

FINE TUNING YOUR ADAPTIVE GROUPS WITH OBJECTIVE QUALITY METRICS

Jan Ozer

www.streaminglearningcenter.com

[jozer@mindspring.com/](mailto:jozer@mindspring.com)

[276-238-9135](tel:276-238-9135)

@janozer

Agenda

- Overview of Objective Quality Metrics
- Configuring your x264 encodes
- Measuring adaptive groups
- Choosing the optimal resolution
- Computer requirements

What Are Objective Quality Metrics

- Mathematical formulas that (attempt to) predict how human eyes would rate the videos
 - Faster and less expensive
 - Automatable
- Examples
 - Peak Signal to Noise Ratio (PSNR)
 - Structural Similarity Index (SSIM)
 - Video Quality Metric (VQM)
 - SSIMPlus

Subjective vs. Objective Visual Quality

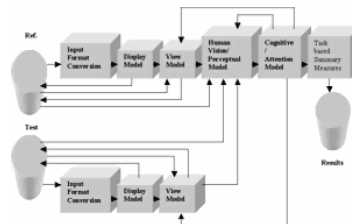
Standards-based



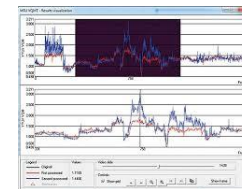
Informal



Perceptual Quality Analyzers

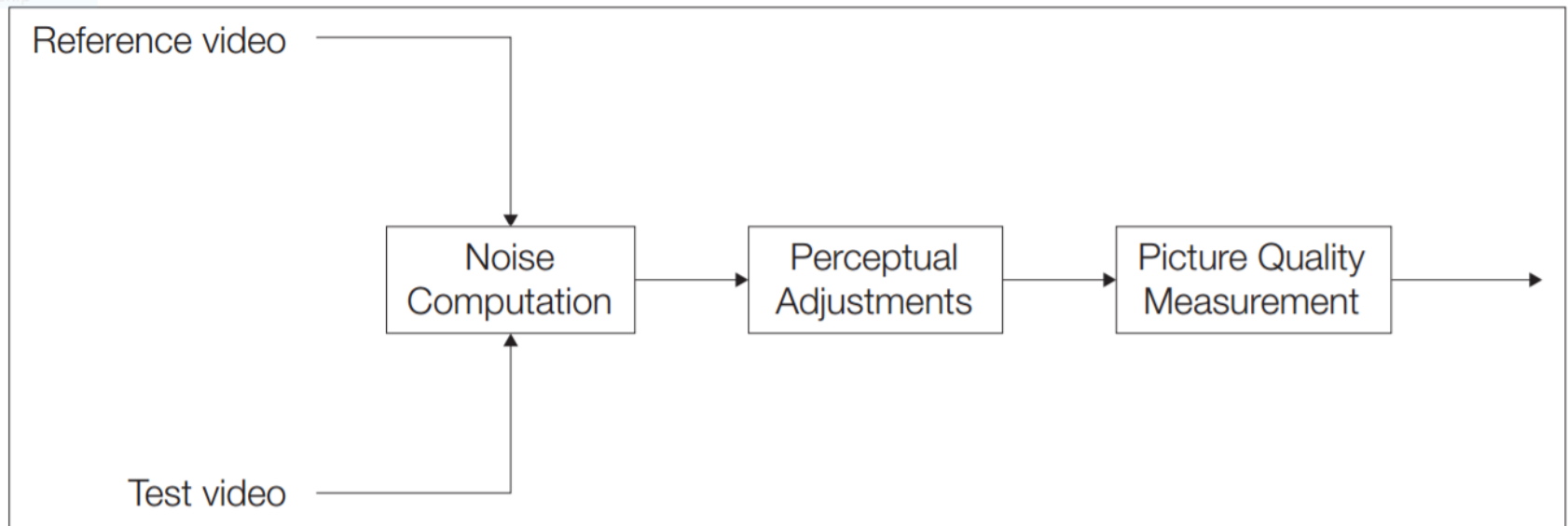


Mathematical (MSE-based)



What are they?	Formal standards	Informal	Perceptual Quality Analyzers	Pure Math-based Quality Models
Example	ITU-T P.910 recommendation	Golden Eye Testing	PQA (Tek), DMOS, SSIMplus, VMAF (Netflix)	PSNR, SSIM
Pros	Gold standard	Accessible	Fast, simple to apply, good correlation to subjective	Fast, simple to apply, cheap
Cons	Time consuming, inappropriate for production	Time consuming	Expensive Some are proprietary	Low correlation with subjective benchmarks

Differentiating Objective Quality Metrics



PSNR
SSIM

MS SSIM

SSIMPlus

PQA
AWDMOS

Measure of Quality Metric

- Role of objective metrics is to predict subjective scores
- Correlation with Human MOS (mean opinion score)
 - Perfect score - objective MOS matched actual subjective tests

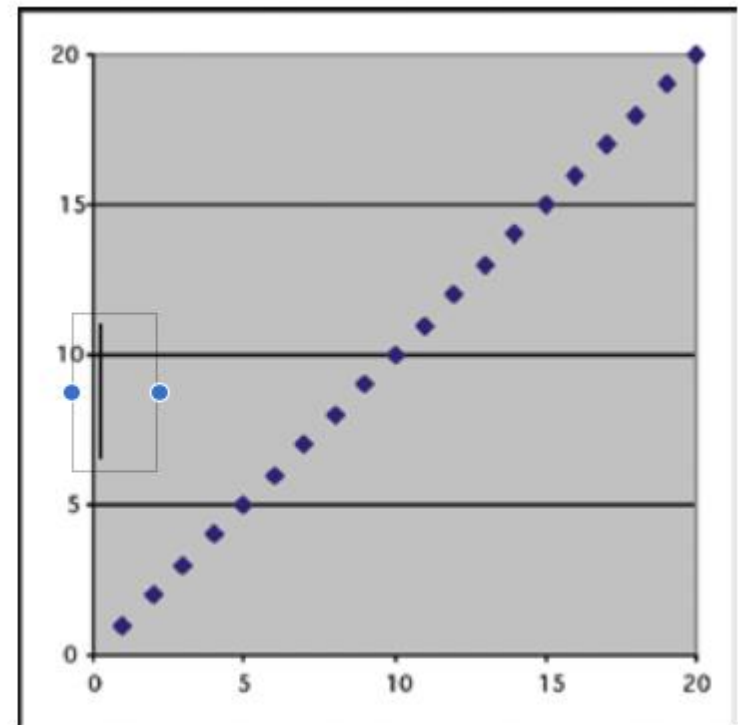
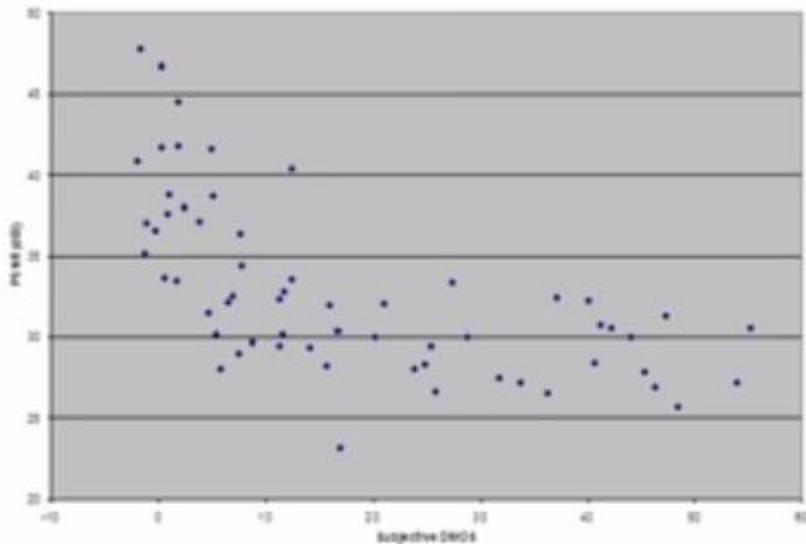


Figure 10. Correlation coefficient: 1.

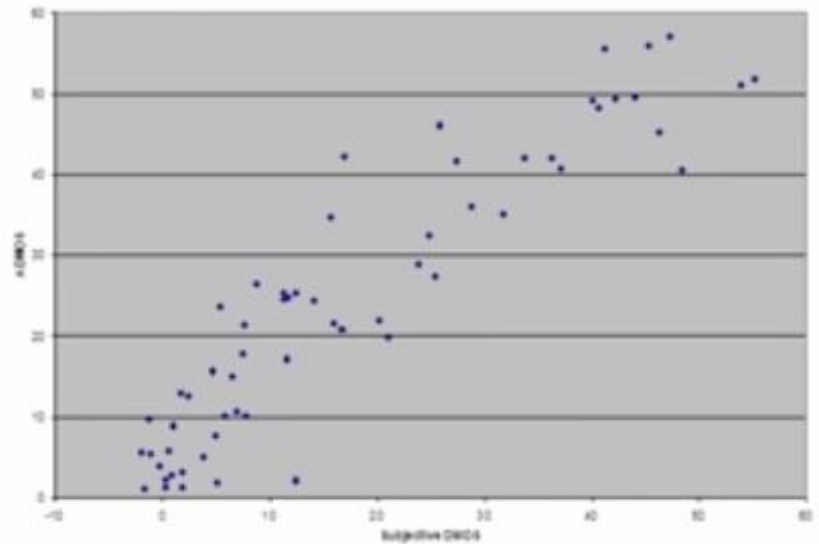
Measure of Quality Metric

- Correlation with Human DMOS (Difference mean opinion score)

