_{v,\theta,\varphi}(x, y, t)$. The frames of the bandpass filter (19 to 24) are arranged from left to right.



(a) 3D-MSCN distribution (b) Distribution of $r_{2,\pi/3,0}(x,y,t)$ (c) Distribution of $r_{2,\pi/3,\pi/2}(x,y,t)$

Fig. 4: Empirical distribution (normalized histogram) of 3D-MSCN coefficients, symmetry and anti-symmetry parts of sub-band responses with $v = 2$ and $\theta = \pi/2$ of a pristine video and its H.264 compressed versions.



(a) 3D-MSCN distribution (b) Distribution of $r_{2,\pi/3,0}(x,y,t)$ (c) Distribution of $r_{2,\pi/3,\pi/2}(x,y,t)$

Fig. 5: Empirical distribution (normalized histogram) of 3D-MSCN coefficients, symmetry and anti-symmetry parts of sub-band responses with $v = 2$ and $\theta = \pi/2$ of a pristine video and its H.264 compression, rate adaptation, temporal dynamics and wireless packet-loss distorted versions.

but consistent improvement in the performance of the proposed approach.

The feature extraction is done separately on both the symmetric and anti-symmetric components of the bandpass filter responses. In this work, we have considered spatiotemporal Gabor filter with 3 speed levels and 4 orientations resulting in 12 filters with the phase offset $\varphi = 0$, and in 12 filters with the phase offset $\varphi = \pi/2$. Effectively, the total number of spatiotemporal filters is 24.

For a given velocity v and orientation θ, the scene statistics of natural video using spatiotemporal Gabor filter responses with phase offset $\varphi = 0$ is represented by the vector $[\gamma^0_{\theta,v}, \beta^0_{l,\theta,v}, \beta^0_{r,\theta,v}, \eta^0_{\theta,v}]^T$, where $\eta^0_{\theta,v}$ is defined as the ratio of $\gamma^0_{\theta,v}$ and $\beta^0_{l,\theta,v} + \beta^0_{r,\theta,v}$. Similarly, for the anti-symmetric part ($\varphi = \pi/2$), the scene statistics are represented by the vector $[\gamma^{\pi/2}_{\theta,v}, \beta^{\pi/2}_{l,\theta,v}, \beta^{\pi/2}_{r,\theta,v}, \eta^{\pi/2}_{\theta,v}]^T$, where $\eta^{\pi/2}_{\theta,v}$ is defined as the ratio of $\gamma^{\pi/2}_{\theta,v}$ and $\beta^{\pi/2}_{l,\theta,v} + \beta^{\pi/2}_{r,\theta,v}$. Therefore, the overall feature vector of a bandpass filter with the phase offset $\varphi = 0$ and $\varphi = \pi/2$ is given by

$$\hat{f}^{\theta,v}_{st} = [\gamma^0_{\theta,v}, \beta^0_{l,\theta,v}, \beta^0_{r,\theta,v}, \eta^0_{\theta,v}, \gamma^{\pi/2}_{\theta,v}, \beta^{\pi/2}_{l,\theta,v}, \beta^{\pi/2}_{r,\theta,v}, \eta^{\pi/2}_{\theta,v}], \tag{10}$$

Since we have 12 symmetric and 12 anti-symmetric spatiotemporal Gabor filters, each filter contributes 4 feature elements. Effectively, the total feature vector length of spatiotemporal Gabor filters becomes 96. After including the four 3D-MSCN feature elements the overall feature vector length becomes 100. Therefore, the overall feature vector is given by

$$\hat{f}_{overall} = [\hat{f}_{mscn}, \hat{f}^{\theta,v}_{st}]^T \text{over all } \theta, v. \tag{11}$$

We designed the proposed NRVQA technique by mapping video level spatiotemporal features to subjective quality scores (MOS/DMOS) using an SVR. Specifically we used an SVR with radial basis function (RBF) as the kernel, since it has shown better performance than simple linear regression and SVR with other kernels.

