Power of Tempospatially Unified Spectral Density for Perceptual Video Quality Assessment

This paper is recognized as a **Finalist of the World's FIRST 10K Best Paper Award**

Mohammed Aabed, Gukyeong Kwon*, and Ghassan AlRegib

(*: Speaker)

Multimedia and Sensors Lab (MSL)
Center for Signal and Information Processing (CSIP)
School of Electrical and Computer Engineering
Georgia Institute of Technology

Outline

I. Introduction

II. Literature Surveys

III. Proposed Method

   a) Feature Extraction Using 3D Power Spectral Density

   b) Mapping Differences in Features to Human Perception

   c) Feature Evaluation

IV. Experiments and Results

V. Conclusion

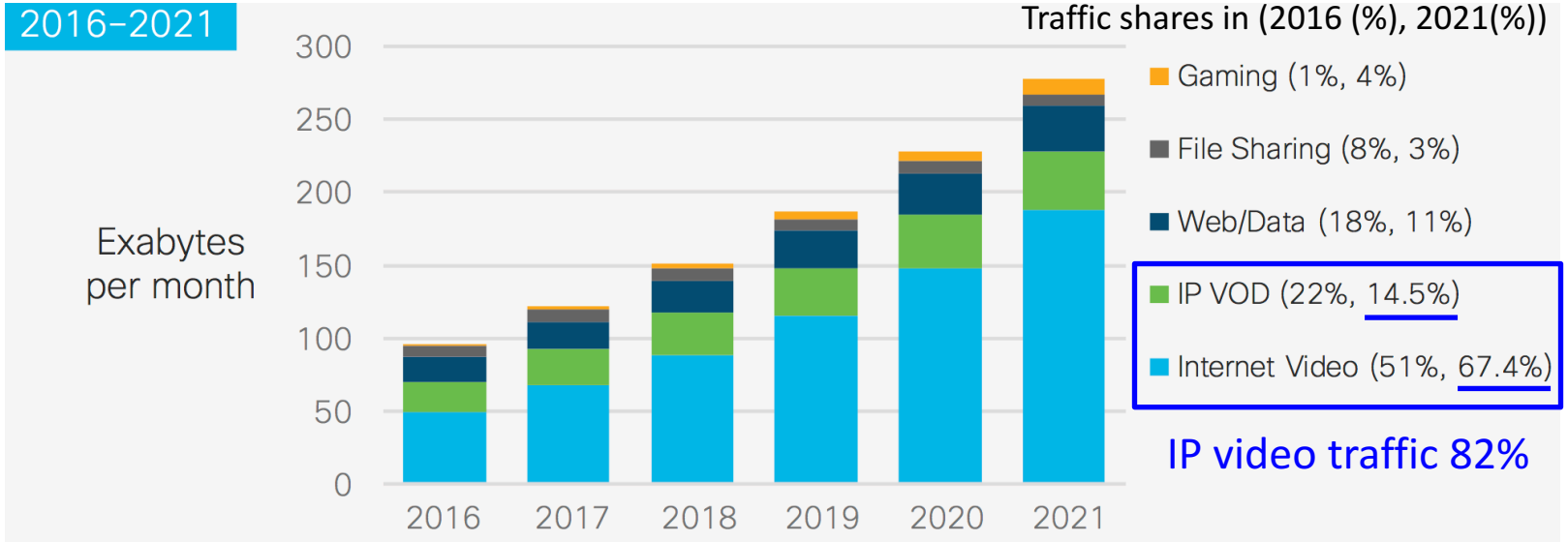
Outline

Drastic Growth of Global IP Video Traffic

- 2017 Cisco Complete Visual Networking Index (VNI) Forecast [1]

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2016	26,600 GB per second
2021	105,800 GB per second

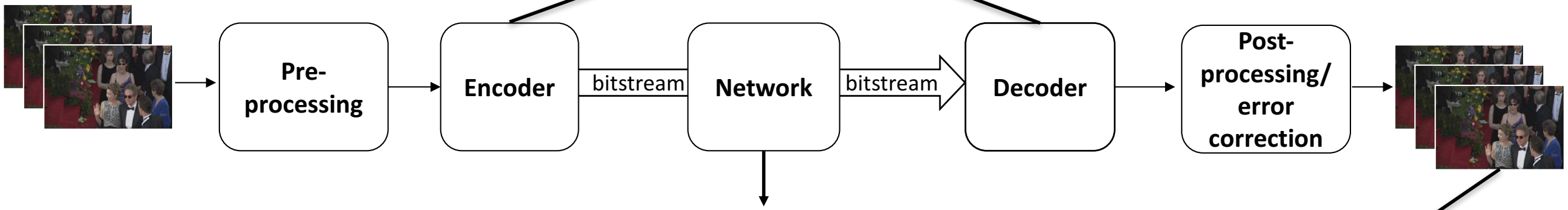
Over 3×



It would take an individual more than 5,000,000 years to watch **the amount of video** that will cross global IP networks **each month in 2021**

[1] "The Zettabyte Era: Trends and Analysis," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf, June 2016

Video Streaming



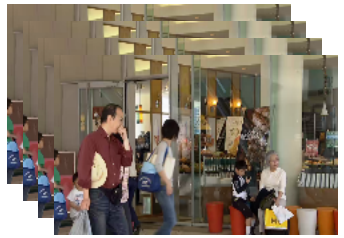
- Channel-induced distortion
- Hardware, bandwidth and other constraints

- Quality?
- Distortion?

A massive number of videos + Distortions from video streaming ⇨ Objective video quality assessment

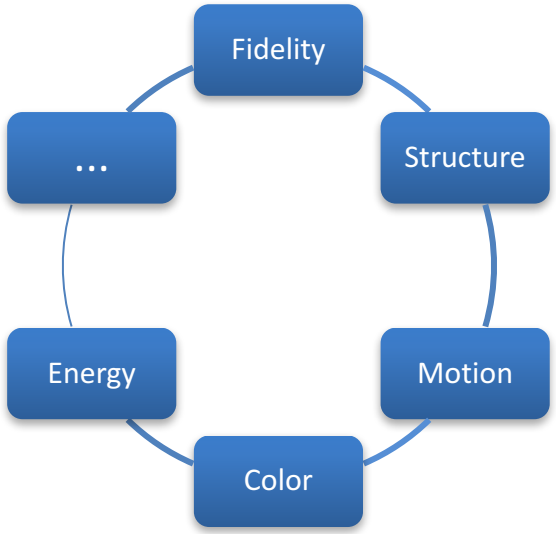
Video Quality Assessment (VQA)

- **Full-reference (w/ anchor)**
- Reduced-reference
- No-reference (w/o anchor)



Video

Spatial/Temporal/Tempospatial features



Objective video quality score

Correlated ?

Perception

Subjective video quality score

How to estimate perceptual video quality score highly correlated with subjective video quality scores?

Outline

I. Introduction

II. Literature Surveys

III. Proposed Method

a) Feature Extraction Using 3D Power Spectral Density

b) Mapping Differences in Features to Human Perception

c) Feature Evaluation

IV. Experiments and Results

V. Conclusion

Video Quality Assessment in the Literature

Multimedia & Sensors LAB

Features \ Algorithm	ST-MAD	ViS3	VRF	Proposed	STAQ	ST-RRED	PeQASO	Video BLIINDS	VIIDEO	Zhu	Yang	Video CORNIA	Dimitrievski	SACONVIA
	Full-reference				Reduced-reference			No-reference						
DCT								■		■			■	
Wavelet					■	■								■
Residuals						■		■	■				■	
Motion	■	■		■	■	■	■	■	■					
Optical Flow	■	■					■							
Codebook												■		
Deep Network														■
Bitstream											■			
Structure				■	■				■			■	■	
3D processing			■	■										■
Power Spectral Density				■										