PSNR vs Subjective rating



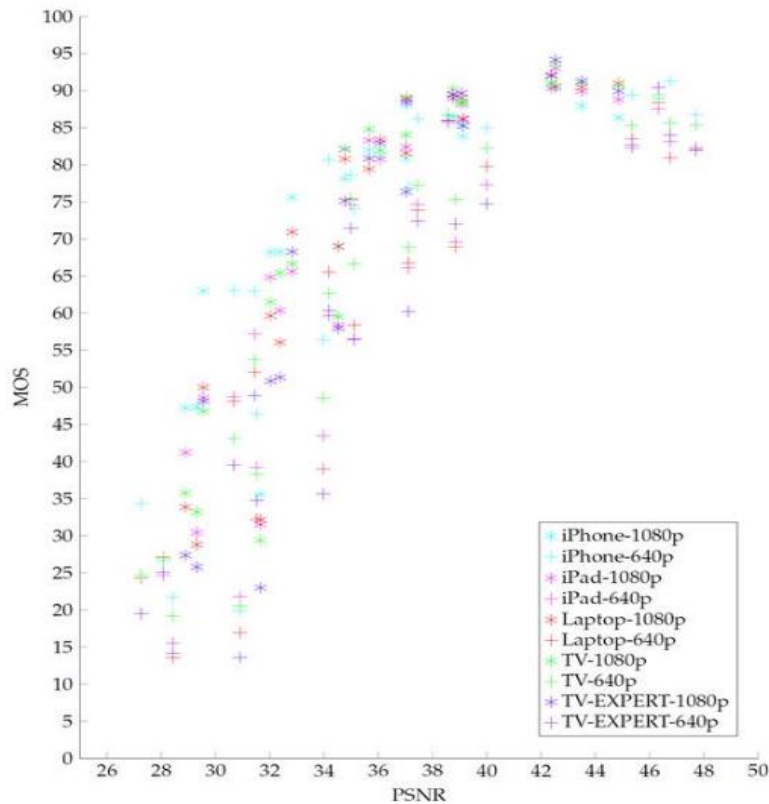
ADMOS vs Subjective rating



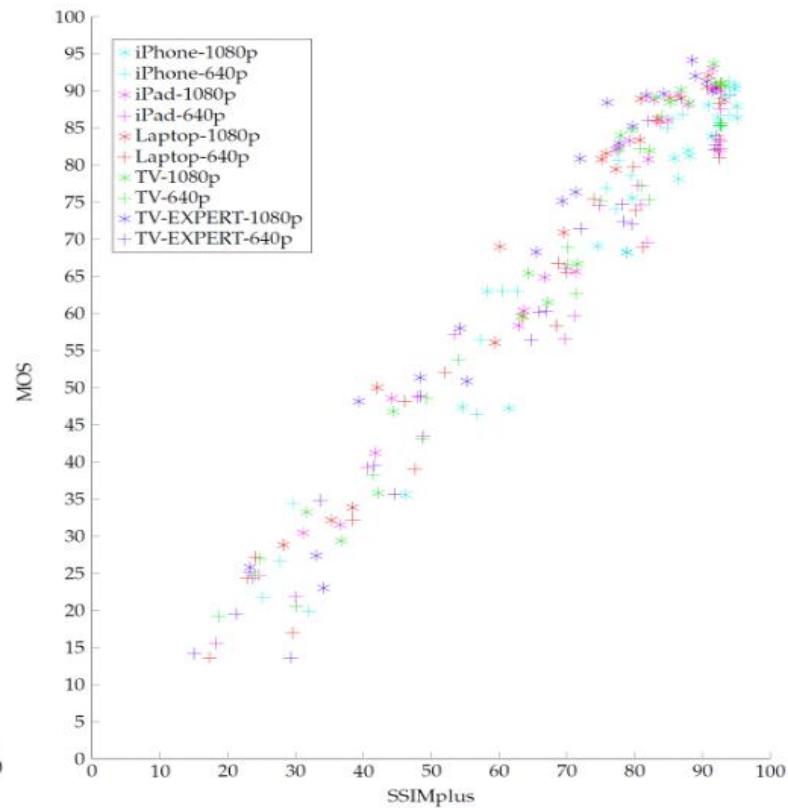
Tektronix

Measure of Quality Metric

PSNR



SSIMplus



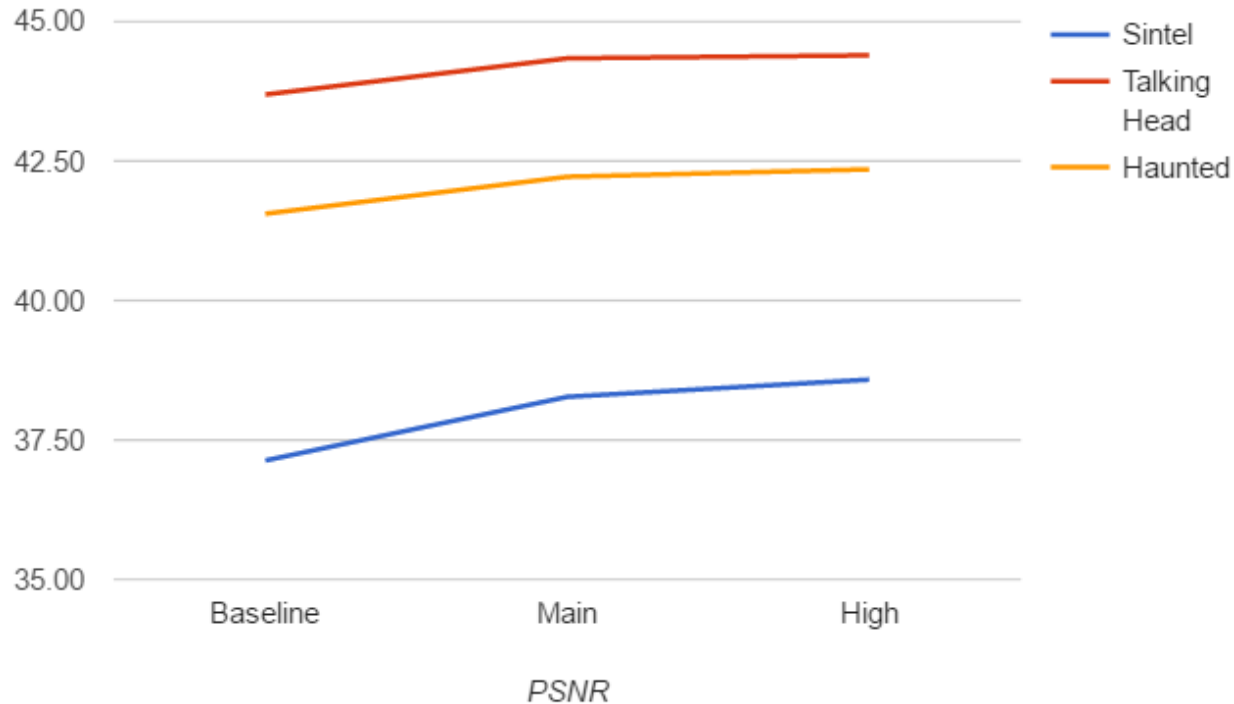
Metrics Taxonomy

	PSNR	MS SSIM	SSIMPlus	PQR	AWDMOS
Basis	Error	Some perceptual	More perceptual	Even More	Even More
Predictive value	Fair	Fair+	Very Good	Very Good	Best
Device specific	No	No	Yes	Yes	Yes
Attention Weighting	No	No	No	Yes	Yes
Score correlation	Some	No	Yes	Yes	Kind of
Cost	Free	\$999	~\$4K	\$19K	\$19K

Comparing the Metrics

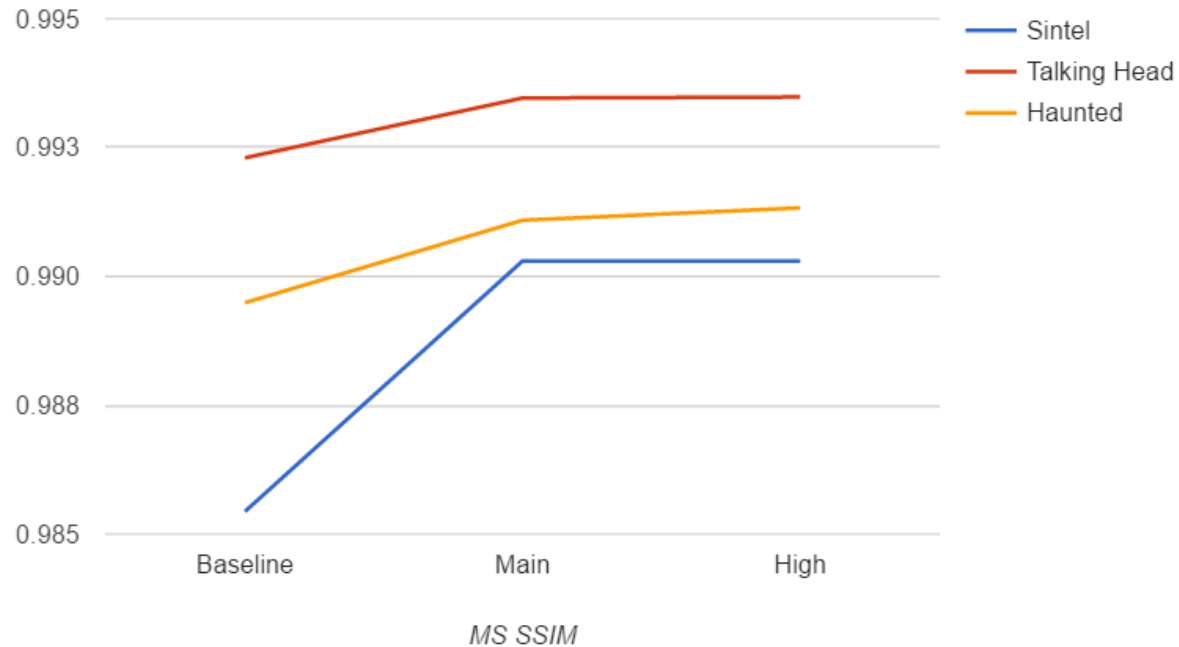
- Encode three files, 720p 1.5 Mbps – 3 Mbps
 - Baseline, Main, High
 - Measure with different tools
 - Draw conclusions about comparative quality

Peak Signal To Noise Ratio



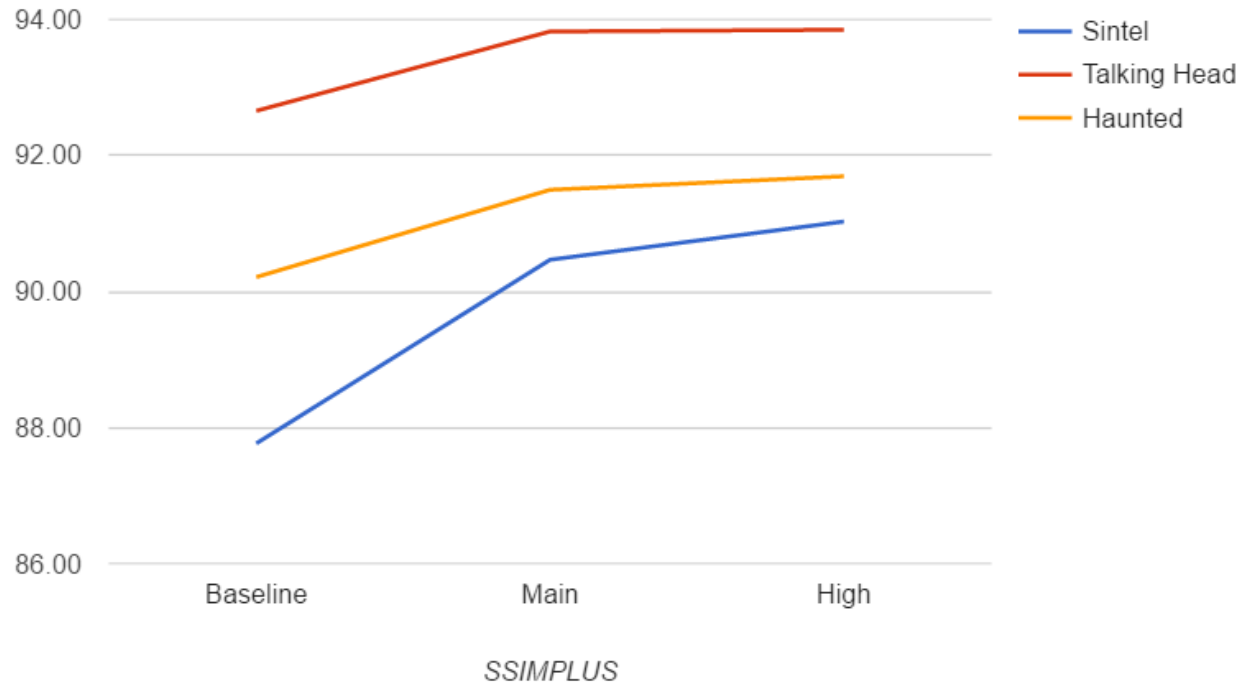
- 0 – 100, Higher scores better
- Interpreting scores
 - Higher than 45 dB undiscernible
 - Lower than 35 usually indicates issues
- Results:
 - Sintel lowest by far
 - Talking head best
 - Difference between profiles not particularly meaningful

