No-reference Video Quality Assessment Using Space-time Chips

Joshua P. Ebenezer[1], Zaixi Shang[1], Yongjun Wu[2], Hai Wei[2], Alan C. Bovik[1]

1 Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin

2 Amazon Prime Video

Presented at IEEE MMSP 2020

Presented by Joshua P. Ebenezer
PhD. Student, LIVE, The University of Texas at Austin
joshuaebenezer@utexas.edu

Introduction

- No-Reference Video Quality Assessment is the task of predicting the quality of a video without the use of a reference video.
- The metric produced must correlate well with human judgments of video quality.

Space-Time (ST) Chips

- ST-chips are localized cuts of video volumes.
- ST-chips are cut from spatial windows such that they are perpendicular to the motion vector at each of those windows.
- The windows are 5x5 and we cut 5 frames back in time.

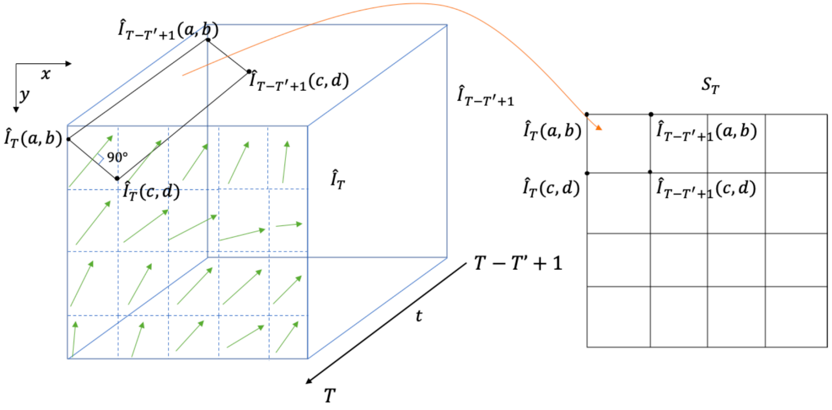


Fig. 1. Extracting ST-chips. On the left is a spatiotemporal volume of frames from time $T - T' + 1$ to T. The green arrows represent motion vectors at each spatial patch at time T. ST-chips are cut perpendicular to these across time and aggregated across spatial patches to form the frame on the right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

ST-Chips

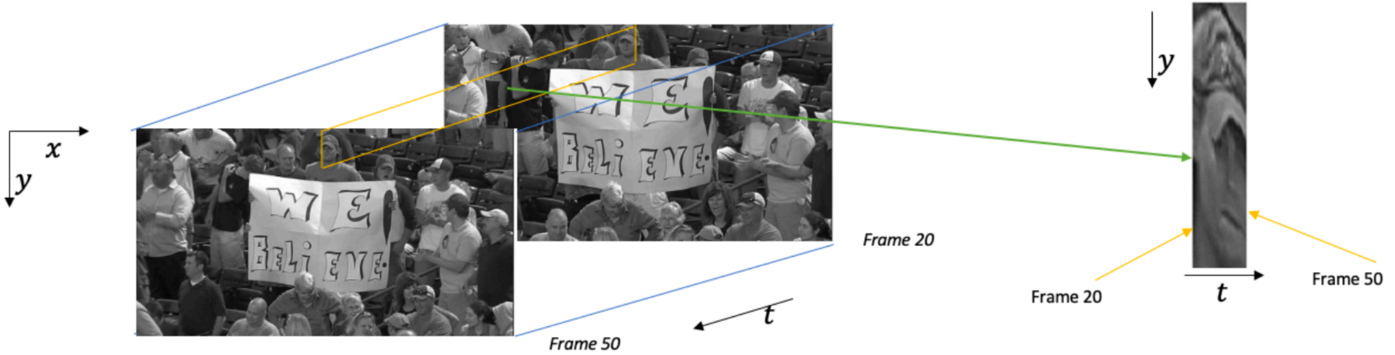


Fig. 2. ST-Chips capture views of objects in motion. In this video, the person in the center moves to their right over time. Consequently, a chip taken in the proximity of their face over this duration and perpendicular to their motion captures their face itself.

Statistics of ST-chips

- It is known that MSCNs $\hat{I}$ of natural images I follow regular statistics.
- MSCNs are defined as $$\hat{I}_T(i,j) = \frac{I_T(i,j) - \mu_T(i,j)}{\sigma_T(i,j) + C}$$

where μ_T is the local mean and σ_T^2 is the local variance.

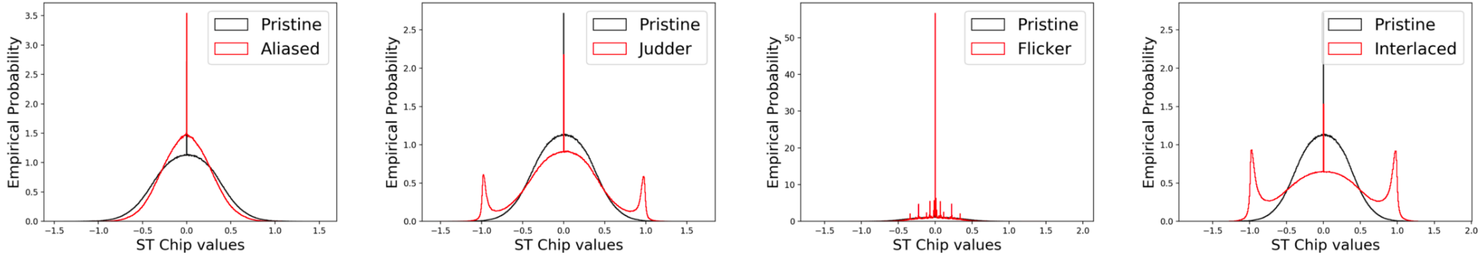
Statistics of ST-chips – first order

- Since ST-chips capture natural objects as they move, we expect ST-chips of MSCNs to follow similar statistics as MSCNs of natural images.
- We find (as expected) that ST-chips of MSCNs follow a generalized Gaussian distribution in the first order:

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

- We extract α and β from the distribution as features for quality assessment.