Outline

I. Introduction

II. Literature Surveys

III. Proposed Method

a) Feature Extraction Using 3D Power Spectral Density

b) Mapping Differences in Features to Human Perception

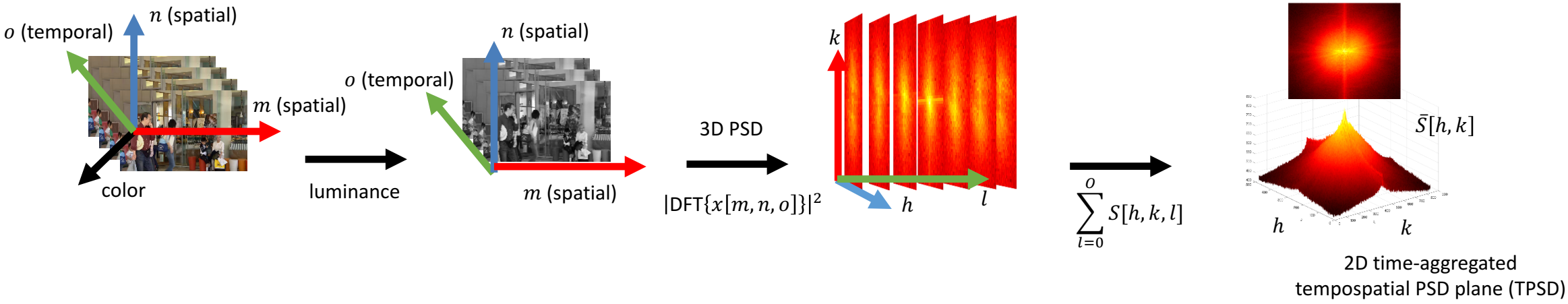
c) Feature Evaluation

IV. Experiments and Results

V. Conclusion

Pipeline

- 3D Power Spectral Density (PSD) Analysis
  - A video is divided into equal size tensor ($M(width)\times N(height)\times O(30\ frames)$)
  - T : Number of tensors in a video
  - For given tensor t $(t = 1,2, \dots, T)$



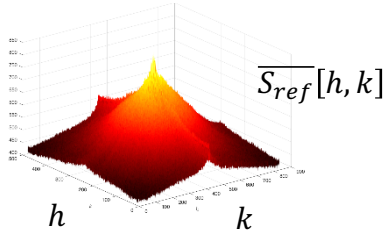
The power spectrum is affected by different types and levels of distortion in a different but **regular** way

Mapping to Human Perception

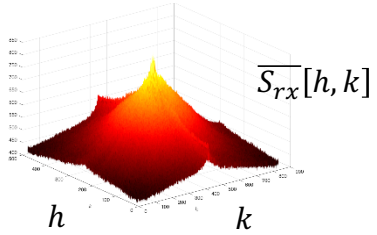
- Local Cross-Correlation
  - Local cross-correlation is calculated in a 11×11 window centered at each pixel
  - Quantifying the masking effect of the original contents in the presence of distortion

    e.g. High correlation → The human visual system (HVS) is not affected by the distortion

Original (Error free)



Distorted



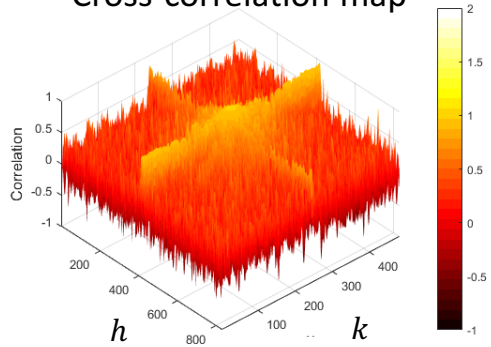
Local cross-correlation

$$\zeta_t[h,k] = \frac{\sigma_{\bar{S}_{ref,t}\cdot\bar{S}_{rx,t}}}{\sigma_{\bar{S}_{ref,t}} \times \sigma_{\bar{S}_{rx,t}}}$$

$\sigma_{\bar{S}_{ref,t}\cdot\bar{S}_{rx,t}}$: Localized covariance

$\sigma_{\bar{S}_{ref,t}}$, $\sigma_{\bar{S}_{rx,t}}$: Localized standard deviation

Cross-correlation map



Mean pooling → $\bar{\zeta}_t$

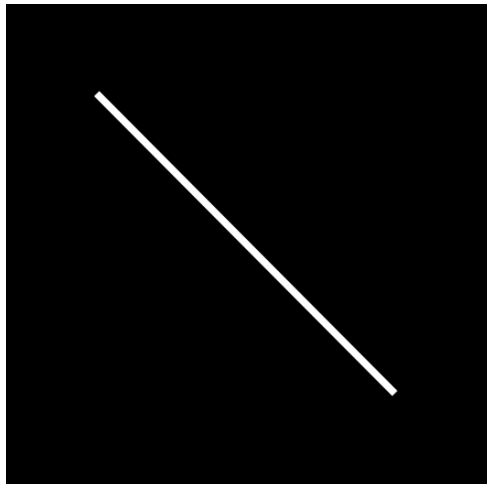
Overall video quality score $\bar{\zeta} = \frac{1}{T}\sum_{\forall t} \bar{\zeta}_t$

Which Video is Better?

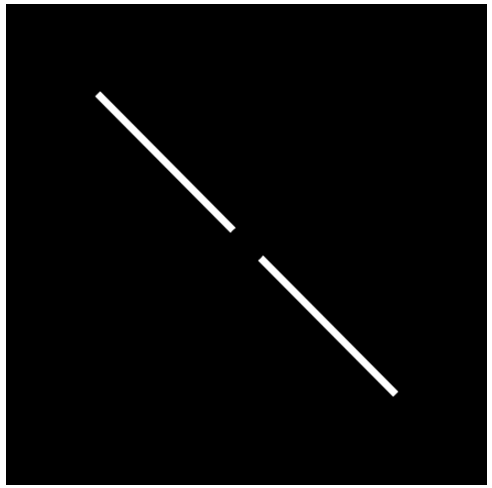
Multimedia & Sensors LAB

No motion

Anchor
(Error free)

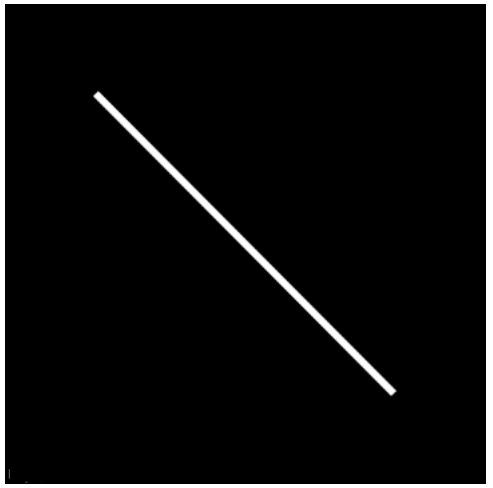


Distorted

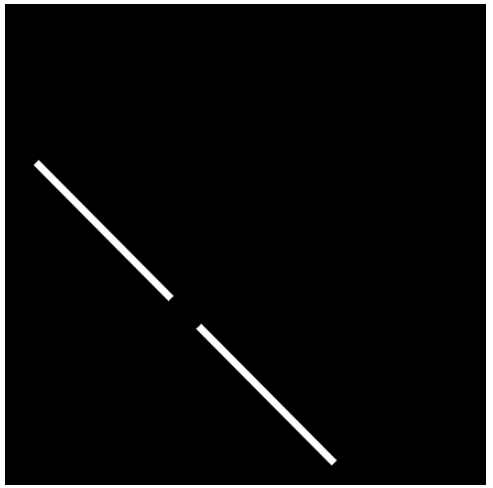


Simple motion

Anchor
(Error free)