Multi Scale Structural Similarity



- 0 – 1 scale, higher scores better
- Interpreting scores
 - Just higher scores better
- Results:
 - Sintel lowest by far
 - Talking head best
 - Sintel
 - Small numerical delta (.05); Baseline to Main, looks steep
 - Other steps not significant

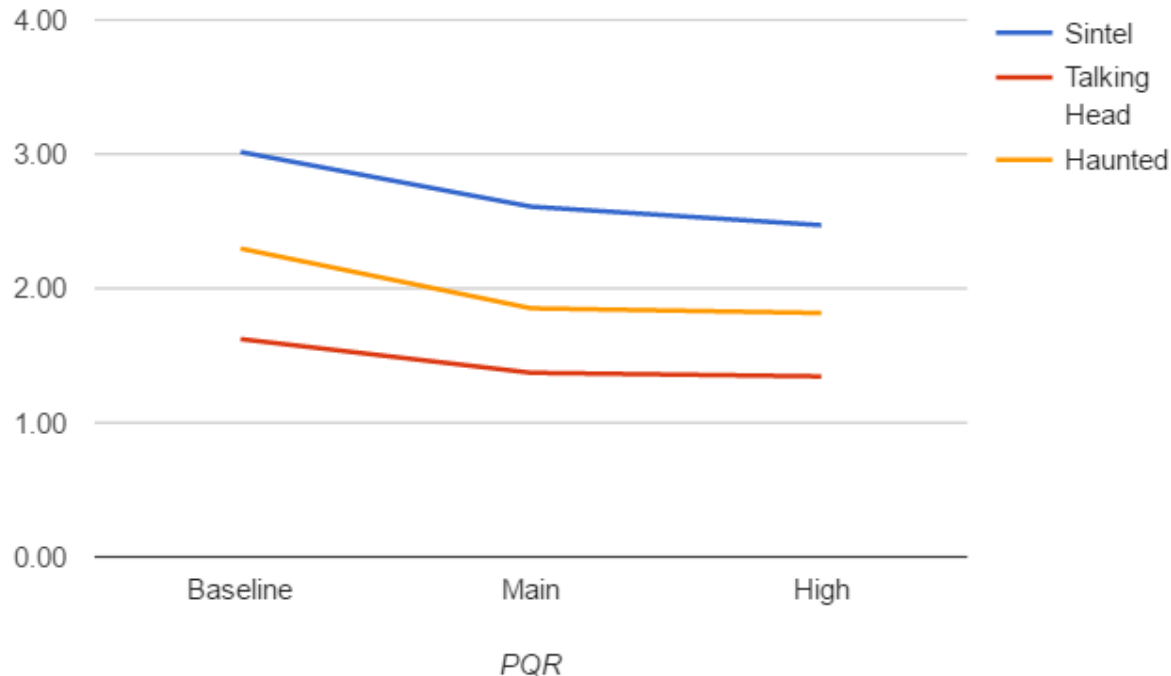
SSIMPlus



- 0 – 100 scale, higher scores better
- Interpreting scores
 - 80 – 100 – s/be perceived as excellent
 - 60 – 80 – good, and so on

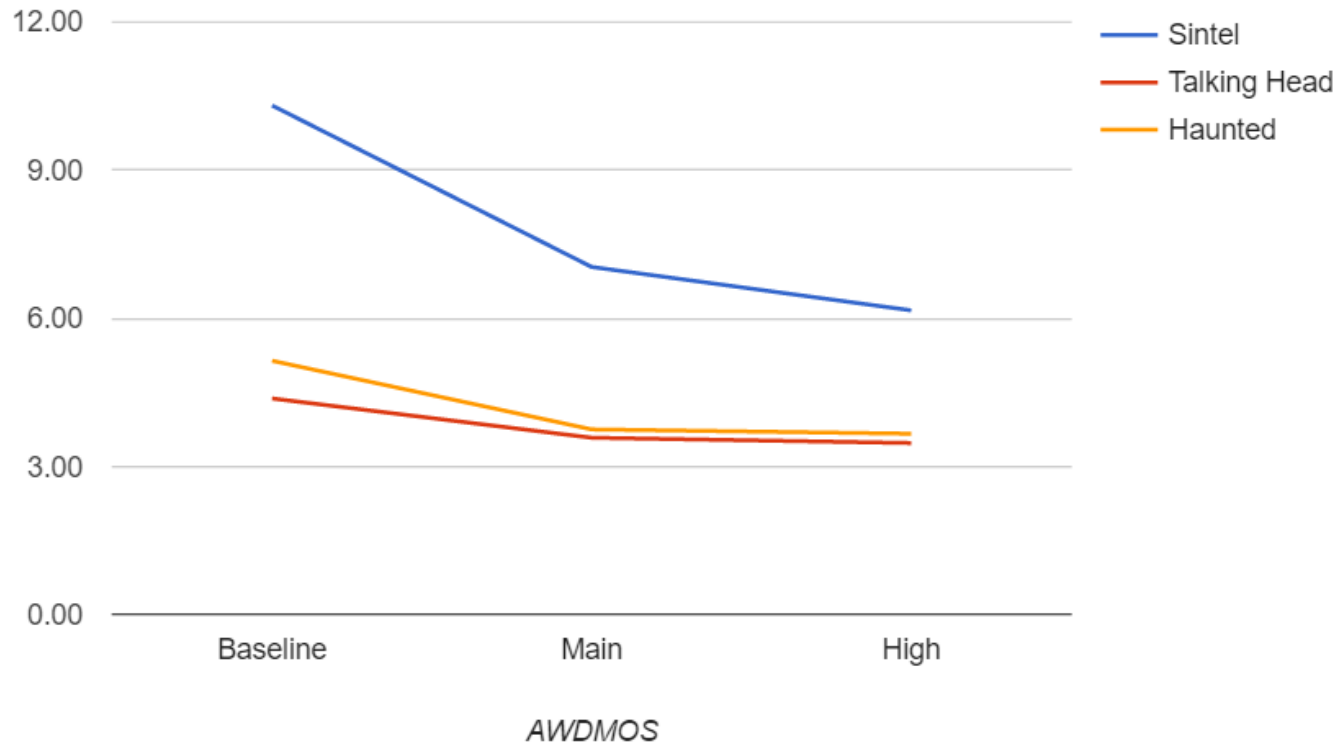
- Results:
 - Sintel lowest by far
 - Talking head best
 - Sintel
 - Small numerical delta (2); Baseline to Main, looks steep
 - All scores comfortably in excellent range

Picture Quality Rating



- 0 – 100 scale, lower scores better
- Interpreting scores
 - 1 PQR = 1 JND – hard to distinguish
 - 2 JND ~ 90% of viewers can tell videos apart
- Results:
 - Sintel lowest by far
 - Talking head best
 - No delta is greater than about .5 JND – most viewers could not tell apart

Attention Weighted DMOS



- 0 – 100 scale, lower scores better
- Interpreting scores
 - It's complicated – DMOS usually 0-100
 - Real subjects seldom rate at extreme ends of scale
 - Don't know if video is absolute best or worst
- Results:
 - Sintel lowest by far
 - Talking head best
 - Largest differential is Sintel, ~ 3 from Baseline to Main
 - Still in excellent range
 - Defer to PQR and say viewers wouldn't notice

The Bottom Line

- In this single test, PSNR delivered results similar to other, higher quality metrics
- Netflix used PSNR for their per-title analysis until mid-2016
- PSNR has many deficits
 - No tuning for specific playback devices
 - No attention weighting (on most tools)
 - No hard correlation to subjective perception

The Bottom Line

- In the land of the blind, the one-eyed man is king
 - Very useful for day-to-day configuration decisions
 - Very affordable and technically accessible
- Would I use Tektronix tool if I had it to keep?
 - Absolutely
 - But I don't have \$19K to spend (for tool + batch capability)
so PSNR/SSIMPlus will have to do

Took Me From Here

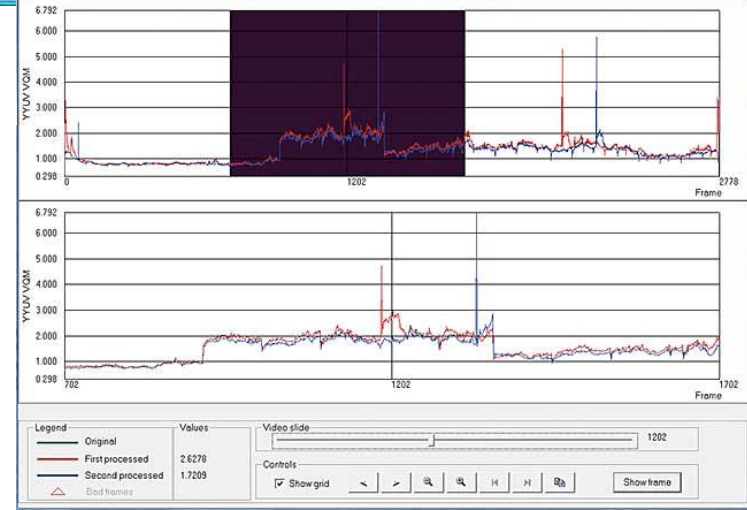
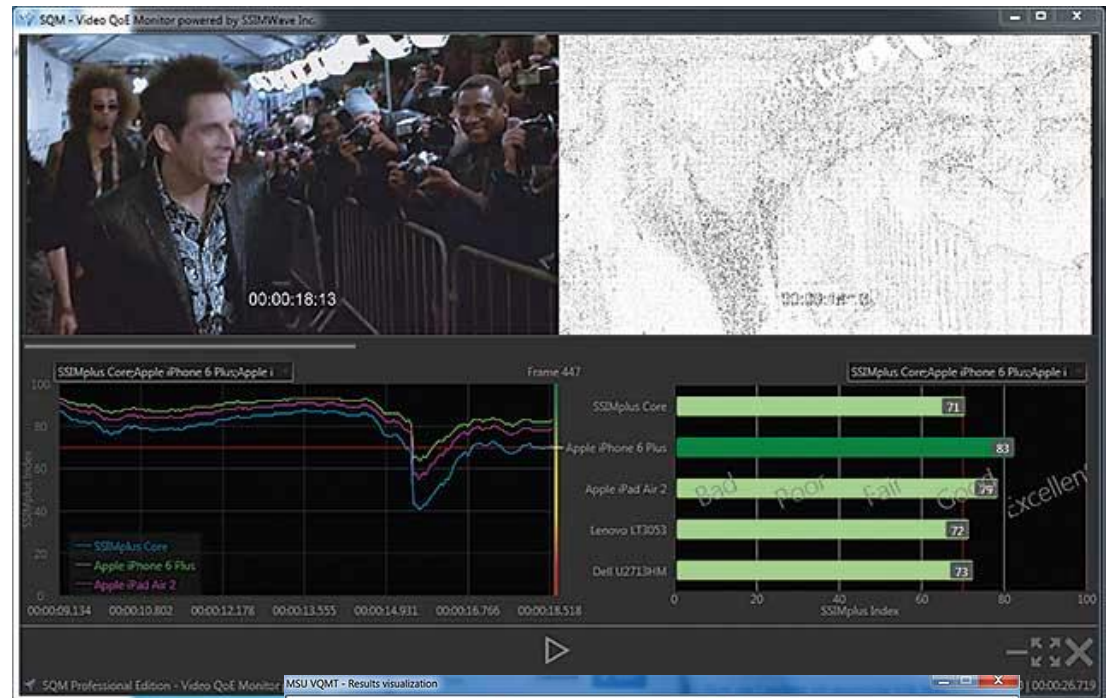


