Different quality metrics analyzed in the context of diverse sequences types

Wojciech Szmyd, Paweł Szulc, Lucjan Janowski

Goal of the project

- Analyse how video quality assessment metrics behave throughout diverse sequences compressed using different codecs and parameters
- Better understand non-reference metrics
- Reduce data centers storage and bandwidth usage
- **Save billions of $$$**

Video sequences

- Netflix *Chimera* raw videos (700 GB total)
- Downloaded from *cdvl.org*, uploaded to Zeus supercomputer and preprocessed:
  - Chrominance downsampling to 4:2:0
  - Cut scenes to single shots
  - Cropping from 4096x2160 to 3840x2160 to get aspect ratio 16:9
  - Changing resolution from 3840x2160 to 1920x1080

Compression

Resolution\Codec	H264	H265
1920x1080	1, 2, 4, 8	0.5, 1, 2, 4
1280x720	0.5, 1, 2, 4	0.25, 0.5, 1, 2
854x480	0.25, 0.5, 1, 2	0.125, 0.25, 0.5, 1

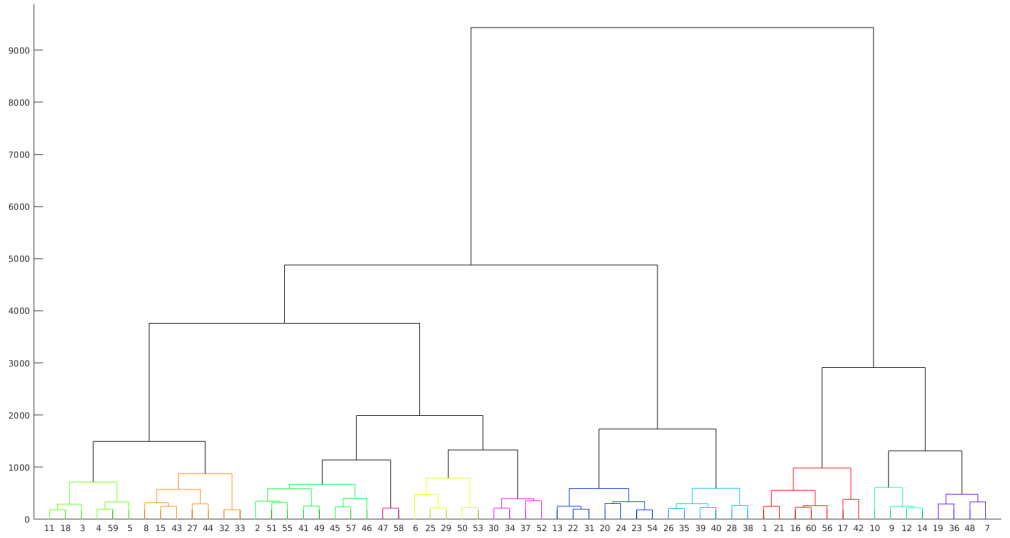
+ I frame interval: 8, 16, 64

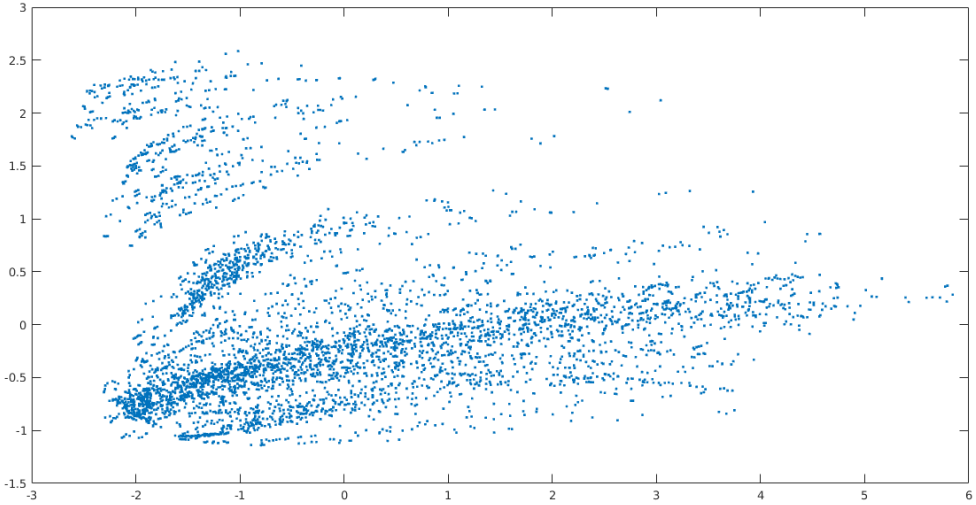
3*2*4*3*109 (number of shots) = 7848 videos