We also studied the efficacy of the proposed features by performing an ablation study where the efficacy of the 3D-MSCN features and the Gabor features for NRVQA was evaluated individually. We then combined both the features in designing the proposed NRVQA technique.

To evaluate the performance of the proposed NRVQA approach, we divided a given VQA dataset randomly in the 80:20 ratio for training and testing respectively. This process is repeated 100 times and the median value of the LCC and SRCC is reported in the Tables. Details about the various publicly available VQA datasets and the performance evaluation procedure are given in the following section.

V. Results and Discussion

The performance of the proposed approach is evaluated on both traditional and authentically distorted VQA datasets.

Distortions generated in a controlled lab setting like compression (H.264/H.265) artifacts, transmission distortions (wireless and IP packet loss) etc. are called traditional distortions. The distortions that occur during acquisition are called in-capture or authentic distortions or realistic distortions. The proposed approach has shown competitiveness and consistency in performance compared to existing state-of-the-art NRVQA techniques specifically on authentic/in-capture distortions. The VQA datasets like LIVE SD [54], EPFL PoliMI [55], LIVE Mobile [51], ECVQ and EVVQ [56] fall into the traditional distortions category while the datasets like CVD2014 [57], LIVE Qualcomm [58] and KoNViD-1K [59] fall into the authentic distortions category.

A. Traditional VQA datasets

1) LIVE SD [54]: This dataset contains 10 pristine and 150 distorted videos. The distorted videos are generated from a set of pristine videos by applying standard distortions like H.264 compression, MPEG-2 compression, transmission of H.264 compressed bit streams through error-prone IP networks and through error-prone wireless networks. These distortions are applied at different levels such that for a given pristine video, 15 corresponding distorted videos are generated.

2) LIVE Mobile [51]: This dataset mainly focuses on studying the distortions that occur in heavily trafficked wireless networks. It contains 10 pristine videos and 200 distorted videos with HD (1280×720) resolution at 30fps. The distortions that are studied in this dataset are 4 levels of compression, 4 levels of wireless packet-loss, 4 levels of frame-freezes, 3 levels of rate-adaptation and 5 levels of temporal dynamics per reference video.

3) EPFL PoliMI [55]: This dataset consists of 156 videos in total, where 12 videos are pristine. Out of the 12 pristine videos, 6 videos are of CIF (352 × 288) resolution and other are of 4CIF (704 × 576) resolution. These 12 pristine videos are encoded with H.264/AVC and corrupted by simulating packet loss. In total, the dataset is composed of 78 videos with CIF resolution and the other 78 with 4CIF resolution.

4) ECVQ and EVVQ [56]: Both the datasets consist of 90 videos each with distortions like H.264/AVC and MPEG-4 Part compression. The spatial resolution of videos in ECVQ and EVVQ is CIF (352 × 288) and VGA (640 × 480) respectively.

B. Authentic VQA datasets

1) CVD2014 [57]: The dataset is composed of 234 videos that are captured using 78 different cameras and with different compression techniques. The resolution of the videos ranges from QCIF to full HD.

2) LIVE-Qualcomm [58]: This dataset is composed of 208 videos with six common in-capture distortions like artifacts, color, exposure, focus, sharpness, and stabilization. In the subjective study, each video is rated with 39 unique subjects. The resolution of the videos of this dataset is 1920 × 1080 at 30 fps.

3) KoNViD-1K [59]: This dataset has 1200 natural videos with different perceptual quality levels. Videos are labelled with MOS and corresponding standard deviation. Videos of this dataset are collected from the publicly available large-scale YFCC-100M [60] video database. The resolution of the videos of this dataset is 960 × 540 at 15 fps of 8-sec duration.