Time consuming and error prone
Subjective comparisons

To Here

VQM (lower is better)					
	Codec A	Codec B	Codec C	High > Low	Codec A > Codec B
Office 1	0.36	0.36	0.37	-3.54%	0.61%
Office 2	0.69	0.61	0.70	-13.51%	12.32%
Office 3	0.28	0.28	0.32	-14.74%	1.32%
Office 4	0.87	0.79	0.87	-9.63%	9.63%
Parking 1	0.68	0.61	0.74	-21.23%	10.90%
Parking 2	0.57	0.55	0.64	-15.47%	3.04%
Parking 3	1.86	1.58	1.76	-17.88%	17.88%
Parking 4	0.47	0.49	0.51	-8.86%	-3.81%
Retail 1	0.56	0.54	0.56	-4.27%	4.27%
Retail 2	0.68	0.66	0.69	-4.45%	3.39%
Retail 3	0.78	0.72	0.76	-8.64%	8.64%
Retail 4	0.73	0.67	0.88	-32.16%	8.52%
Traffic 1	0.55	0.50	0.58	-15.89%	9.14%
Traffic 2	0.34	0.32	0.38	-17.79%	6.39%
Traffic 3	0.52	0.49	0.55	-11.42%	5.29%
Traffic 4	0.68	0.61	0.66	-11.56%	11.56%
Total	10.61	9.78	10.96		
7.84%	Difference between Codec A and Codec B				
-3.34%	Difference between Codec A and Codec C				
-12.13%	Difference between Codec B and Codec C				
	0.61				
	Green equals best in category				
	Orange means worst in category				
	Difference greater than 7.5%				

Statistically meaningful comparisons



With Objective Quality Metrics You Get

- More data
 - Can run many more tests in much less time
- Better data
 - Mathematical models can measure smaller changes than your eye can easily discern
- High level operation
 - Input source and test file(s)
 - Test program delivers a score

Trust, But Verify



- Never rely solely on objective test results
- Compare files yourself to verify comparisons
 - Still image comparisons
 - Side by side real time playback

The Tools I use

- Moscow University Visual Quality Comparison Tool (VQMT)
 - Developed by same group that outputs H.264/HEVC comparisons
 - Typically use PSNR
- SSIMWave Video Quality-of-Experience Monitor (SQM)
 - From one of the inventors of SSIM metric

VQMT Workflow

The screenshot shows the MSU Video Quality Measurement Tool interface. It is divided into three main steps: Step 1: File selection, Step 2: Metric Selection, and Step 3: Output Selection. Step 1 includes fields for 'Original file' and 'Processed (compressed)' files, with 'Browse' and 'Preview' buttons. It also has checkboxes for 'Comparative analysis', 'Use mask file', and 'Use black mask'. Step 2 shows a dropdown for 'Metric Selection' set to 'VQM', with 'Settings' and 'Online metric info' buttons. It also has radio buttons for 'Color component' (Y-YUV, U-YUV, V-YUV, L-LUV, R-RGB, G-RGB, B-RGB). Step 3 includes checkboxes for 'Save CSV file', 'Save metric visualization video / image', and 'Save "bad frames"', with 'Advanced' buttons and a 'More options' button. At the bottom, there is a 'Process' button, 'Website', 'Feedback', 'Help', and 'Exit' buttons. A 'Ready' status indicator and a 'Show results visualization' checkbox are also present. The interface has a blue header with the tool's name and version (5.1 PREBETA PROFESSIONAL). A logo for 'GRAPHICS & MEDIA LAB VIDEO GROUP' is in the bottom right.

MSU Video Quality Measurement Tool

Version 5.1 PREBETA PROFESSIONAL

Step 1: File selection

Original file:
(video file or image sequence) E:\TOS\TOS_720p.mp4 ... Browse Preview

Processed (compressed): E:\TOS\720p\TOS_720p_24_x264_720p_ ... Browse Preview

☒ Comparative analysis

Second processed (another codec): E:\TOS\720p\TOS_720p_24_x264_720p_ ... Browse Preview

Advanced

☐ Use mask file: ... Browse Preview

☐ Use black mask

Step 2: Metric Selection

VQM Settings Online metric info

Color component

☒ Y-YUV ☐ U-YUV ☐ V-YUV ☐ L-LUV ☐ R-RGB ☐ G-RGB ☐ B-RGB

Step 3: Output Selection

☒ Save CSV file Advanced

☐ Save metric visualization video / image Advanced

☐ Save "bad frames" Advanced

More options

Ready ☒ Show results visualization

Process Website Feedback Help Exit

Load Source
File

Load one or
two encoded
files

Choose Metric

Press Process

Results Visualization

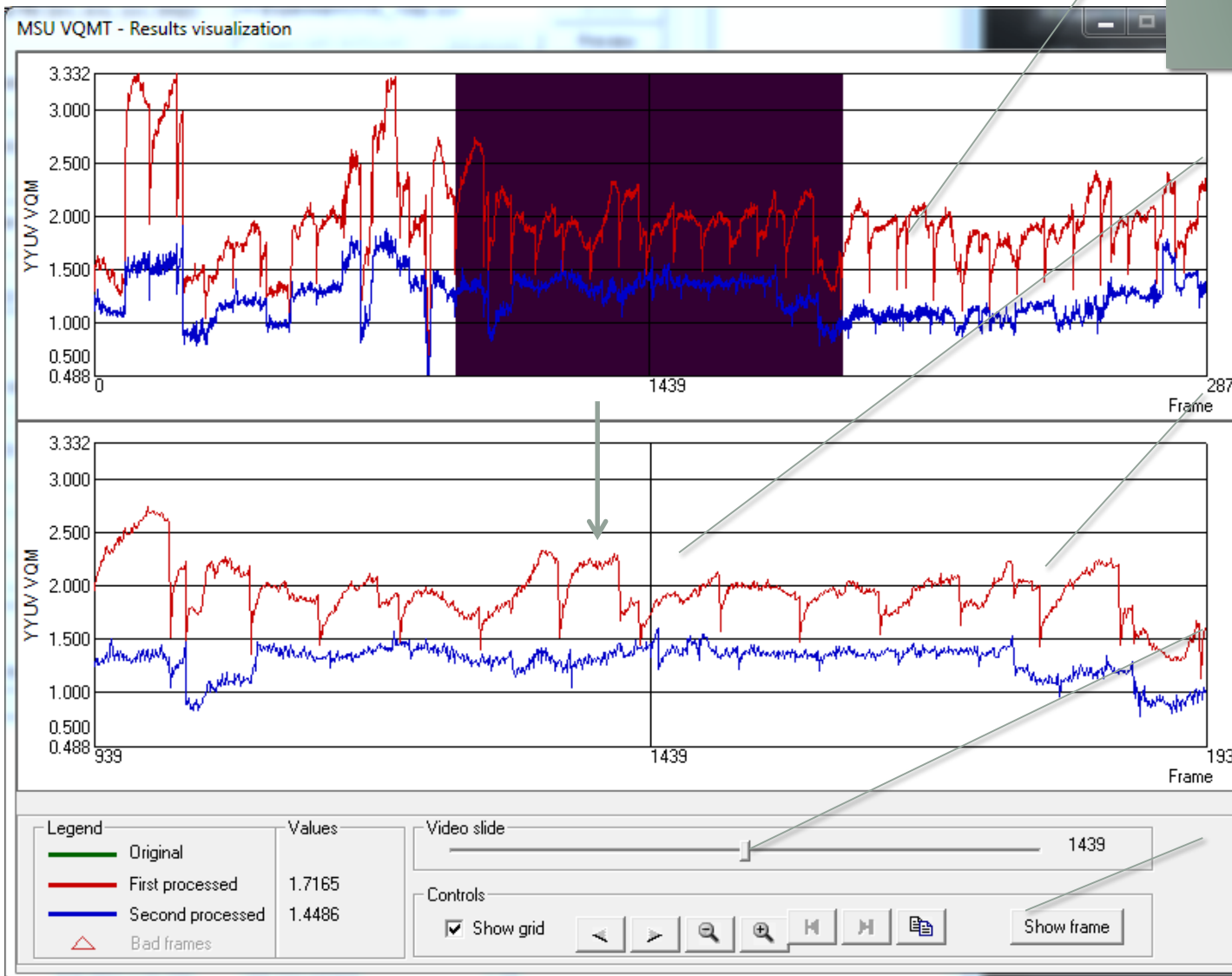
Score entire comparison

Zoom in of black area

Red – first file
Blue – second

Slide through frames

Click to Show Actual Frames



S



Toggle through
source, test
files

Can Zoom In

MSU VQMT

Pros

- Affordable (~\$995)
- Very visual – easy to see test results in actual frames
- Multiple algorithms – PSNR, VQM, SSIM, MS SSIM
- My review of VQMT
 - bit.ly/VQMT_review

Cons

- Can only compare files of:
 - Like resolution
 - Like frame rate
- Scores don't directly correlate to subjective perception
 - Can make some assumptions
- Scores don't correlate to any playback platform (mobile, computer, OTT)

SQM Overview

- Based on SSIMplus Algorithm
- Rates videos on scale that corresponds with human perception
 - 80 – 100 – Excellent
 - 60 – 80 – Good
 - 40 – 60 – Fair
 - 20 – 40 – Poor
 - < 20 – Bad
- Predicts ratings on multiple devices
 - Phones, TVs, monitors, etc.
- Separate command line tool for Windows/Linux
- My review
 - http://bit.ly/SQM_review