Metrics analysis

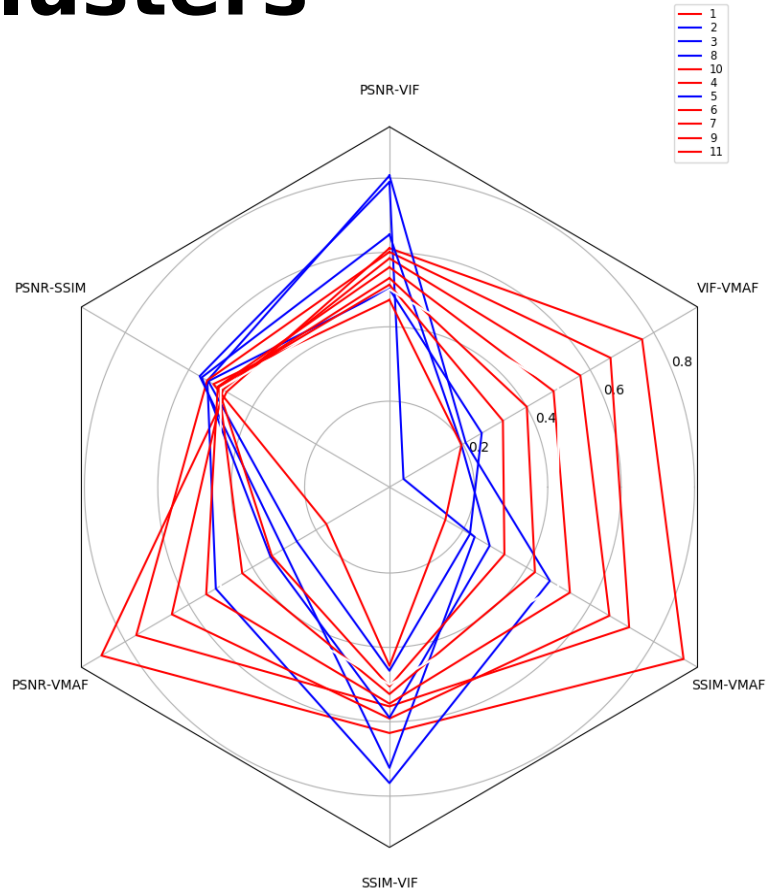
- Calculating FR metrics (VMAF, VIF, PSNR, SSIM) on upscaled decompressed videos and NR metrics (Temporal Activity, Spatial Activity, Blur, Brightness...) on original content
- Linearization of VIF, PSNR and SSIM with respect to MOS (VMAF considered as linear) and normalization
- Calculating differences between every pair of metrics for each sequence and clustering them in this six dimensional space
- Analysis of NR metrics and compression parameters in FR-defined clusters