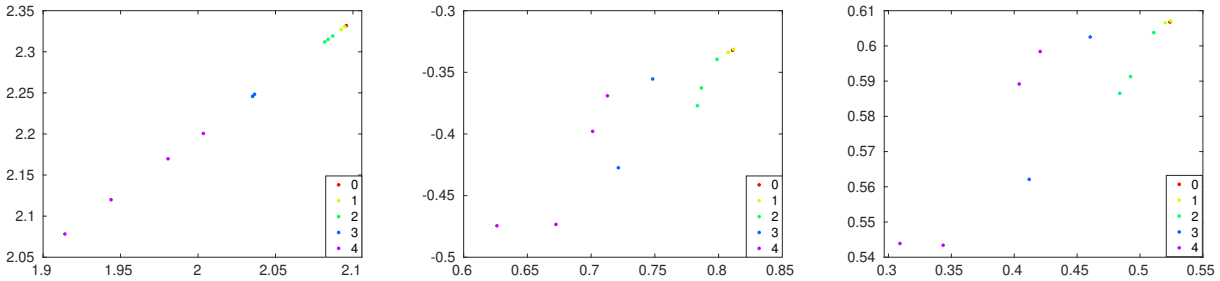
The performance of proposed approach is evaluated on the datasets described above. We quantify the performance of the proposed method using the Linear Correlation Coefficient (LCC) and Spearmen Rank Order Coefficient (SRCC) between subjective scores and objective scores of the proposed NRVQA technique after a non-linear logistic fitting. As per Video Quality Expert Group (VQEG) [1] recommendations for adjusting scaling and non-linearity effect between subjective scores and objective scores we used a non-linear transform $f(x)$ given by

$$f(x) = \frac{\tau_1 - \tau_2}{1 + e^{-(\frac{x-\tau_3}{|\tau_4|})}} + \tau_2. \tag{12}$$

Results are reported by dividing the a given VQA dataset into 80:20 ratio for training and testing sets randomly. This process is repeated 100 times and the median value of the LCC and SRCC is reported in the Tables I, II, III, IV, V and VI.

Tables I and II show distortion-wise and overall performance of the proposed approach on traditionally distorted videos in the LIVE Mobile VQA dataset. We compared the proposed approach with FRIQA methods like SSIM [5] and MS-SSIM [6], NRIQA methods like BRISQUE [11] and NIQE [8], FRVQA methods like MOVIE [14] and VQM [16], and NRVQA methods like V-BLIINDS [29], VIIDEO [44] and TLVQM [34]. It is clear from these tables that the proposed approach delivers competitive performance across distortions and datasets. Similarly, Tables III and IV show distortion-wise and overall performance of the proposed metric on authentically distorted videos in the LIVE Qualcomm VQA dataset [58]. On this dataset we did not compare with FRIQA and FRVQA techniques because of the unavailability of the reference videos. Again, the proposed approach clearly shows competitive performance.

Tables V and VI show the performance of the proposed approach on larger VQA datasets namely LIVE SD [54], EPFL PoliMI [55], LIVE Mobile [51], ECVQ and EVVQ [56], CVD2014 [57], LIVE Qualcomm [58] and KoNViD-1K [59]. Most of the existing techniques do not deliver consistent performance across the spectrum of datasets. They perform extremely well on a few datasets while doing poorly on other datasets. Specifically, existing techniques give good performance on traditional distortions and fail on authentically distorted datasets. However, the proposed approach shows competitive performance on traditional (synthetic) distorted datasets and acceptable performance on authentic distorted datasets. Most of the existing NRVQA techniques have not been evaluated on such a large collection of datasets. Also, in many cases, the source codes are not available publicly. Therefore, we could not compare the performance of the proposed approach with existing NRVQA techniques over all the considered datasets.



(a) Visualization of 3D-MSCN model parameters.

(b) Visualization of spatiotemporal Gabor bandpass filter response model parameters.

(c) Visualization of 3D-MSCN and spatiotemporal Gabor bandpass filter response model parameters.

Fig. 6: Feature visualizations of 3D-MSCN and spatiotemporal based model parameters on videos from EPFL PoliMi [55] VQA dataset using t-SNE [61]. The features are semantically separable according to their quality levels. This suggests that the features extracted from statistical modeling the natural videos are well suited for the VQA task. Each video is visualized as a point and video belonging to the same quality levels have the same color. Best viewed on color display with zoom.