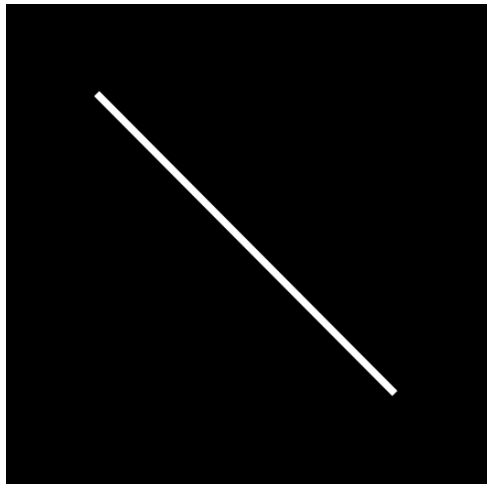


Distorted

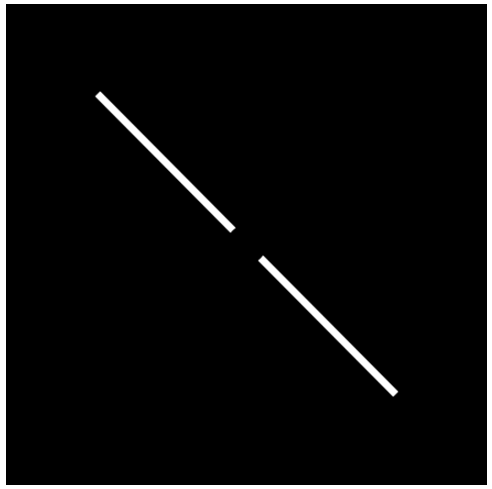


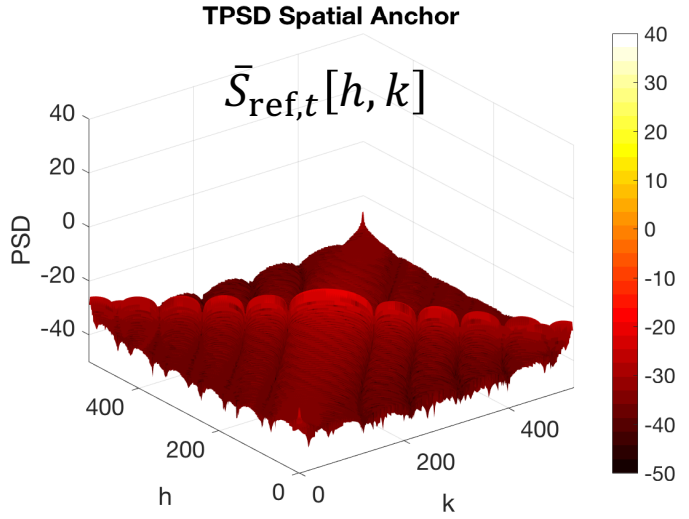
Simple Example – No Motion

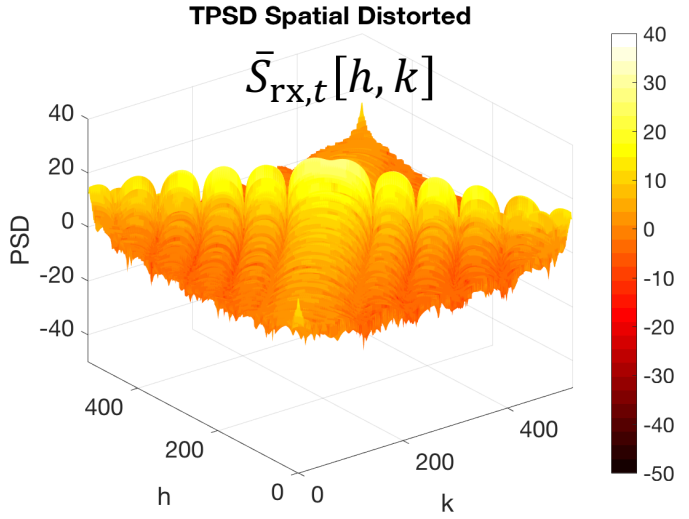
Anchor
(Error free)

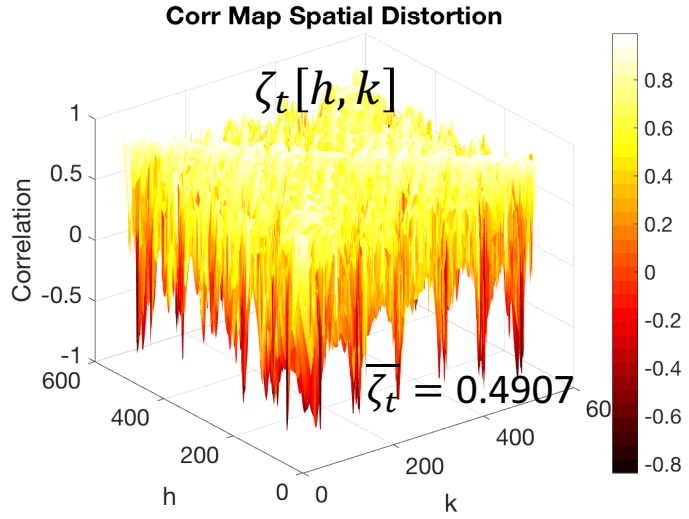


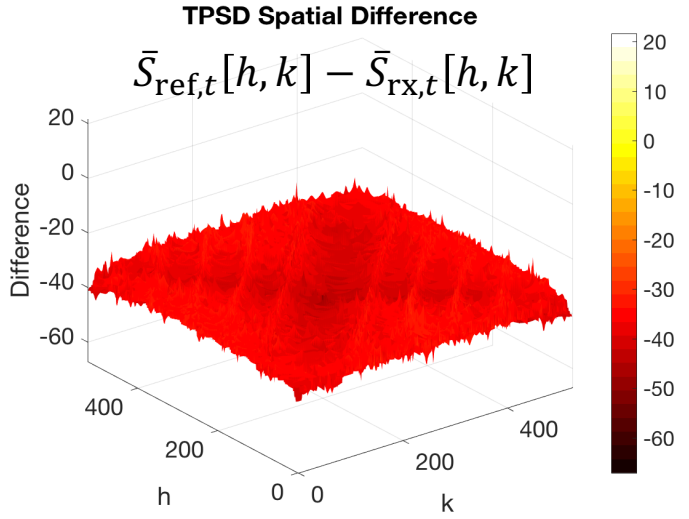
Distorted





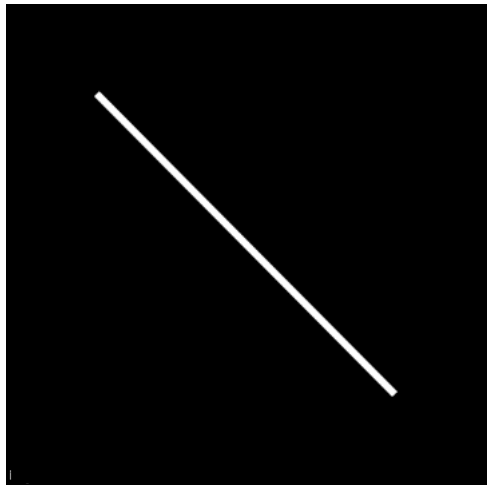




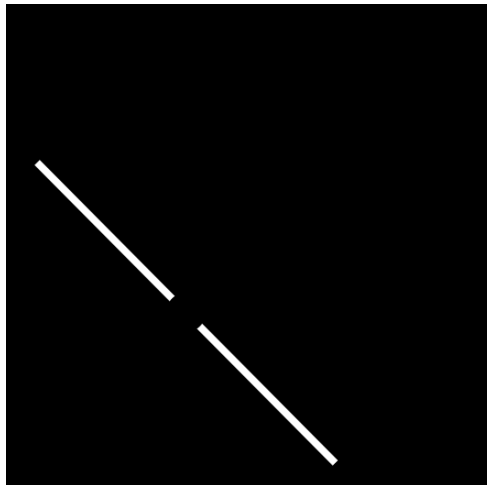


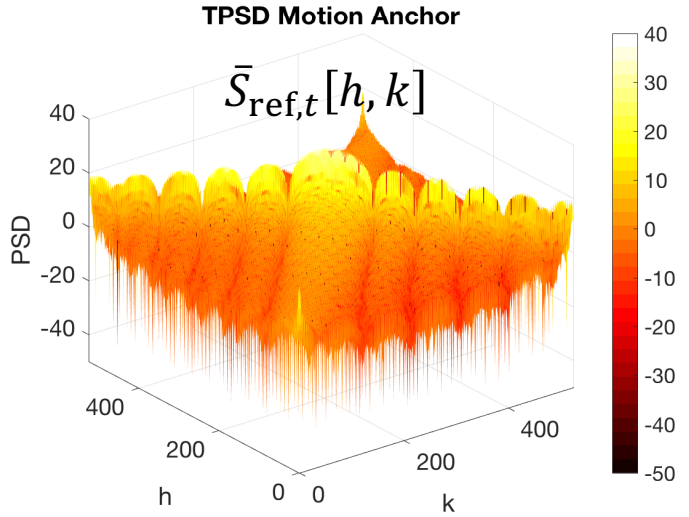
Simple Example – Simple Motion

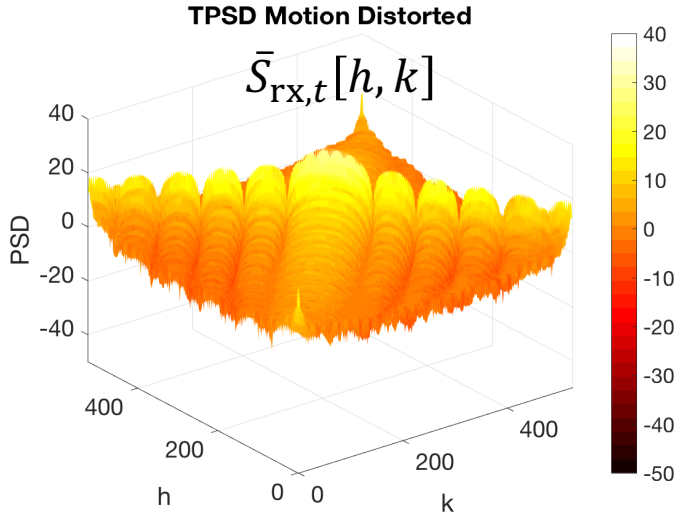
Anchor
(Error free)