SQM Workflow

Load Test File

Load Source File

Load reference and test video files

Test video file Frame offset: 1 Reference video file Frame offset: 0

H:\VP9vsHEVC\265_2015_final\New_1280_x265.hevc H:\VP9vsHEVC\New_1280.mp4

Format: YUV 4:2:0 Format: YUV 4:2:0 Progressive
Resolution: 1280x720 Resolution: 1280x720
Frame rate: 29.970fps Frame rate: 29.970fps
Time duration: 00h:01m:36s Time duration: 00h:01m:36s
Number of frames: 2880 Process frames: 1440 Number of frames: 2880

Settings profile

Quality indices Also include: ☐ PSNR-YUV ☐ PSNR-Y

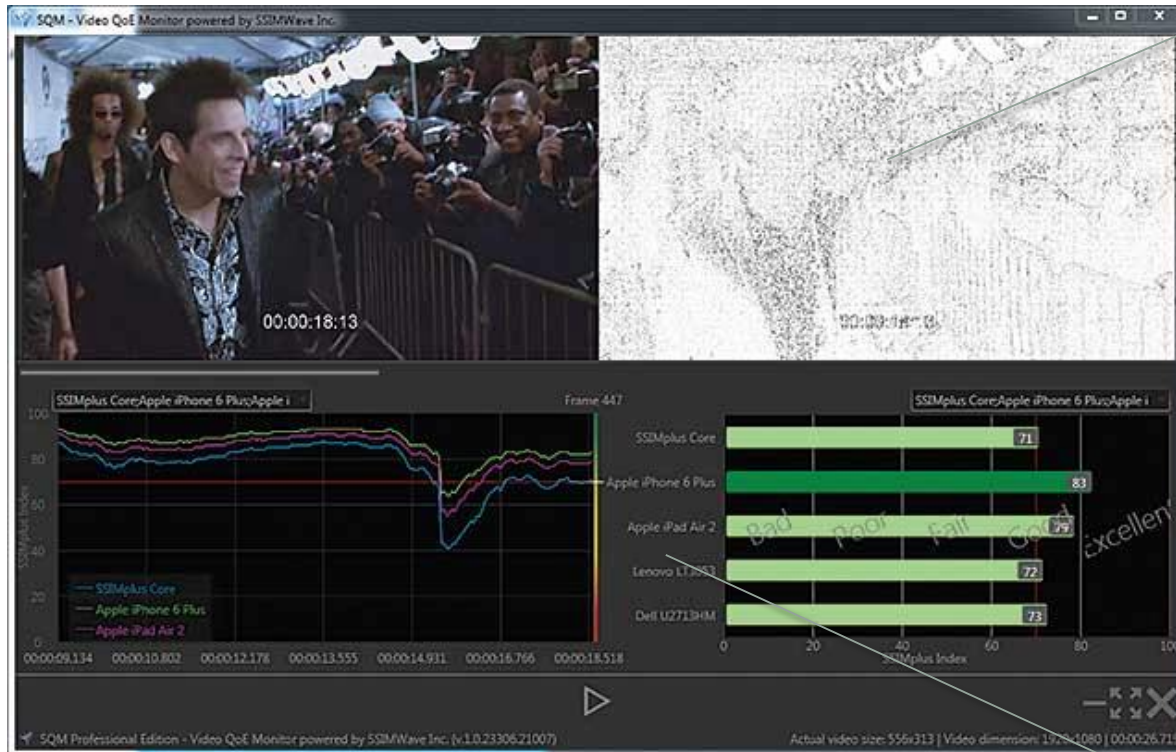
	Manufacturer	Device name	Size	Resol	Category
<input type="checkbox"/>	HTC	One (M8)	5	1920x1080	Phone
<input type="checkbox"/>	Samsung	F8500	64	1920x1080	TV
<input type="checkbox"/>	Panasonic	VT60	50	1920x1080	TV
<input type="checkbox"/>	Sony	W8 (Expert)	50	1920x1080	TV
<input type="checkbox"/>	Samsung	H7150	55	1920x1080	TV

5 more can be selected for playback session.

Cancel Continue

Choose viewing platforms to score

SQM Workflow



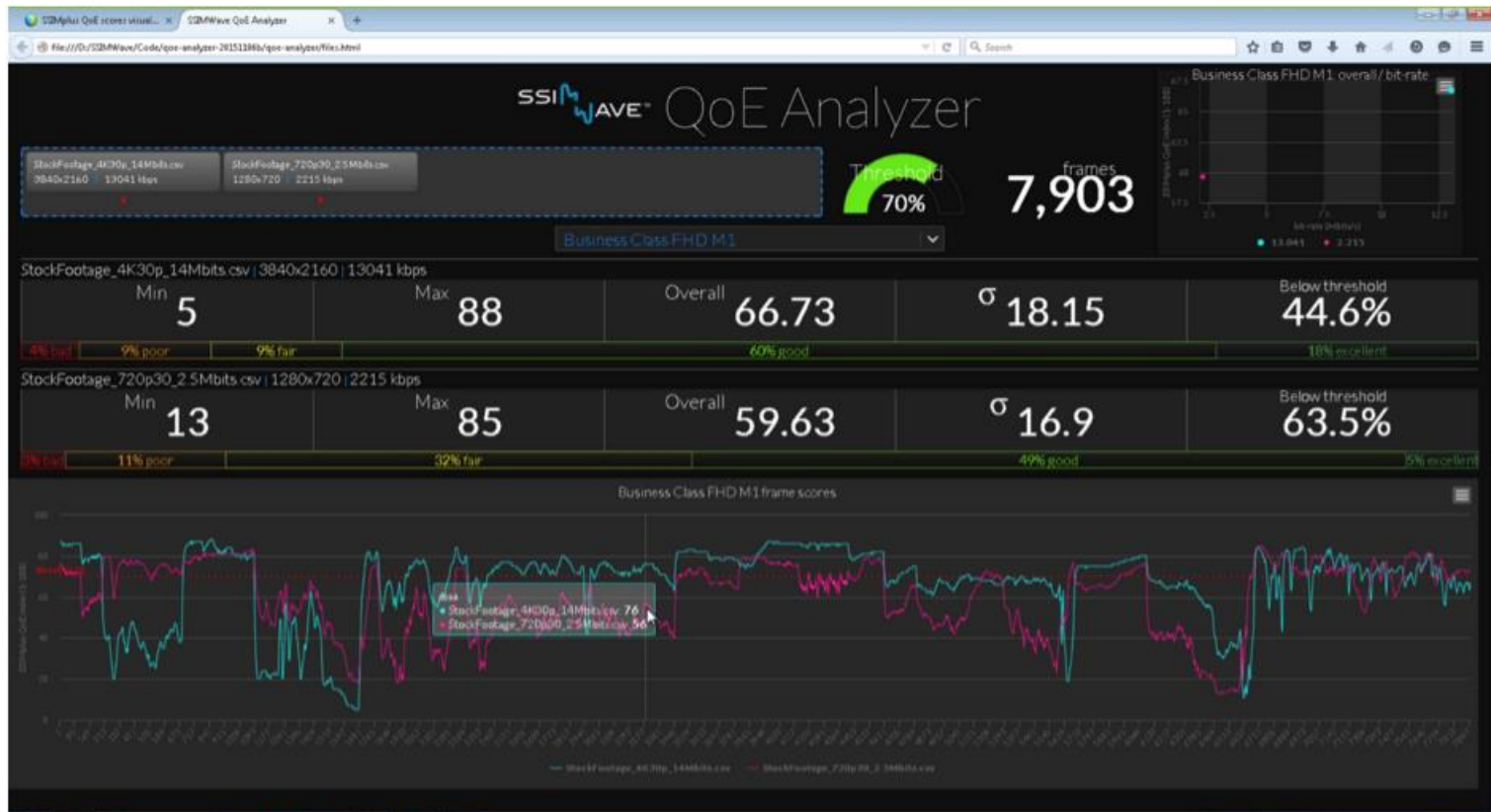
Shows local perceptual quality in compressed file (can toggle to source file)

Device ratings over time

Results Presented in CSV File

SSIMplus QoE Analysis Summary				
Statistics	SSIMplus Core	Sony X9 (Expert)	Samsung HU9000	Lenovo LT3053
Maximum	98	92	97	94
Minimum	78	71	80	74
Mean	95.202	87.002	94.393	90.329
Standard Deviation	1.964	1.949	1.961	1.961
Excellent (81-100)	99.94%	97.14%	99.97%	99.89%
Good (61-80)	0.03%	2.11%	0.00%	0.11%
Fair (41-60)	0.00%	0.00%	0.00%	0.00%
Poor (21-40)	0.00%	0.00%	0.00%	0.00%
Bad (0-20)	0.00%	0.00%	0.00%	0.00%
Below threshold (0-70)	0.00%	0.00%	0.00%	0.00%
Reference video frames c	0			
Test video frames offset	0			

Graphical Results Comparison Tool



Browser based tool for multiple file visualizations

With the Ability to Compare Files

SSIM_wAVE™ Qo

TestVideo_MBR_Lvl3.csv
854x480 | 297 kbps

X

TestVideo_MBR_Lvl4.csv
854x480 | 148 kbps

X

TestVideo_MBR_Lvl1.csv
854x480 | 1379 kbps

X

TestVideo_MBR_Lvl2.csv
854x480 | 630 kbps

X

SSIMplusCore



SSIMWave SQM

Pros

- Unique quality algorithm (SSIMplus)
- Scores correlate with viewer perceptions
- Multiple devices
- Multiple resolutions
- Multiple frame rates (soon)

Cons

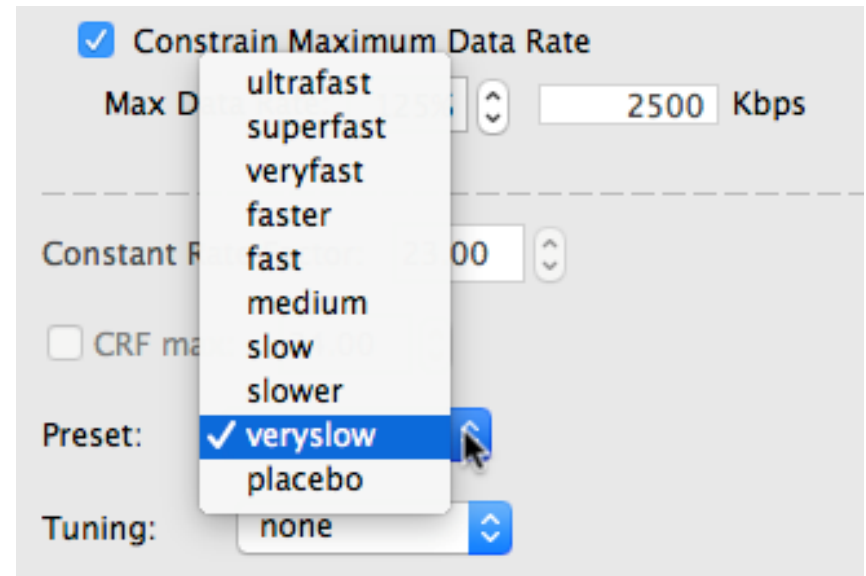
- More expensive (~\$2,400)
- Limited algorithms (SSIM/SSIMPlus/PSNR)
 - Visualization tools not quite as accessible