TABLE I: LCC based performance evaluation of proposed approach on LIVE Mobile dataset [51]. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	Compression	Rate Adaptation	Temporal Dynamics	Wireless	ALL
SSIM [5]	*0.4947*	*0.3679*	*0.0773*	*0.5609*	*0.4300*
MS-SSIM [6]	*0.6602*	*0.4821*	*0.1400*	*0.6451*	*0.5678*
MOVIE [14]	*0.7744*	*0.7714*	*0.0658*	*0.8451*	*0.6792*
VQM [16]	*0.6316*	*0.4357*	*0.0515*	*0.6692*	*0.5552*
BRISQUE [11]	*0.6441*	*0.5707*	*0.5796*	*0.6295*	*0.2923*
NIQE [8]	*0.7247*	*0.6771*	*0.6485*	*0.7306*	*0.4748*
V-BLIINDS [29]	-	-	-	-	*0.4373*
VIIDEO [44]	0.3473	0.4067	0.4042	0.3442	0.2451
3D-MSCN	0.7965	**0.9031**	0.6765	0.7742	0.7582
ST-Gabor	0.8859	0.8421	0.7427	**0.8630**	**0.8417**
3D-MSCN + ST-Gabor	**0.8979**	0.8200	**0.7783**	0.7978	0.8405

TABLE II: SRCC based performance evaluation of proposed approach on LIVE Mobile dataset [51]. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	Compression	Rate Adaptation	Temporal Dynamics	Wireless	ALL
SSIM [5]	*0.7092*	*0.6303*	*0.3429*	*0.7246*	*0.6498*
MS-SSIM [6]	*0.8044*	*0.7378*	*0.3974*	*0.8128*	*0.7425*
MOVIE [14]	*0.7738*	*0.7198*	*0.1578*	*0.6508*	*0.6420*
VQM [16]	*0.7717*	*0.6475*	*0.3860*	*0.7758*	*0.6945*
BRISQUE [11]	*0.5941*	*0.4172*	*0.5230*	*0.6141*	*0.2622*
NIQE [8]	*0.6647*	*0.6377*	*0.5688*	*0.6429*	*0.4503*
V-BLIINDS [29]	-	-	-	-	*0.4392*
VIIDEO [44]	0.2997	0.4508	0.1934	0.2923	0.2164
3D-MSCN	0.7262	**0.8286**	0.6000	0.7143	0.7151
ST-Gabor	0.7719	0.8034	0.6251	**0.7623**	**0.8330**
3D-MSCN + ST-Gabor	**0.8103**	0.7717	**0.6731**	0.7022	0.8073

TABLE III: LCC based performance evaluation of proposed approach on LIVE Qualcomm dataset [58]. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	ALL
BRISQUE [11]	*0.6402*	*0.3392*	*0.6042*	*0.4550*	*0.5371*	*0.6940*	*0.5788*
NIQE [8]	*0.6078*	*0.2904*	*0.4625*	*0.5371*	*0.5595*	*0.6015*	***0.6802***
V-BLIINDS [29]	***0.8386***	*0.6645*	*0.6900*	*0.8077*	*0.6845*	*0.7138*	*0.6653*
VIIDEO [44]	*0.2888*	*0.3312*	*0.2073*	*0.2515*	*0.3012*	*0.3697*	*0.0982*
3D-MSCN	0.4608	**0.6819**	0.6378	0.6958	0.5026	0.6064	0.3569
ST-Gabor	0.7677	0.5942	0.7026	0.7840	0.8540	**0.7664**	0.6219
3D-MSCN + ST-Gabor	0.7111	0.5812	**0.7102**	**0.8091**	**0.8733**	0.7588	0.6283