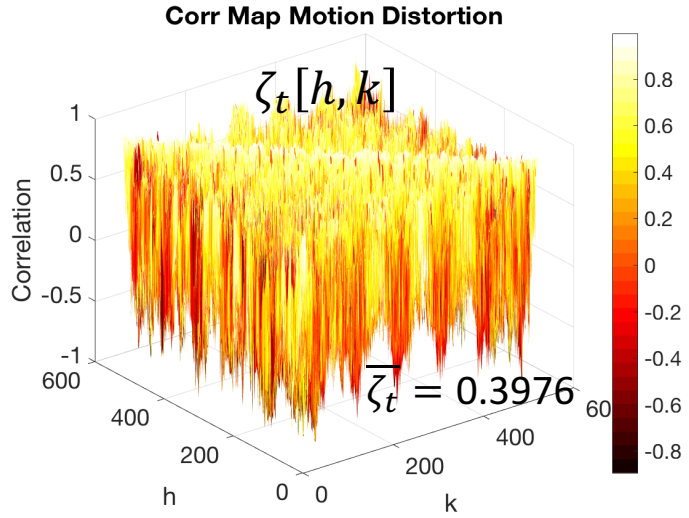


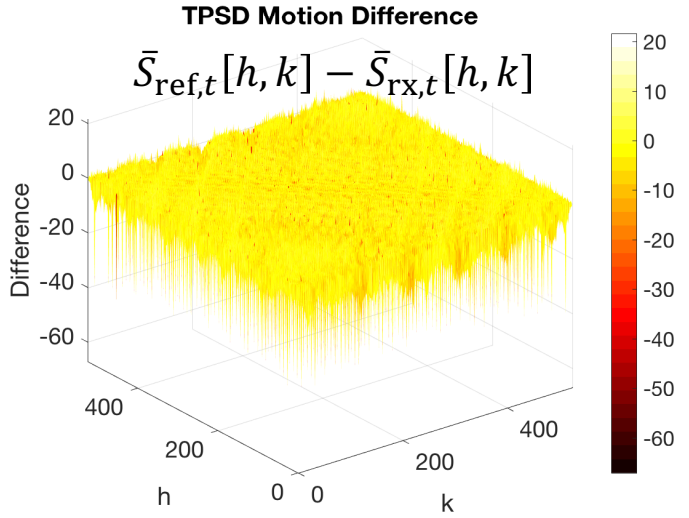
Distorted





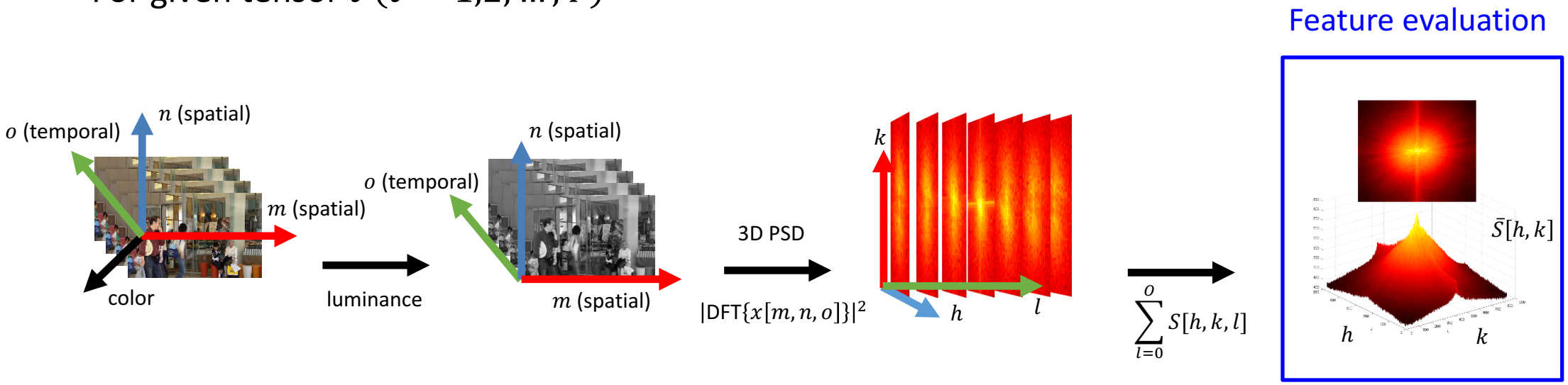






Feature Evaluation

- 3D Power Spectral Density Analysis
  - A video is divided into equal size tensor ($M \times N \times O$)
  - T : Number of tensors in a video
  - For given tensor t ($t = 1,2, \ldots, T$)



$$\sum_{l=0}^{O} S[h,k,l]$$

Incremental Change of Distortion Level

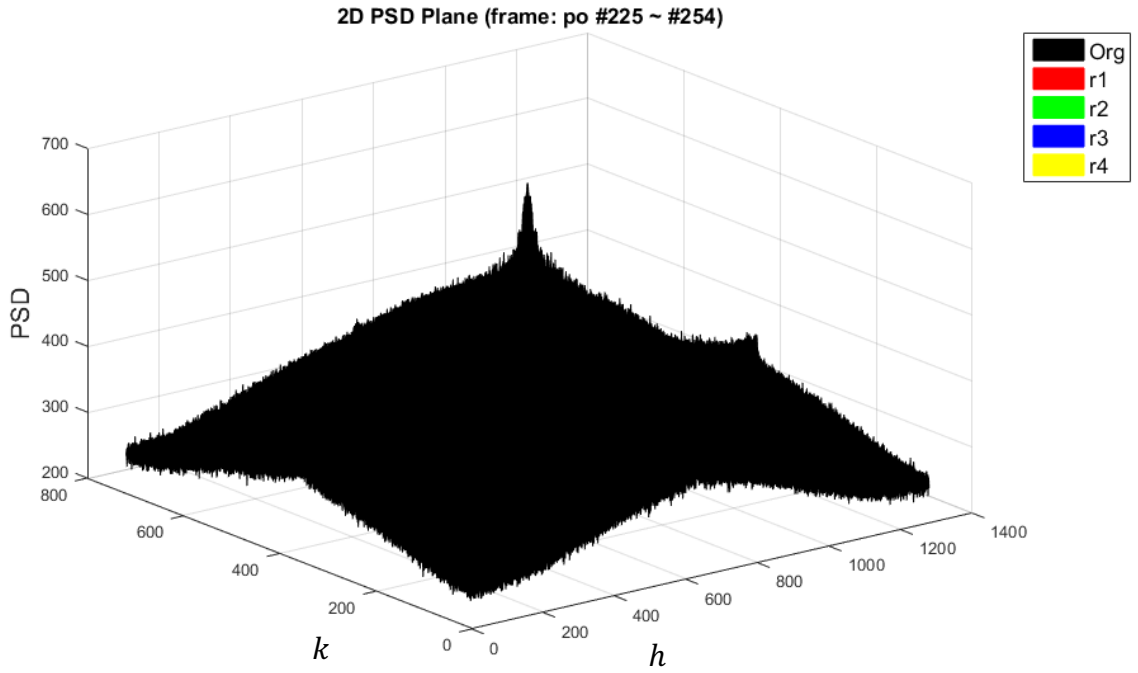
- Example
  - LIVE Mobile Video Quality Assessment Database
  - Sequence: Panning Under Oak (PO) (Frame #225 ~ #254 (30 frames))
  - Cube size: 1280(width) × 720(height) × 30(Frames)