Configuring Your x264 Encodes

- Taking the guesswork out of:
 - Preset selection
 - Key Frame Interval
 - Data rate control
 - Building your encoding ladder
- All tests performed:
 - FFmpeg/x264
 - 720p files
 - Data rates vary by video file
 - 110% constrained VBR
 - Keyframe of 3 seconds
 - B-frame of 3
 - Reference 5

X264 Preset

- What are presets
 - Simple way to adjust multiple parameters to trade off encoding speed vs. Quality
 - Used by virtually all x264 encoders
 - Medium is generally the default preset



Test Description

- Eight files
 - 1 movie (Tears of Steel)
 - 2 animations (Sintel, BBB)
 - Two general purpose (concert, advertisement)
 - One talking head
 - Screencam
 - Tutorial (PPT/Video)
- Encode to all presets
- Time encoding
- PSNR

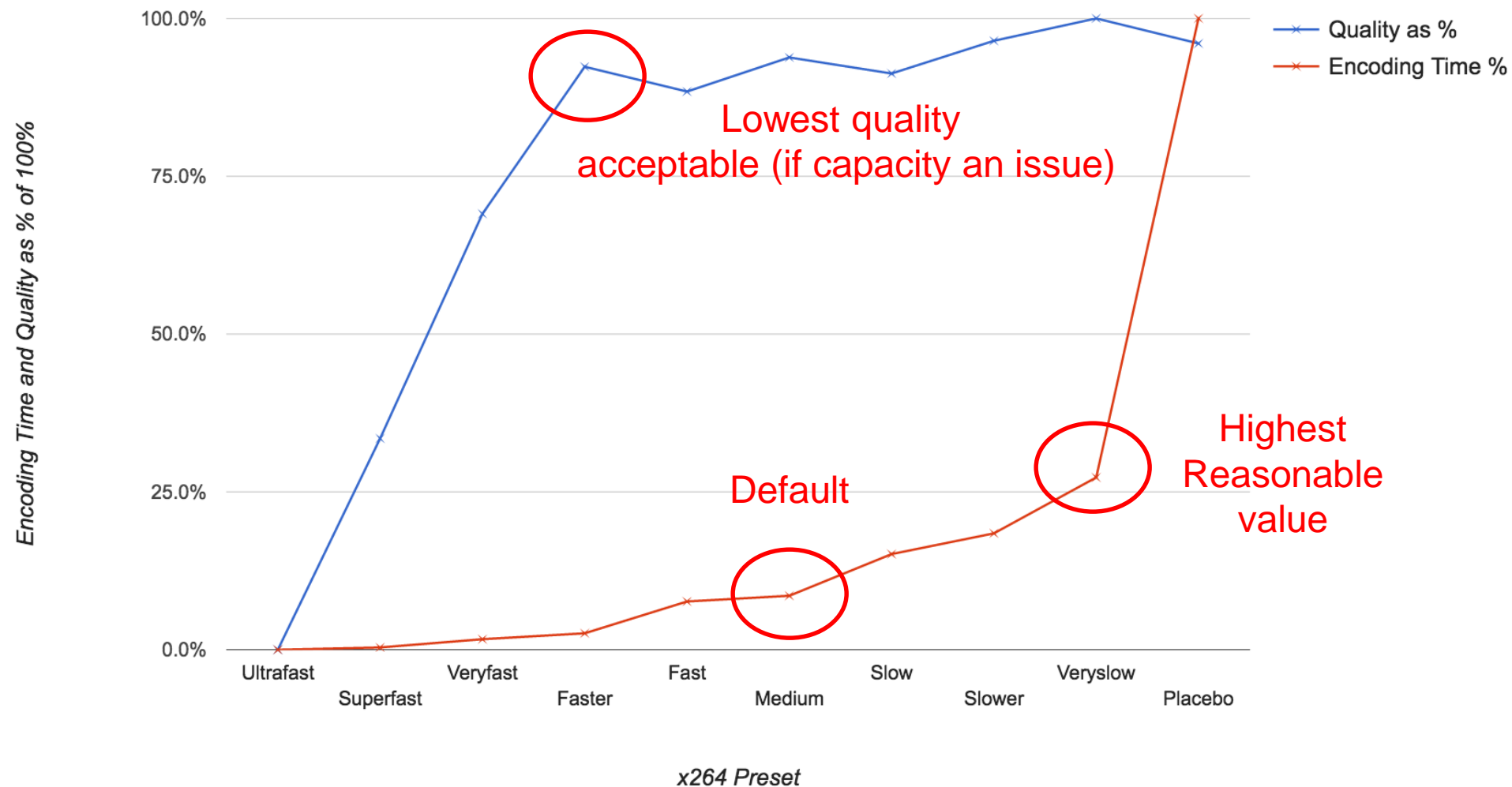
Results Please

	Ultrafast	Superfast	Veryfast	Faster	Fast	Medium	Slow	Slower	Veryslow	Placebo	Total Delta
Tears of Steel	36.07	37.82	38.51	39.23	39.26	39.33	39.27	39.41	39.47	39.40	9.43%
Sintel	35.14	36.71	37.42	38.40	38.43	38.46	38.40	38.55	38.57	38.47	9.75%
Big Buck Bunny	35.19	37.65	38.82	39.49	39.51	39.56	39.50	39.61	39.64	39.54	12.62%
Talking Head	43.38	43.38	44.06	44.39	44.28	44.28	44.21	44.34	44.39	44.29	2.34%
Freedom	38.46	39.26	40.01	40.41	40.32	40.58	40.55	40.69	40.85	40.77	6.22%
Haunted	41.13	41.30	41.89	42.20	42.07	42.27	42.25	42.27	42.35	42.31	2.98%
Screencam	44.46	45.67	46.68	47.12	46.82	46.96	46.95	47.06	46.88	46.76	5.99%
Tutorial	38.47	41.83	43.62	44.50	44.37	44.30	43.99	44.14	44.07	43.91	15.68%
Average	38.23	39.35	40.12	40.69	40.64	40.75	40.70	40.81	40.88	40.80	8.13%

- Red is lowest quality
- Green highest quality
- Very slow averages best quality
 - But only 8% spread between best and worst

Results Please

Videos and Animations: Encoding Time and Quality by Preset



Key Frame Interval

	20 sec	10 sec	5 sec	3 sec	2 sec	1 sec	Total Q
TOS	0.936	0.938	0.949	0.964	0.977	1.024	-9.35%
Sintel	0.926	0.932	0.948	0.955	0.969	1.014	-9.59%
Big Buick Bunny	0.525	0.533	0.525	0.541	0.563	0.616	-17.19%
Screencam	0.478	0.478	0.478	0.480	0.493	0.551	-15.09%
Tutorial	0.671	0.673	0.674	0.674	0.675	0.680	-1.25%
Talking Head	0.567	0.569	0.571	0.572	0.569	0.576	-1.72%
Freedom	1.013	1.014	1.014	1.014	1.019	1.022	-0.93%
Haunted	1.665	1.667	1.669	1.669	1.670	1.677	-0.68%

- Encode with interval of 1, 2, 3, 5, 10, 20 second
- Measure quality with VQM
- Green is best, red is worst
- Anyone using keyframe interval of 1 out there?
 - Difference is modest, but why?
- Recommend 3 for ABR (shorter if shorter chunk size)
- Max 10 for other footage

Reference Frames

- What are they?
 - Frames from which the encoded frame can find redundant information
- What's the trade-off?
 - Searching through more frames takes more time, lengthening the encoding cycle
 - Since most redundancies are found in frames proximate to the encoded frame, additional reference frames deliver diminishing returns

How Much Quality?

720p-110CVBR	1 Ref	5 Ref	10 Ref	16 Ref	Max Delta	10 - 16 Delta	16 - 5 Delta
Tears of Steel	39.34	38.99	39.47	39.49	1.28%	-0.04%	-1.26%
Sintel	38.45	38.54	38.58	38.59	0.35%	-0.02%	-0.12%
Big Buck Bunny	38.38	38.48	38.52	38.51	0.36%	0.03%	-0.08%
Talking Head	44.27	44.36	44.39	44.40	0.29%	-0.03%	-0.10%
Freedom	40.68	40.80	40.85	40.87	0.47%	-0.06%	-0.19%
Haunted	42.24	42.32	42.35	42.36	0.26%	-0.02%	-0.08%
Average - 720p	40.56	40.58	40.69	40.70	0.34%	-0.02%	-0.30%

- 16 is best
 - Miniscule difference between 16 and 10 (.02%)
 - .3% delta between 5 and 16

How Much Time?

Encoding Time	1 Ref	5 Ref	10 Ref	16 Ref	Max Delta	10 - 16 Delta	16 - 5 Delta
Tears of Steel	39	49	72	91	133%	-21%	-46%
Sintel	40	53	71	76	90%	-7%	-30%
Big Buck Bunny	41	53	68	85	107%	-20%	-38%
Talking Head	37	47	61	77	108%	-21%	-39%
Freedom	99	142	200	263	166%	-24%	-46%
Haunted	47	65	93	123	162%	-24%	-47%
Average - 720p	51	68	94	119	136%	-21%	-43%

- 16 is ~ 2.5 x longer than 1 reference frame
 - Cutting to 5 reduces encoding time by 43% (close to doubling capacity)
 - Reduces quality by .3%