TABLE IV: SRCC based performance evaluation of proposed approach on LIVE Qualcomm dataset [58]. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	Artifacts	Color	Exposure	Focus	Sharpness	Stabilization	ALL
BRISQUE [11]	*0.6071*	*0.3571*	*0.5536*	*0.3929*	*0.4821*	*0.6429*	*0.5585*
NIQE [8]	*0.5000*	*0.3214*	*0.3929*	*0.3393*	*0.5000*	*0.2143*	*0.5451*
V-BLIINDS [29]	***0.7321***	*0.6071*	*0.6429*	***0.8036***	*0.6786*	*0.6607*	***0.6177***
VIIDEO [44]	*-0.1786*	*0.1429*	*-0.0714*	*0*	*-0.1786*	*-0.1071*	*-0.1414*
3D-MSCN	0.4286	**0.6607**	0.5714	0.6429	0.4643	0.5357	0.2612
ST-Gabor	0.6319	0.5759	**0.6452**	0.7143	**0.7674**	0.6786	0.5831
3D-MSCN + ST-Gabor	0.6789	0.5734	0.6433	0.6429	0.7500	**0.7321**	0.5892

TABLE V: LCC based performance evaluation of proposed approach on datasets with traditional and authentic distortions. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	LIVE SD [54]	EPFL PoliMi [55]	LIVE Mobile [51]	ECVQ [56]	EVVQ [56]	CVD2014 [57]	LIVE Qualcomm [58]	KoNViD - 1K [59]
NIQE [8]	0.2668	0.516	0.4748	0.4960	0.5757	*0.61*	*0.6802*	*0.34*
FRIQUEE [12]	-	0.2950	-	*0.280*	*0.296*	*0.83*	0.7379	*0.74*
V-BLIINDS [29]	***0.8810***	*0.7520*	*0.4373*	*0.283*	*0.622*	*0.71*	*0.6653*	0.5650
VIIDEO [44]	*0.651*	*0.184*	0.2451	*0.280*	*0.296*	-	*0.0982*	-0.0150
SACONVA [32]	*0.8714*	-	-	-	-	-	-	-
V-MEON [33]	-	-	-	*0.767*	*0.841*	-	-	-
TLVQM [34]	0.6849	0.8960	**0.8949**	**0.8209**	0.7397	***0.85***	***0.8100***	***0.7800***
3D-MSCN	0.4275	0.9060	0.7582	0.5500	0.8667	0.5260	0.3569	0.3952
ST-Gabor	0.5886	0.9257	0.8417	0.7772	0.8807	0.6064	0.6219	0.6385
3D-MSCN + ST-Gabor	0.5979	**0.9282**	0.8405	0.8024	**0.8867**	0.6525	0.6283	0.6531

TABLE VI: SRCC based performance evaluation of proposed approach on datasets with traditional and authentic distortions. Bold indicates best scores and the numbers in italics indicate that they are taken from the literature.

Algorithm	LIVE SD [54]	EPFL PoliMi [55]	LIVE Mobile [51]	ECVQ [56]	EVVQ [56]	CVD2014 [57]	LIVE Qualcomm [58]	KoNViD - 1K [59]
NIQE [8]	0.225	0.4998	0.4503	0.4469	0.5170	*0.58*	*0.5451*	*0.34*
FRIQUEE [12]	-	0.2836	-	*0.150*	*0.357*	*0.82*	0.6795	*0.74*
V-BLIINDS [29]	*0.7590*	*0.8070*	*0.4392*	*0.343*	*0.684*	*0.70*	*0.6177*	0.5720
VIIDEO [44]	*0.624*	*0.2052*	0.2164	*0.15*	*0.357*	-	*-0.1414*	0.0310
SACONVA [32]	***0.8569***	-	-	-	-	-	-	-
V-MEON [33]	-	-	-	*0.639*	*0.800*	-	-	-
TLVQM [34]	0.5042	**0.8966**	**0.8679**	0.7989	0.7386	***0.83***	***0.7800***	***0.7800***
3D-MSCN	0.3661	0.8537	0.7151	0.4964	0.8627	0.4747	0.2612	0.3655
ST-Gabor	0.5772	0.8821	0.8330	0.7874	0.8772	0.5687	0.5831	0.6251
3D-MSCN + ST-Gabor	0.5875	0.8828	0.8073	**0.8080**	**0.8720**	0.6146	0.5892	0.6417

In Table V and VI we also present the importance of the 3D-MSCN and spatiotemporal Gabor filters based features separately using an ablation test. Ablation test is a popular technique to study the importance of a feature or a set of features on performance in their absence. We observe that the performance of the proposed NRVQA technique is superior using spatiotemporal Gabor filter based features when compared to simple 3D-MSCN based features.