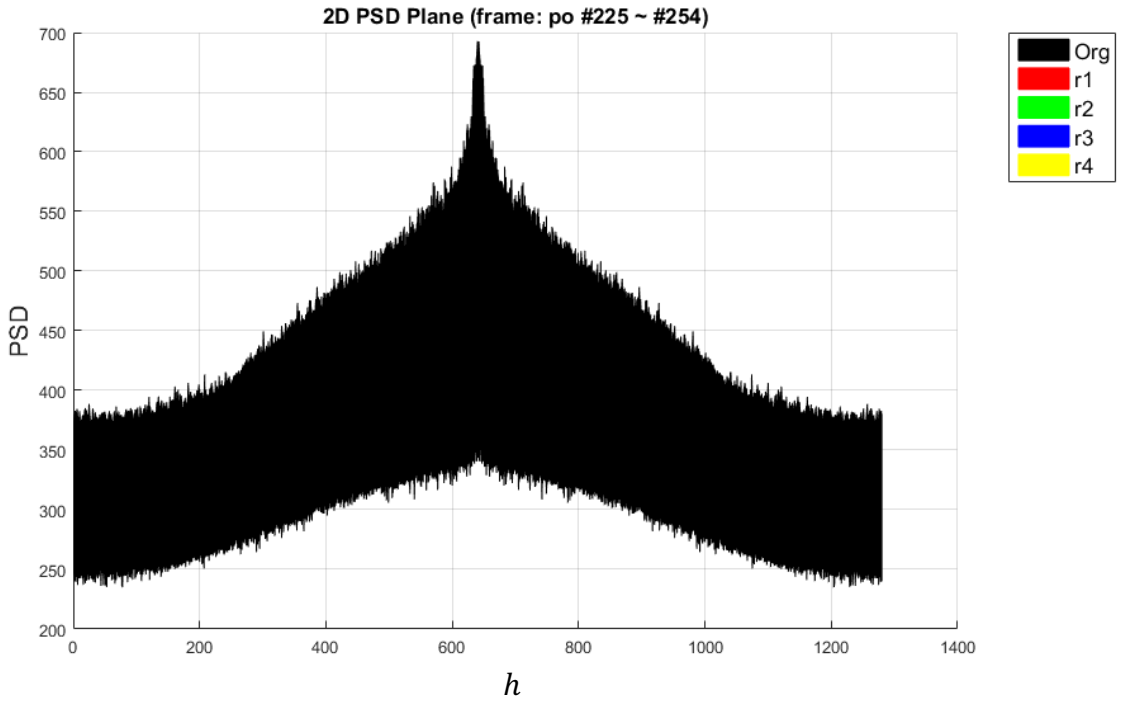
<Panning Under Oak frame #225>

<Panning Under Oak frame #254>

Incremental Change of Distortion Level

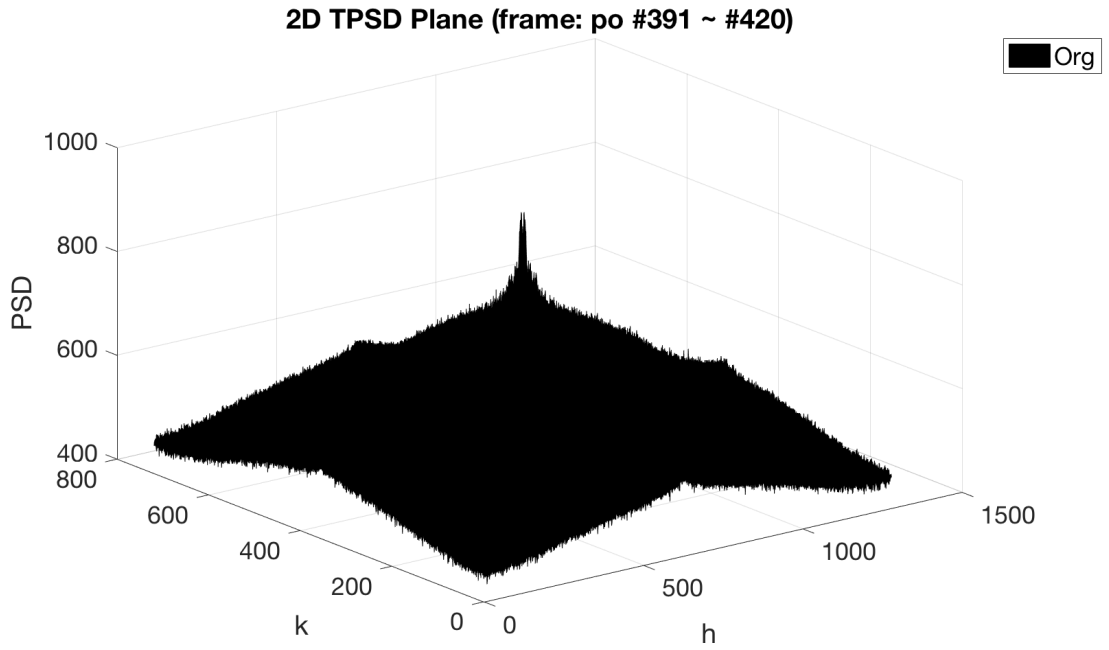
- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4