Reference Frames

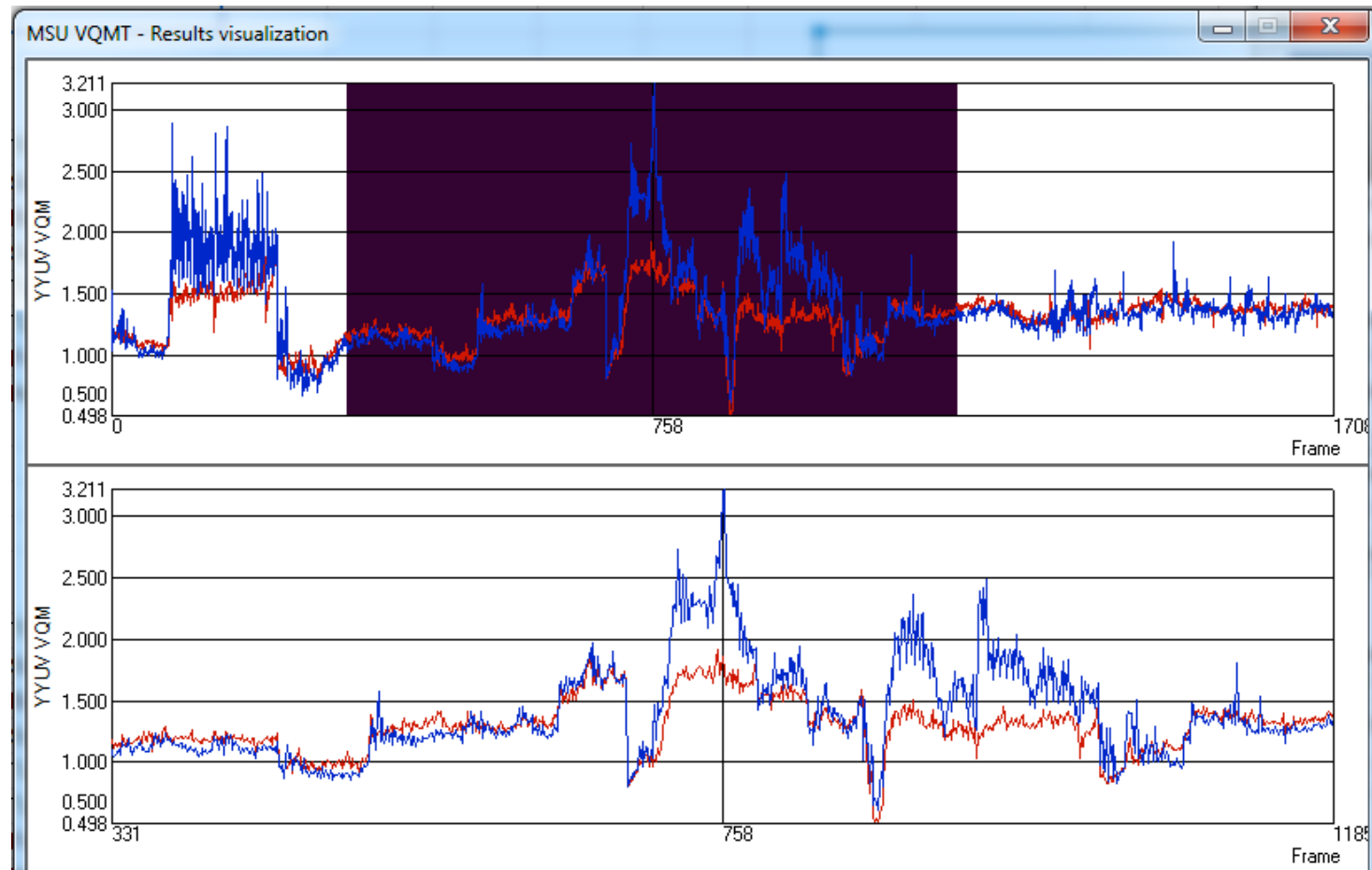
- Recommend 5 as best blend of performance and quality
 - Can increase encoding capacity by ~40% over 16 with no discernable impact on quality

VBR or CBR?

	200% VBR	150% VBR	125% VBR	CBR 2Pass	CBR 1Pass	Total Quality Delta
TOS	1.278	1.278	1.297	1.379	1.507	-18%
Sintel	1.211	1.212	1.209	1.306	1.439	-19%
Big Buick Bunny	0.994	0.995	0.996	1.073	1.164	-17%
Screencam	0.480	0.485	0.501	0.654	0.696	-45%
Tutorial	0.845	0.845	0.845	0.869	0.850	-1%
Talking Head	0.561	0.562	0.561	0.582	0.621	-11%
Freedom	1.620	1.618	1.621	1.639	1.682	-4%
Haunted	1.669	1.665	1.667	1.676	1.710	-2%

- Encode using 200%, 150, and 125% constrained VBR; 1 & 2 pass CBR
- Measure quality with VQM
- Green is best, red worst
- It gets even worse

Some Files will Show Quality Glitches



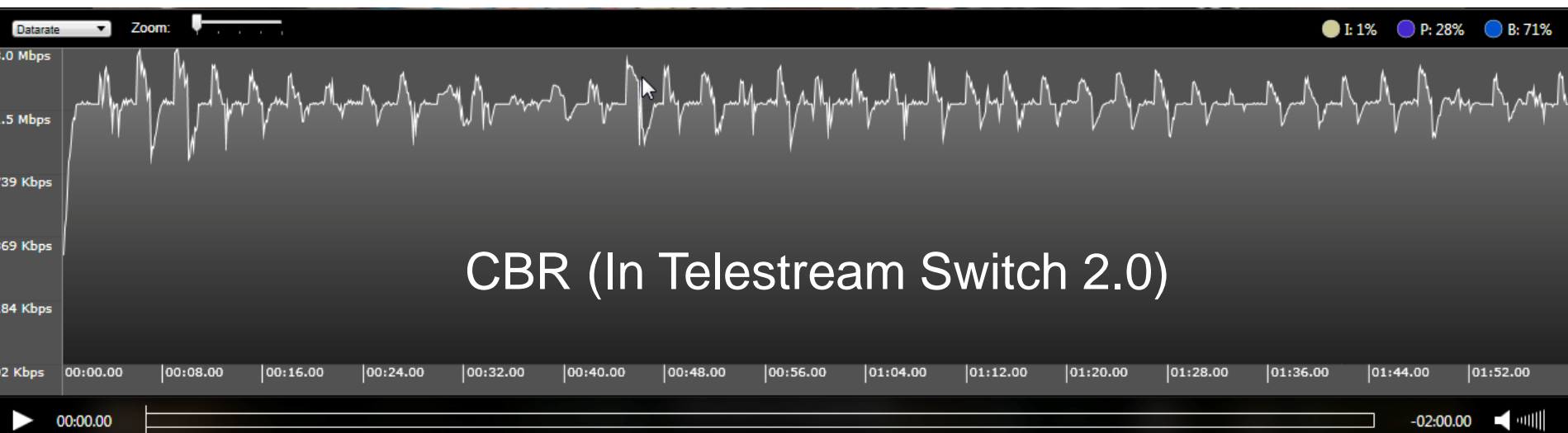
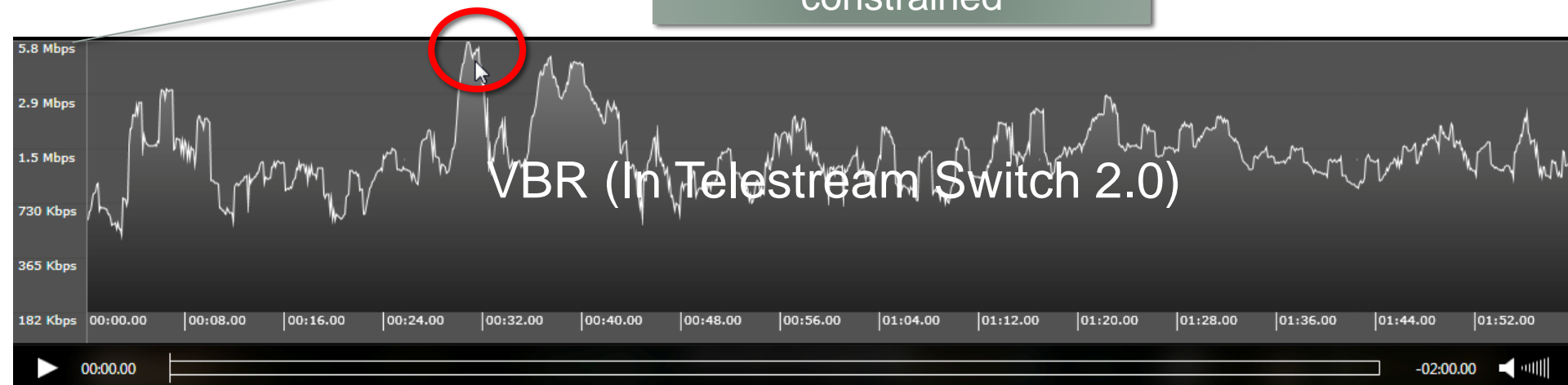
Files very close most of the time with notable exceptions

Transient Quality Issues



Definitely Can Be Smoothness Issues

Supposed to be 125%
constrained



CBR vs VBR

- Big issue:
 - Overall quality
 - Transient quality
- Deliverability is a huge issue with VBR
 - http://bit.ly/VBR_CBR_QOE
- I recommend 110% constrained VBR; best blend of *quality* and *deliverability*

Building Your Encoding Ladder

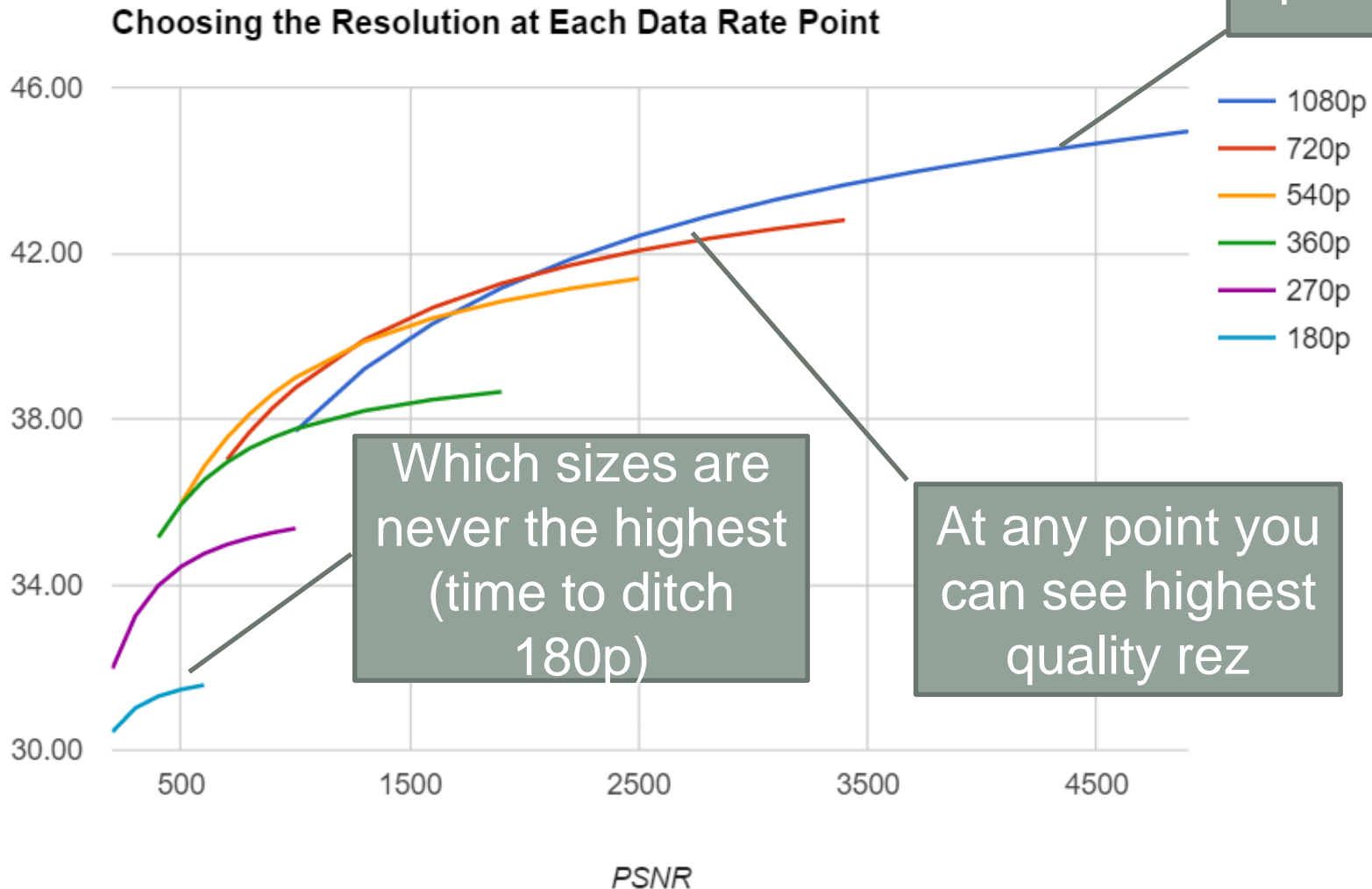
- Step 1: Choose lowest rate for mobile 200 kbps
- Step 2: Choose highest supported data rate (cost issue) 500 kbps
- Step 3: Choose data rate around 3 mbps (highest sustainable) 1000 kbps
- Step 4: fill in the blanks (between 150/200% apart) 1600 kbps
- 2100 kbps
- 3100 kbps
- 4600 kbps