We now discuss the reasons for the effectiveness and consistency of our NRVQA algorithm. Figure 4 shows the empirical distribution of the 3D-MSCN coefficients, along with the symmetric and anti-symmetric parts of sub-band responses with $v = 2$ and $\theta = \pi/2$ of a pristine video and its H.264 compressed versions. Figure 5 shows the empirical distribution of the 3D-MSCN coefficients, along with the symmetric and anti-symmetric parts of sub-band responses with $v = 2$ and $\theta = \pi/2$ of a pristine video and its distorted versions at different perceptual quality levels. From Figures 4 and 5, we can clearly observe that the empirical distributions of distorted video are deviating from pristine video (bf_{org}) distributions in accordance to their perceptual quality levels. These observations provide empirical evidence for the effectiveness of the proposed NSS features. In Figure 6, we show the feature visualization of the model parameters that are estimated using 3D-MSCN coefficients and spatiotemporal Gabor filter responses of the natural videos. Specifically, video *foreman.yuv* from the EPFL PoliMi [55] dataset with 12 different quality levels is used. These videos are labeled with DMOS scores. To visualize the feature using t-SNE [61] plot, we rounded the DMOS scores to its nearest integer. From the figure we can conclude that the proposed features discriminate well according to their quality levels. This implies that the proposed features based on natural video modeling can be used to design robust NRVQA algorithms.

In Table VII, we also studied the cross dataset evaluation of the proposed approach. The proposed NRVQA approach trained on the KoNViD-K [59] dataset and tested on the LIVE Qualcomm [58] dataset. We specifically chose these two datasets because they share common distortions (such as authentic/in-capture distortions). We observed that the proposed approach is not able to deliver competitive performance in cross dataset setting. This could be attributed to the different resolutions of the training and testing datasets. The generalization challenges of NRVQA continues to be an open problem.

TABLE VII: Cross dataset validation of the proposed NRVQA algorithm, trained on KoNViD-K [59] dataset and tested on LIVE Qualcomm [58] dataset.

Algorithm	LCC	SRCC
3D-MSCN	0.3398	0.3233
ST-Gabor	0.3778	0.3493
3D-MSCN + ST-Gabor	0.4388	0.4055

Table VIII shows the effect of some of the design parameters, specifically velocity (v) and orientation (θ) of the Gabor filter on performance (by fixing other parameters). We observed improvement in performance of the proposed approach when velocities $v = \{0, 1\}$ to $v = \{0, 1, 2\}$ and similarly with orientation parameter θ. Due to increased time complexity we did not experiment with other higher velocities. To have a trade-off between time complexity and performance, we conducted our experiments with $v = \{0, 1, 2\}$ and $\theta = \{0, 60, 120, 180\}$. The choice of other parameters like half response bandwidth (b), spatial aspect ratio (γ), time mean time (μ_t) and standard deviation (τ) of time varying Gaussian are based on the previous works of Petkov et al. [62]–[65]. These studies have shown that the selected parameters model well the human visual system and also consider some of the restrictions found in the experimental data.

It is worth reiterating the influence of the parameter σ of the Gabor filter on the time complexity of the proposed approach. The higher the value of σ, the higher the filter size. This increases the time complexity of the 3D convolution operation which is crucial operation in proposed approach. We also studied the computational complexity aspect of the proposed approach. It is always a difficult task to compare the computational complexity of different NRVQA algorithms quantitatively because most of the existing approaches are not optimized for computational efficiency. Rather, the main focus is on designing effective objective video quality assessment techniques. However, we compared the computational complexity of different NRVQA technique using publicly available *Matlab* implementations. The time complexity comparison is made by running all these algorithms (and our proposed method) on the same computer. Specifically, we ran time complexity check on videos of resolution QCIF and CIF, which are borrowed from EPFL PoliMi [55] and running time on per-frame basis is reported in Table IX. The hardware and software specification of the computer is Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz, 24GB RAM, NVIDIA-TITAN Xp GPU, and MATLAB R2018a.