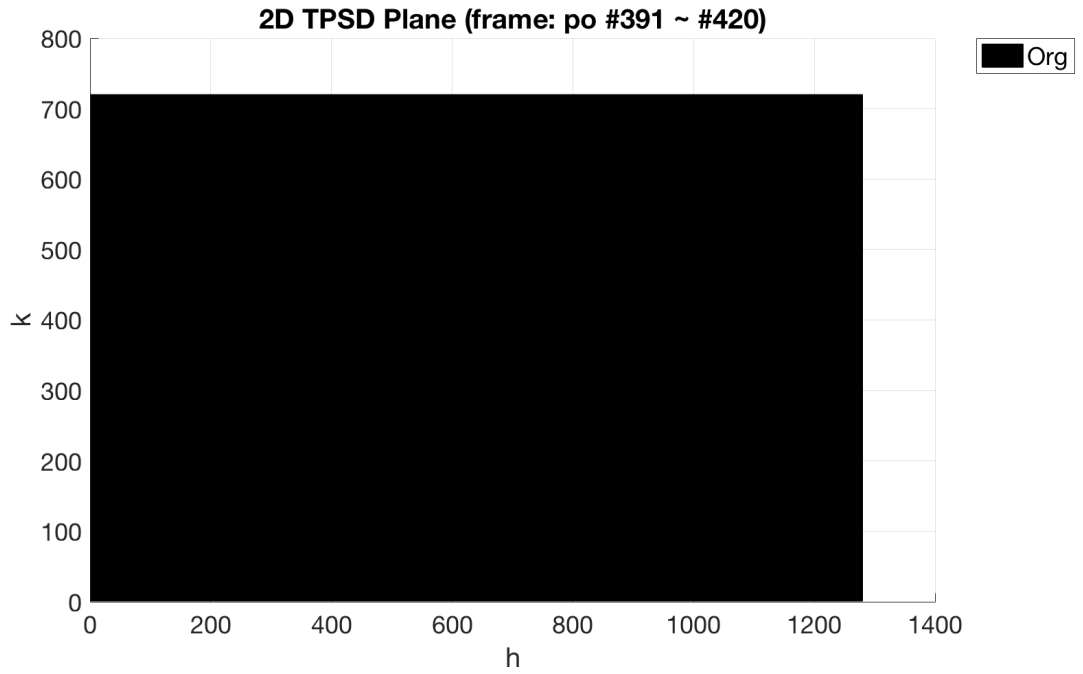




Incremental Change of Distortion Level

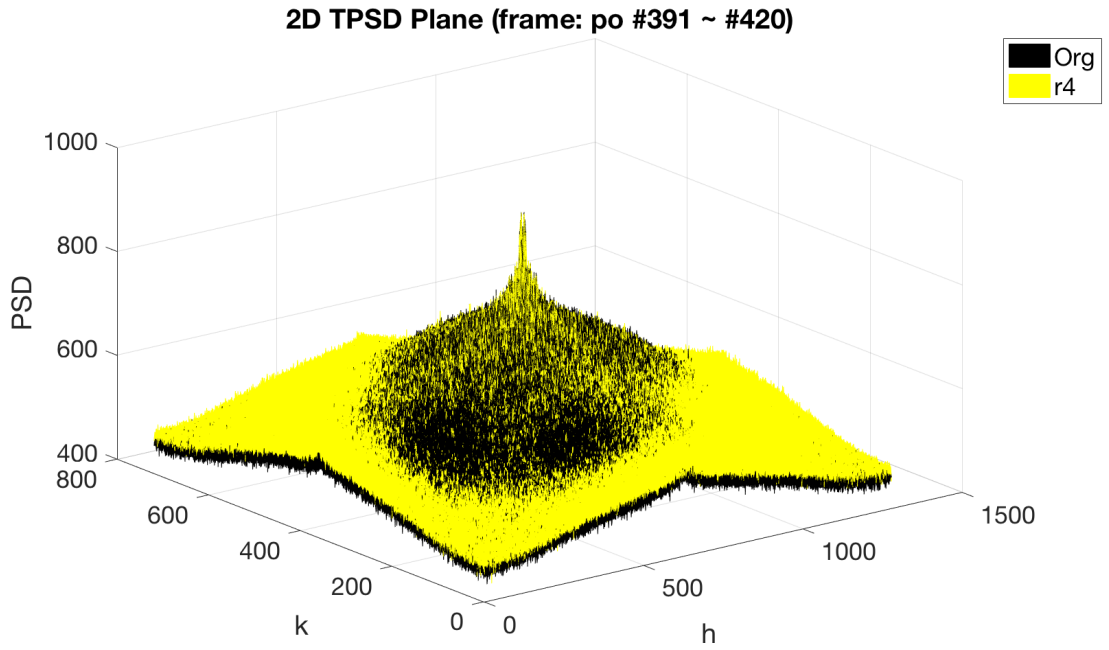
- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4

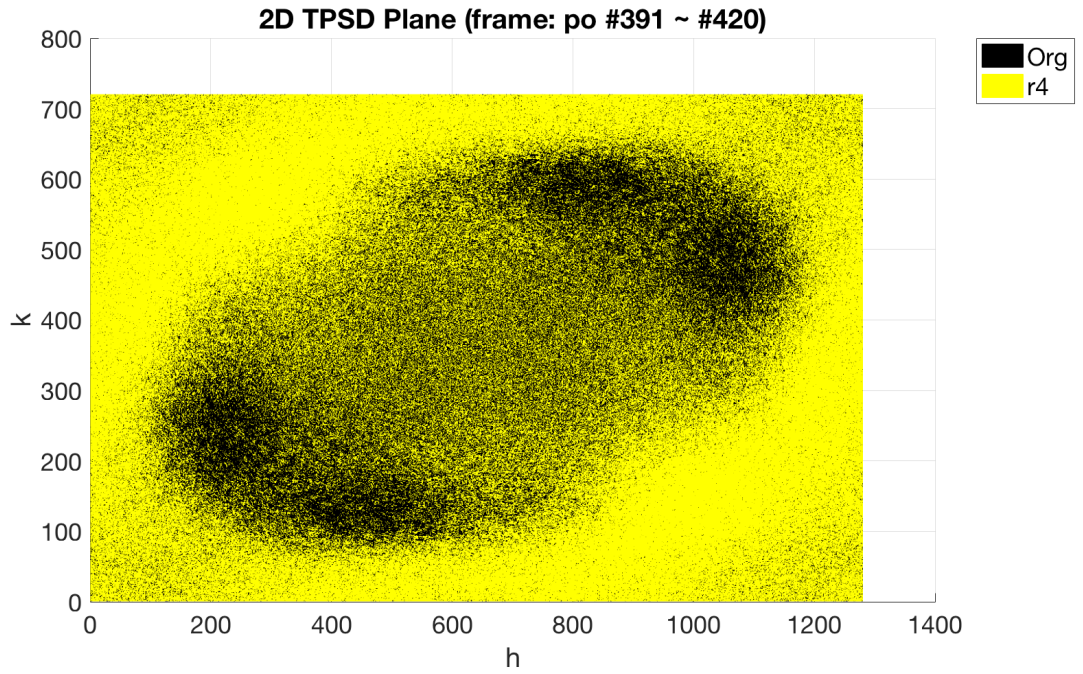




Incremental Change of Distortion Level

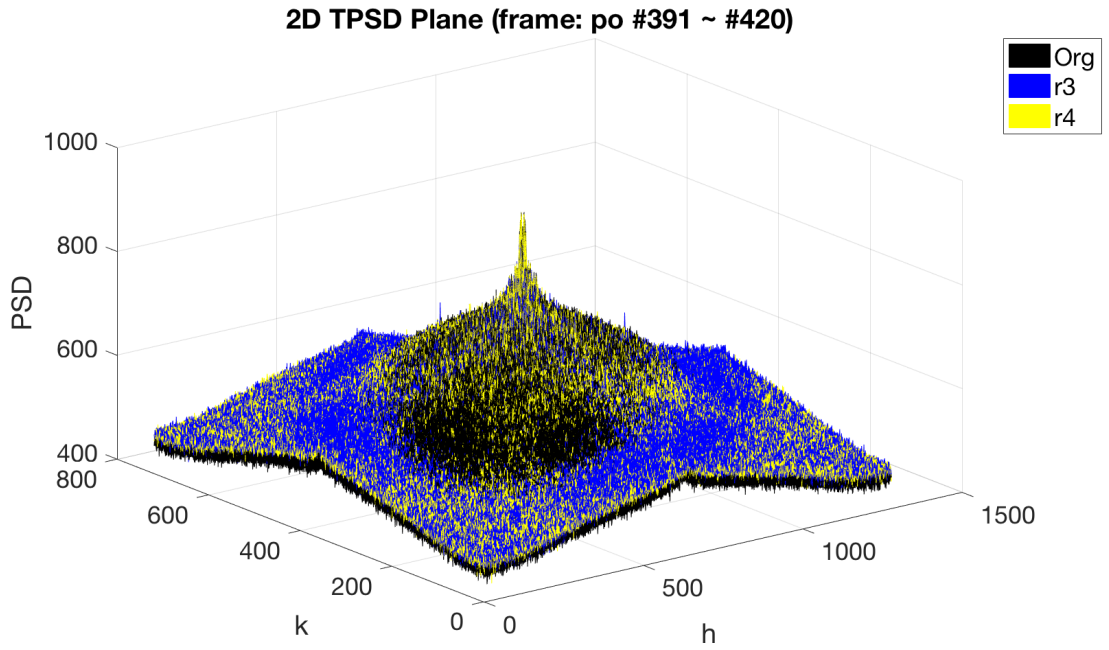
- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4