Then Question is:

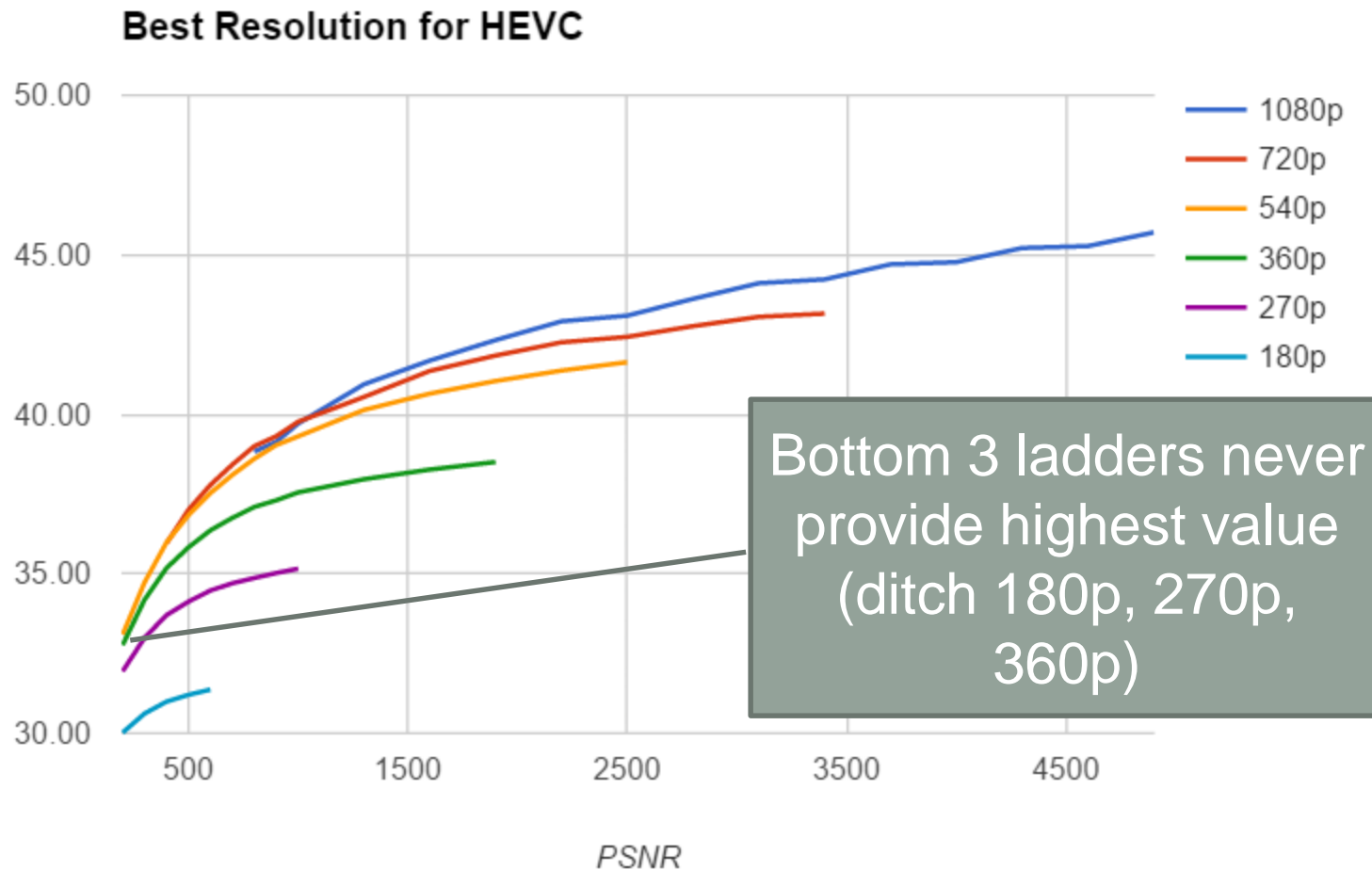
- Best resolution at each data rate
- Similar to per-title approach used by Netflix

PSNR	1080p	720p	540p	360p	270p	180p
4900	44.94					
4600	44.73					
4300	44.50					
4000	44.24					
3700	43.96					
3400	43.65	42.80				
3100	43.30	42.59				
2800	42.89	42.35				
2500	42.42	42.07	41.39			
2200	41.85	41.71	41.15			
1900	41.16	41.27	40.84	38.65		
1600	40.30	40.69	40.43	38.47		
1300	39.20	39.91	39.87	38.20		
1000	37.70	38.75	39.00	37.76	35.35	
900		38.27	38.60	37.55	35.25	
800		37.69	38.12	37.29	35.13	
700		37.01	37.54	36.95	34.97	
600			36.85	36.52	34.74	31.57
500			35.97	35.93	34.43	31.47
400				35.14	33.97	31.30
300					33.24	31.02
200					31.97	30.44

Choosing the Best Resolution

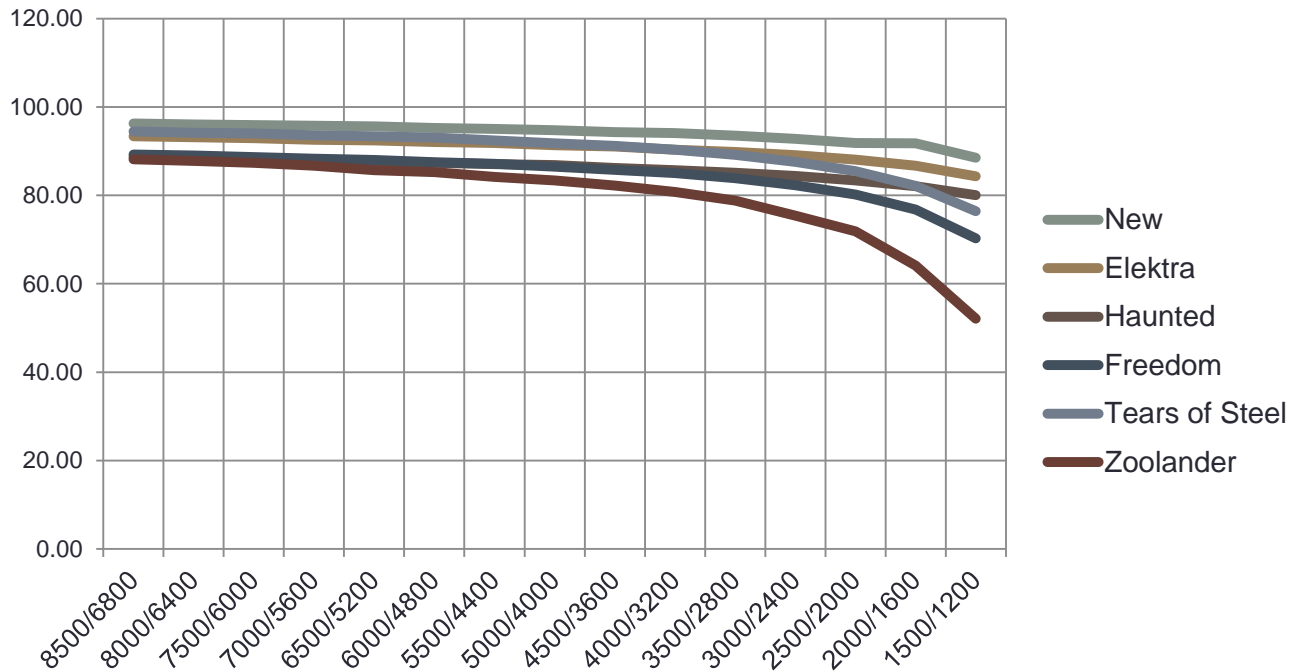


Choosing the Best Resolution HEVC



How Low Can You Go?

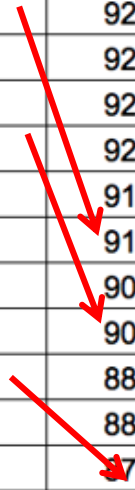
SQM Scores by Data Rate for Real World Content



- SQM – Higher is better
- Here we see Zoolander drop below 80 right around 4 mbps
- Others stay in excellent range throughout

What About Animation?

VQM	SQM	
	Real World	Animated
8500/6800	91.71	92.84
8000/6400	91.48	92.59
7500/6000	91.19	92.38
7000/5600	90.84	92.19
6500/5200	90.49	92.06
6000/4800	90.10	91.68
5500/4400	89.63	91.40
5000/4000	89.12	90.88
4500/3600	88.49	90.33
4000/3200	87.72	88.77
3500/2800	86.74	88.83
3000/2400	85.28	87.62
2500/2000	83.50	85.92
2000/1600	80.62	83.26



SQM Level	Real World Data Rate	Animated Data Rate	Delta
91.71/91.68	8500	6000	2500
90.84/90.88	7000	5000	2000
90.10/90.33	6000	4500	1500
87.72/87.62	4000	3000	1000

- Animated scores achieved similar quality levels to real world at much lower data rates
- Should be able to produce the same quality on animated content at a much lower data rate

To Run These Tests

Overall Performance			
Analysis	Z840	Z800	% Decrease
Convert to YUV	56	367	85%
MSU VQMT	860	1,701	49%

- Computer/disk speed matters
- Use the fastest computer you have
- Use an SSD drive if at all possible
- HP Z840 have been awesome for me

Questions?

- Questions