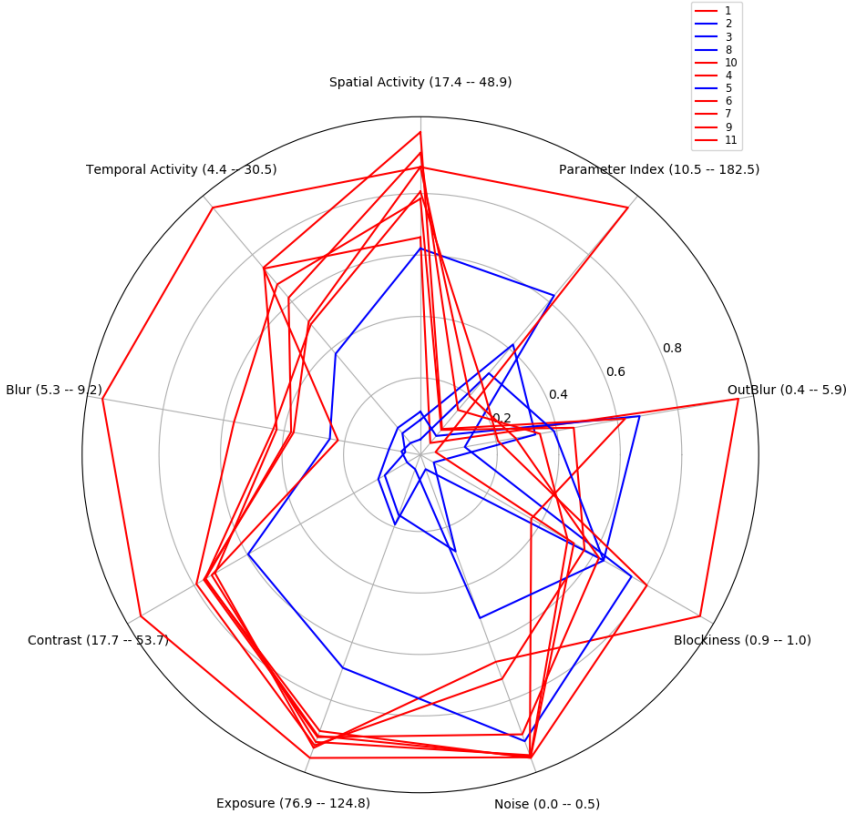
Clusters


9000
8000
7000
6000
5000
4000
3000
2000
1000
0
11 18 3 4 59 5 8 15 43 27 44 32 33 2 51 55 41 49 45 57 46 47 58 6 25 29 50 53 30 34 37 52 13 22 31 20 24 23 54 26 35 39 40 28 38 1 21 16 60 56 17 42 10 9 12 14 19 36 48 7