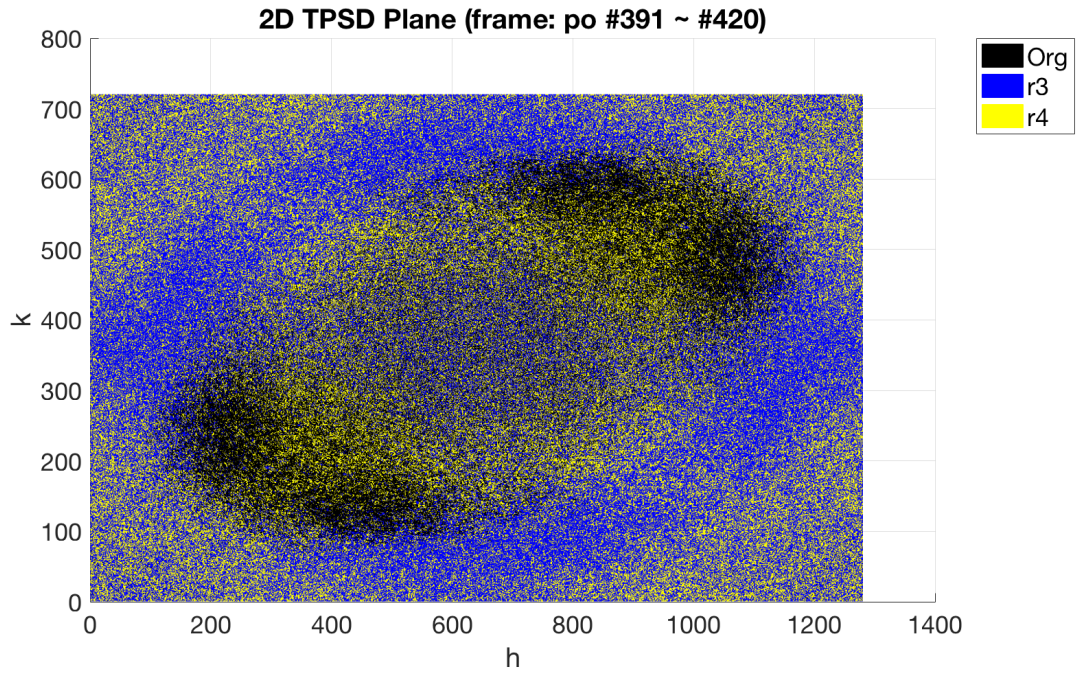




Incremental Change of Distortion Level

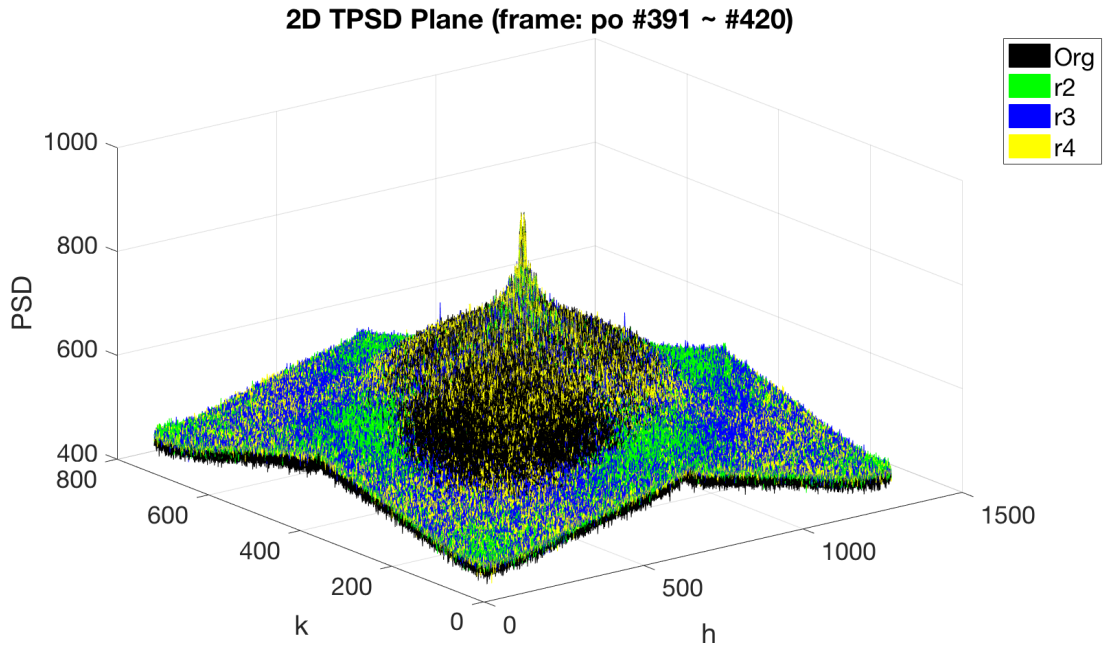
- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4

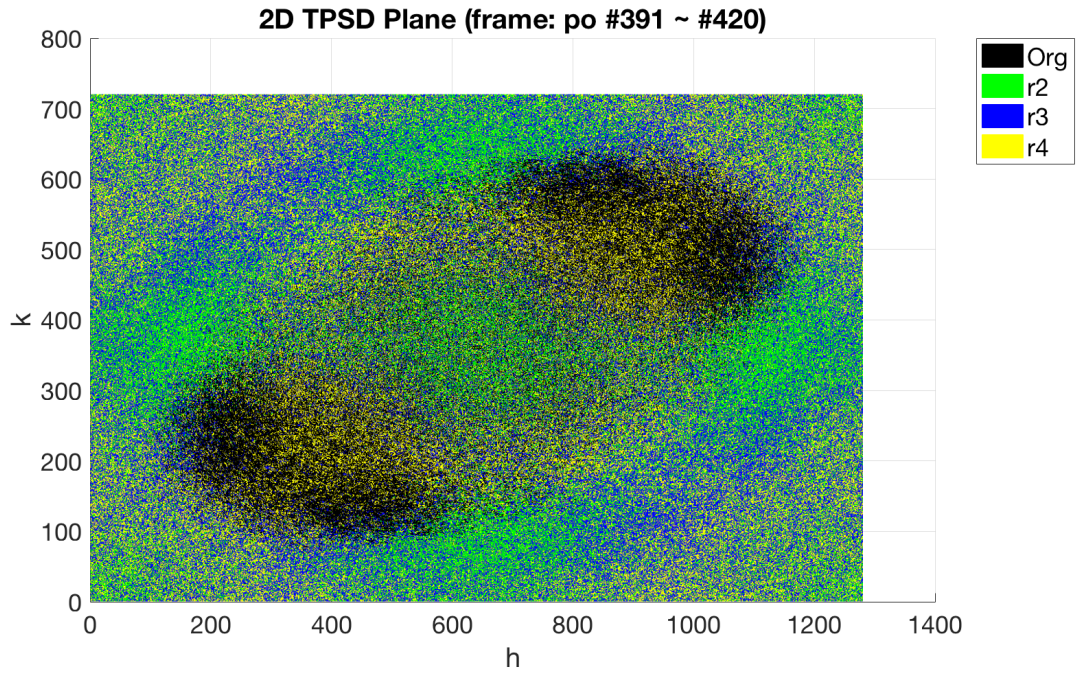




Incremental Change of Distortion Level

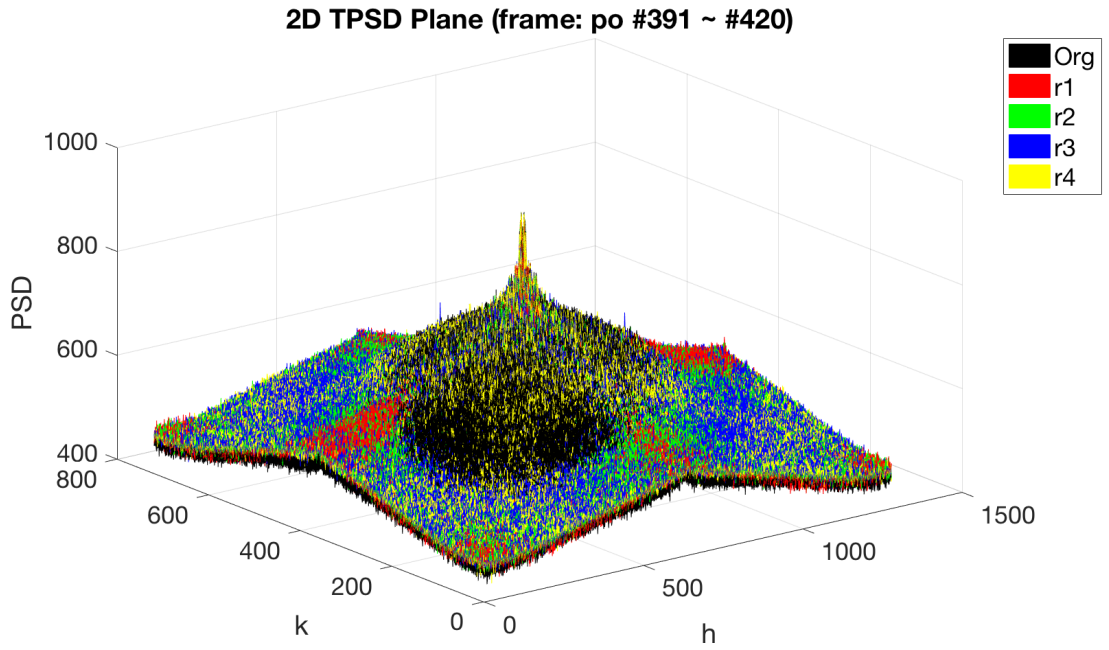
- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4