VI. Conclusions

In this paper, we systematically studied the spatiotemporal statistics of natural videos in the spatial domain using 3D-MSCN coefficients and in the frequency domain using spatiotemporal Gabor filters. We proposed an AGGD to model the statistics of both 3D-MSCN and bandpass filter coefficients of natural videos. We then proposed an NRVQA technique based on the parameters of the AGGD model. It was demonstrated that the proposed algorithm delivers competitive performance on traditional (synthetic) distortions and acceptable performance on authentic distortions. To the best of our knowledge, this is the first work to study the video level scene statistics using spatiotemporal Gabor filters. The proposed scene statistics has other potential applications like action recognition, object tracking etc. As future work, we plan to use spatiotemporal scene statistics to design a completely blind VQA technique and other video related tasks.

VII. Acknowledgement

This work is supported under Visvesvaraya Ph.D. scheme by the Media Asia Lab, Ministry of Electronics and Information Technology, Government of India. We also gratefully acknowledge the support of NVIDIA Corporation for the donation of a Titan Xp GPU used for this research. The authors would like thank the anonymous reviewers for their insightful comments and suggestions that helped improve the quality of this paper. We also thank Qualcomm Technologies for their generous financial support through Qualcomm Innovation Fellowship India 2019.

References

[1] "(2000) final report from the video quality experts group on the validation of objective quality metrics for video quality assessment, http://ftp://vqeg.its.bldrdoc.gov/documents/vqeg˙approved˙final˙reports/."
[2] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.
[3] R. Sekuler, S. N. Watamaniuk, and R. Blake, "Motion perception," 2002.
[4] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Josa a*, vol. 2, no. 2, pp. 284–299, 1985.
[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, Ieee, 2003.
[7] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
[8] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
[9] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.
[10] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Processing: Image Communication*, vol. 29, no. 7, pp. 725–747, 2014.
[11] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
[12] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," in *SPIE/IS&T Electronic Imaging*, pp. 93940J–93940J, International Society for Optics and Photonics, 2015.
[13] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995–1002, 2013.
[14] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.
[15] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016.
[16] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 637–649, 2014.
[17] D. J. Heeger, "Optical flow using spatiotemporal filters," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 279–302, 1988.

TABLE VIII: Study of the proposed approach on ECVQ [56] and KoNViD [59] with different velocities v and orientations θ. Bold indicates best scores.

Data set		$v = \{0, 1\}$				$v = \{0, 1, 2\}$			
		$\theta = \{0, 60, 120, 180\}$		$\theta = \{0, 45, 90, 135, 180\}$		$\theta = \{0, 60, 120, 180\}$		$\theta = \{0, 45, 90, 135, 180\}$	
		LCC	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**
ECVQ [56]	**ST-Gabor**	0.7635	0.7689	0.7703	0.7841	0.7772	0.7874	0.7759	0.7869
	3D-MSCN + ST-Gabor	0.7866	0.7871	0.7931	0.8013	0.8024	0.8080	0.8031	0.8073
KoNViD [59]	**ST-Gabor**	0.6259	0.6162	0.6344	0.6173	0.6361	0.6244	0.6359	0.6265
	3D-MSCN + ST-Gabor	0.6432	0.6358	0.6489	0.6379	0.6519	0.6412	0.6539	0.6435

TABLE IX: Time complexity comparison of the proposed approach with existing NRVQA techniques on a per-frame basis with spatial resolution QCIF and CIF.

Algorithm	QCIF	CIF
NIQE [8]	0.3862 sec	0.4273 sec
FRIQUEE [12]	6.6260 sec	27.267 sec
V-BLIINDS [29]	0.0622 sec	0.1979 sec
VIIDEO [44]	0.1605 sec	0.6683 sec
TLVQM [34]	0.0863 sec	0.3152 sec
3D-MSCN	0.0201 sec	0.0805 sec
ST-Gabor	0.6913 sec	1.6779 sec
3D-MSCN + ST-Gabor	0.7114 sec	1.7584 sec

[18] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *4th International Conference on Computer Vision*, pp. 231–236, IEEE, 1993.