Statistics of ST-chips – first order



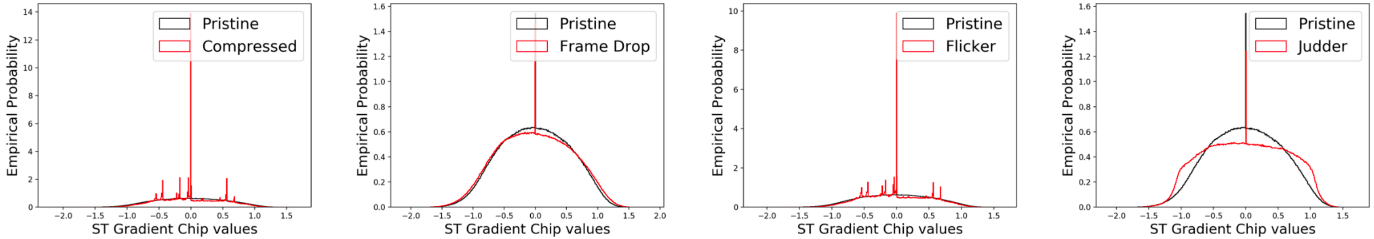
(a) Aliased and pristine (b) Judder and pristine (c) Flicker and pristine (d) Interlaced and pristine

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

ST Gradient chips

- Gradients can capture distortions that affect edges and contrast, which are very important.
- We also find the gradient magnitude of the video and compute the MSCNs of the gradient magnitude as well.
- We then find the ST-chips of the MSCNs of the gradient magnitude.

Statistics of ST Gradient chips – first order



(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.

- We find that ST Gradient chips of MSCNs follow a generalized Gaussian distribution in the first order.
- We extract α and β from the distribution as features for quality assessment.

Statistics of ST-chips – second order

- We find the pairwise products of ST-chips

$$H_T(i,j) = S_T(i,j)S_T(i,j+1)$$

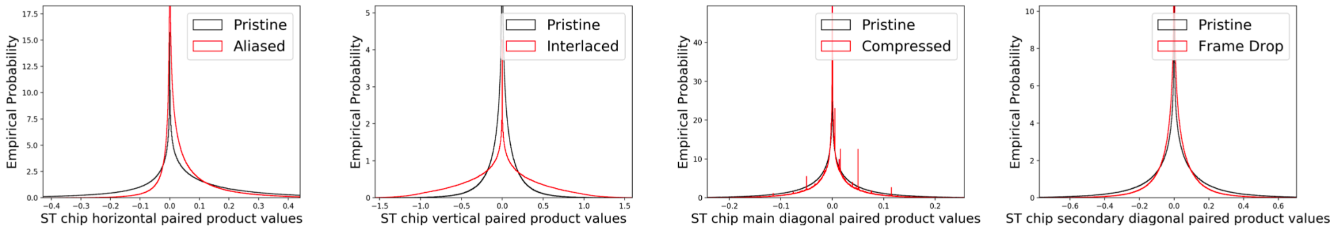
$$V_T(i,j) = S_T(i,j)S_T(i+1,j)$$

$$D1_T(i,j) = S_T(i,j)S_T(i+1,j+1)$$

$$D2_T(i,j) = S_T(i,j)S_T(i+1,j-1)$$

- We find that these follow an asymmetric generalized Gaussian distribution.

Statistics of ST-chips – second order



(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.

- We find that both ST-Chips and ST-Gradient chips follow an AGGD in the second order.

Statistics of ST-chips – second order

- AGGDs are of the form

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

- We extract the parameters $\eta, \nu, \sigma_l^2, \sigma_r^2$ from the distribution.
- $\eta = (\beta_r - \beta_l)\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}$. σ_l^2 and σ_r^2 are the variances of each side of the distribution.

Features in ChipQA-0

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.	$f_{73} - f_{109}$

- We train an SVR with these features.
- For all databases, we perform an 80:20 train-test split and use cross-validation to find the best parameters for the SVR.

Experiments

- We evaluate our algorithm on 4 large databases:
  - LIVE-APV Livestream VQA – *Mix of spatial and temporal distortions. 4K content shown on 4K TV.*
  - LIVE Mobile - *Mix of spatial and temporal distortions. Study was on mobile devices.*
  - Konvid 1k – *User generated content. Mostly spatial distortions.*
  - LIVE Video Quality Challenge (VQC) – *User generated content. Mostly spatial distortions.*

Results

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**

Results

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

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

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

Computation time

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

Conclusion

- We presented a novel, quality-aware feature space.
- 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.
- ChipQA-0 is highly competitive with other state-of-the-art models on several databases.
- We are working on developing this idea further to do away with optical flow and improve performance.