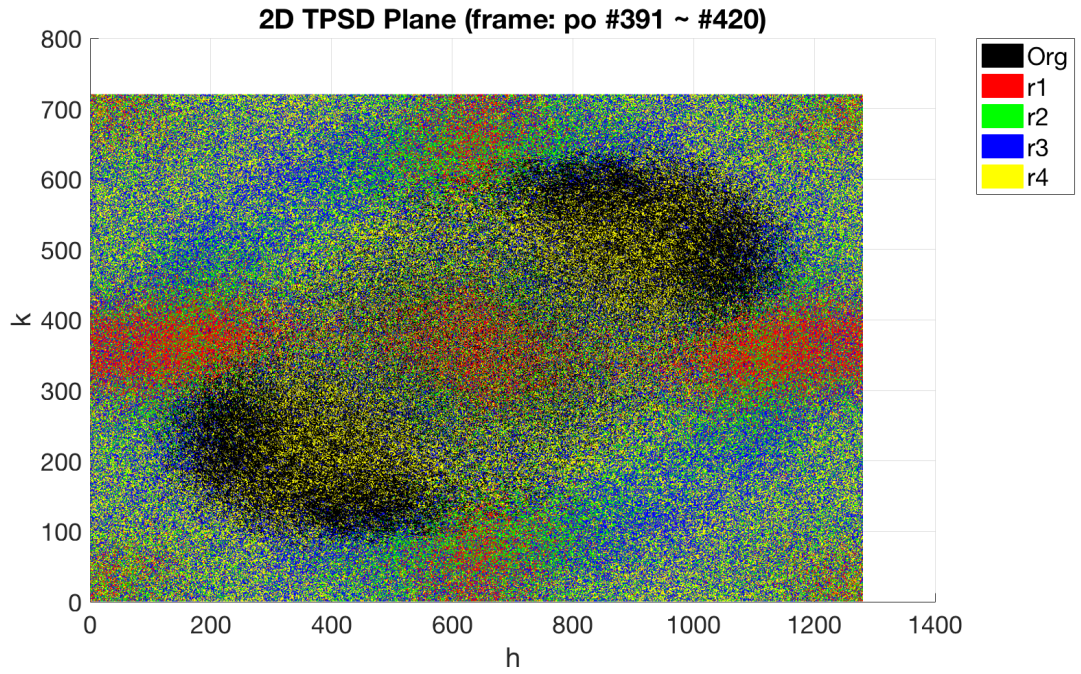




Incremental Change of Distortion Level

- Example
  - Distortion level : r1>r2>r3>r4>Org (r: compression artifact)
  - PSNR, SSIM: r1<r2<r3<r4





Outline

I. Introduction

II. Literature Surveys

III. Proposed Method

   a) Feature Extraction Using 3D Power Spectral Density

   b) Mapping Differences in Features to Human Perception

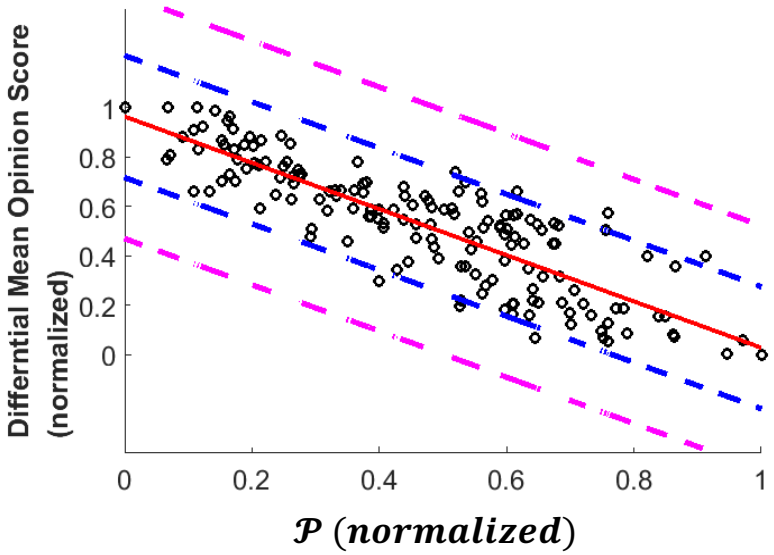
   c) Feature Evaluation

IV. Experiments and Results

V. Conclusion

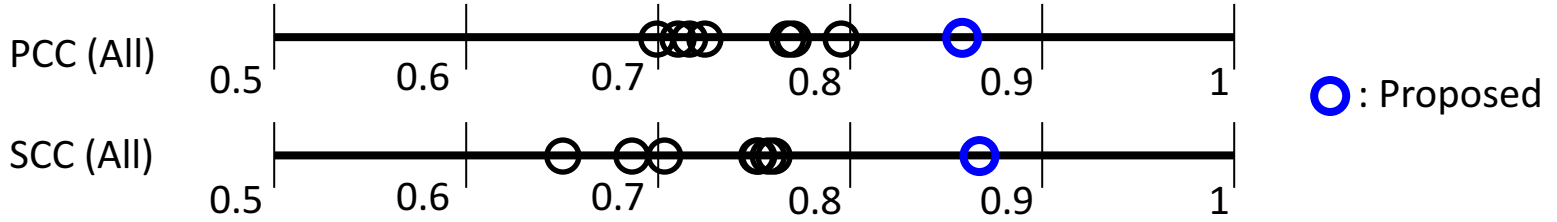
Results

- LIVE Mobile Video Quality Assessment Database (160 Distorted Videos)
  - Co: Compression artifact, Wl: Wireless packet loss, Ra: Rate adaptation, Td: Temporal dynamics



Metric	Pearson Correlation Coefficient (PCC)				
	Co	Wl	Ra	Td	All
PSNR	0.784	0.762	0.536	0.417	0.691
VQM	0.782	0.791	0.591	0.407	0.702
MOVIE	0.810	0.727	0.681	0.244	0.716
MS-SSIM	0.766	0.771	0.709	0.407	0.708
VIF	0.883	0.898	0.664	0.105	0.787
VSNR	0.849	0.849	0.658	0.427	0.759
NQM	0.832	0.874	0.677	0.365	0.762
Proposed	**0.951**	**0.949**	**0.856**	**0.800**	**0.850**

Metric	Spearman Correlation Coefficient (SCC)				
	Co	Wl	Ra	Td	All
PSNR	0.819	0.793	0.598	0.372	0.678
VQM	0.772	0.776	0.648	0.386	0.695
MOVIE	0.774	0.651	0.720	0.158	0.642
MS-SSIM	0.804	0.813	0.738	0.397	0.743
VIF	0.861	0.874	0.639	0.124	0.744
VSNR	0.874	0.856	0.674	0.317	0.752
NQM	0.850	0.899	0.678	0.238	0.749
Proposed	**0.959**	**0.952**	**0.879**	**0.811**	**0.858**



Results

- Computation Time
  - Sequence: ‘harmonicat’ in LIVE Mobile VQA DB (#201 ~ #320, total 120 frames)
  - PC information: Core™ i7-6700K CPU @ 4.00GHz, 32.0 GB RAM, MATLAB R2015(b)
  - Proposed method requires only 5.88% of computation time required by VIF and 25.26% of computation time require by NQM

Metric	Computation time		
	VIF	NQM	Proposed
Time [sec]	255.729	59.490	**15.030**

Since **3D DFT** is simple and fast domain transform, the proposed method is computationally inexpensive

Outline

I. Introduction

II. Literature Surveys

III. Proposed Method

   a) Feature Extraction Using 3D Power Spectral Density

   b) Mapping Differences in Features to Human Perception

   c) Feature Evaluation

IV. Experiments and Results

V. Conclusion

Conclusion

- We propose a full-reference perceptual video quality assessment metric through 3D PSD analysis
  - 3D processing incorporates spatial and temporal features simultaneously
  - Power spectrum is affected by different types and levels of distortions

- This work does not make any assumption on coding conditions or video sequence

- The proposed metric has a low computational complexity
  - Simple 3D DFT operation

Thank you for attention

{maabed, gukyeong.kwon*, alregib}@gatech.edu

Codes to reproduce the results in this work are available in our group website:

ghassanalregib.com