[19] G. Farneback, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, pp. 171–177, IEEE, 2001.

[20] B. Ortiz-Jaramillo, A. Kumcu, L. Platisa, and W. Philips, "A full reference video quality measure based on motion differences and saliency maps evaluation," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, pp. 714–722, IEEE, 2014.

[21] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 219–234, 2018.

[22] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Citeseer, 2005.

[23] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 25–26, 2007.

[24] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 71–83, 2008.

[25] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 684–694, April 2013.

[26] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1010–1013, March 2010.

[27] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[28] T. Brandão and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437–1447, 2010.

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

[30] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3329–3342, 2016.

[31] K. Manasa and S. S. Channappayya, "An optical flow-based no-reference video quality assessment algorithm," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2400–2404, IEEE, 2016.

[32] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, 2016.

[33] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 546–554, ACM, 2018.

[34] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, pp. 5923–5938, Dec 2019.

[35] P. M. Shabeer, S. Bhati, and S. S. Channappayya, "Modeling sparse spatio-temporal representations for no-reference video quality assessment," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1220–1224, Nov 2017.

[36] M. Aharon, M. Elad, A. Bruckstein, *et al.*, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, p. 4311, 2006.

[37] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *2014 IEEE international conference on image processing (ICIP)*, pp. 491–495, IEEE, 2014.

[38] M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, "Deep learning for quality assessment in live video streaming," *IEEE signal processing letters*, vol. 24, no. 6, pp. 736–740, 2017.

[39] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[40] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2349–2353, IEEE, 2019.

[41] J. E. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *Visual Communications and Image Processing*, vol. 5150, pp. 621–633, International Society for Optics and Photonics, 2003.

[42] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *IEEE International Conference on Image Processing 2005*, vol. 3, pp. III–141–4, Sept 2005.

[43] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 685–688, 2005.

[44] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[45] D. W. Dong and J. J. Atick, "Statistics of natural time-varying images," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 345–358, 1995.

[46] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 265, no. 1412, pp. 2315–2320, 1998.

[47] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, vol. 1, pp. I–41, IEEE, 2003.

[48] Z. Wang and Q. Li, "Statistics of natural image sequences: Temporal motion smoothness by local phase correlations," in *Human Vision and Electronic Imaging XIV*, vol. 7240, p. 72400W, International Society for Optics and Photonics, 2009.

[49] G. Varghese and Z. Wang, "Video denoising based on a spatiotemporal gaussian scale mixture model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 7, pp. 1032–1040, 2010.

[50] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Asilomar Conference Signals, Systems and Computers*, pp. 723–727, IEEE, 2011.

[51] 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 Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[52] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biological Cybernetics*, vol. 97, no. 5-6, pp. 423–439, 2007.

[53] N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *International Conference on Image Processing*, pp. 2281–2284, IEEE, 2009.

[54] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, pp. 1427–1441, June 2010.

[55] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 2430–2433, IEEE, 2010.

[56] M. Vranješ, S. Rimac-Drlje, and D. Vranješ, "Ecvq and evvq video quality databases," in *Proceedings ELMAR-2012*, pp. 1–5, IEEE, 2012.

[57] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "Cvd2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[58] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[59] 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 *International Conference on Quality of Multimedia Experience*, pp. 1–6, IEEE, 2017.

[60] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: the new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[61] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[62] N. Petkov and P. Kruizinga, "Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells," *Biological cybernetics*, vol. 76, no. 2, pp. 83–96, 1997.

[63] P. Kruizinga and N. Petkov, "Nonlinear operator for oriented texture," *IEEE Transactions on image processing*, vol. 8, no. 10, pp. 1395–1407, 1999.

[64] N. Petkov and M. A. Westenberg, "Suppression of contour perception by band-limited noise and its relation to nonclassical receptive field inhibition," *Biological cybernetics*, vol. 88, no. 3, pp. 236–246, 2003.

[65] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE Transactions on image processing*, vol. 12, no. 7, pp. 729–739, 2003.