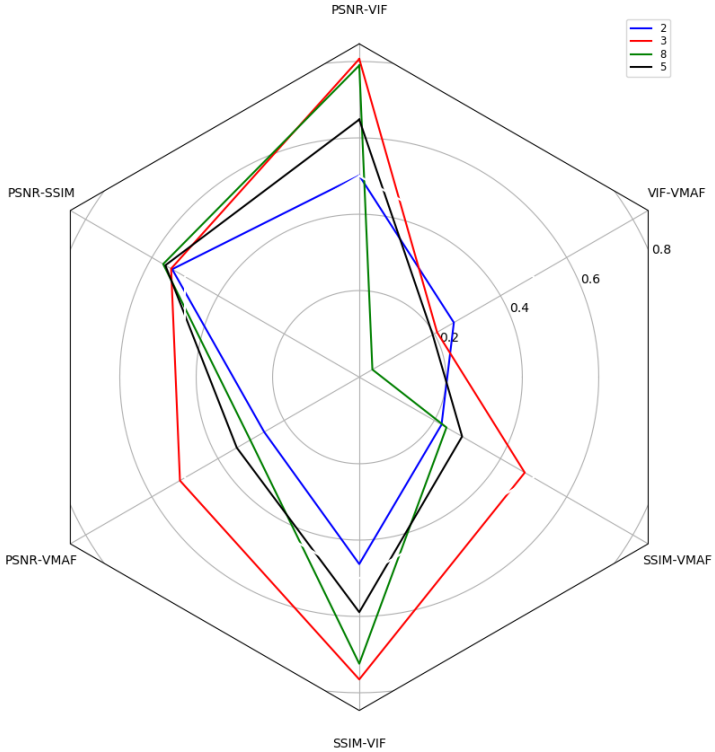

3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-3 -2 -1 0 1 2 3 4 5 6

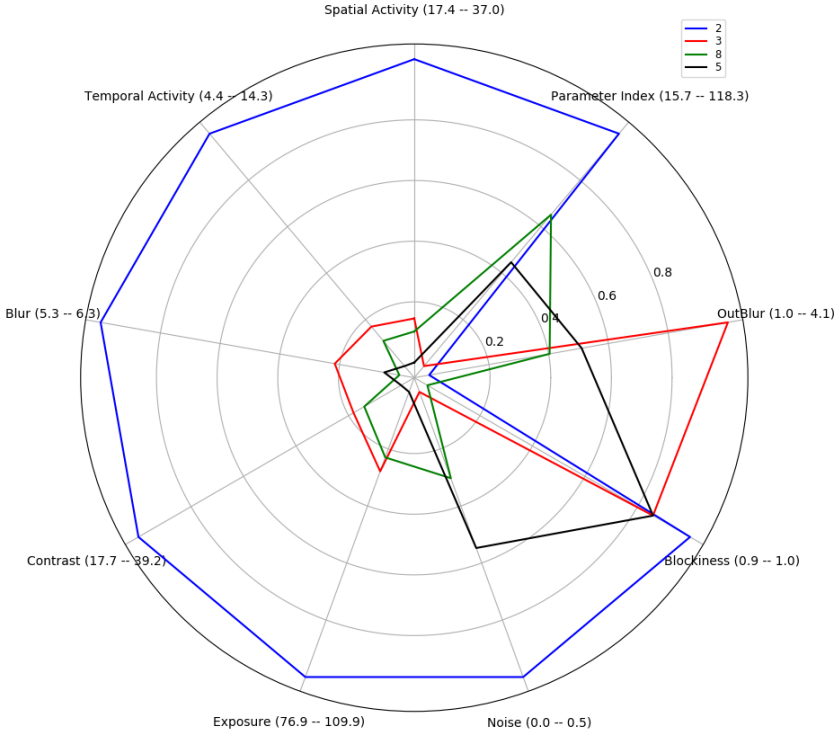
Clusters


1
2
3
8
10
4
5
6
7
9
11
PSNR-VIF
VIF-VMAF
SSIM-VMAF
SSIM-VIF
PSNR-VMAF
PSNR-SSIM
0.2
0.4
0.6
0.8


1
2
3
8
10
4
5
6
7
9
11
Spatial Activity (17.4 -- 48.9)
Parameter Index (10.5 -- 182.5)
OutBlur (0.4 -- 5.9)
Blockiness (0.9 -- 1.0)
Noise (0.0 -- 0.5)
Exposure (76.9 -- 124.8)
Contrast (17.7 -- 53.7)
Blur (5.3 -- 9.2)
Temporal Activity (4.4 -- 30.5)
0.2
0.4
0.6
0.8

Dark clusters





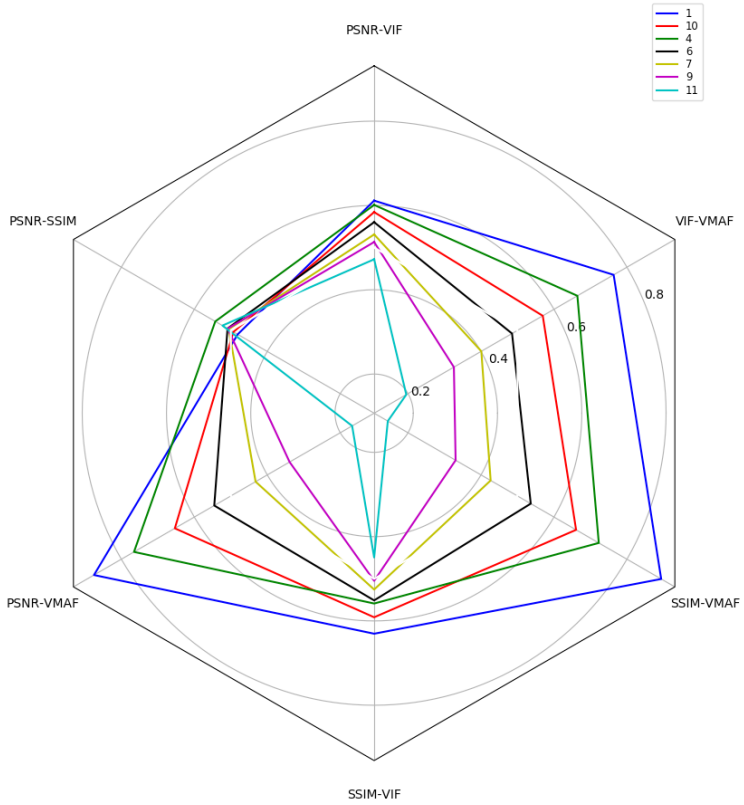
Dark clusters - cluster details

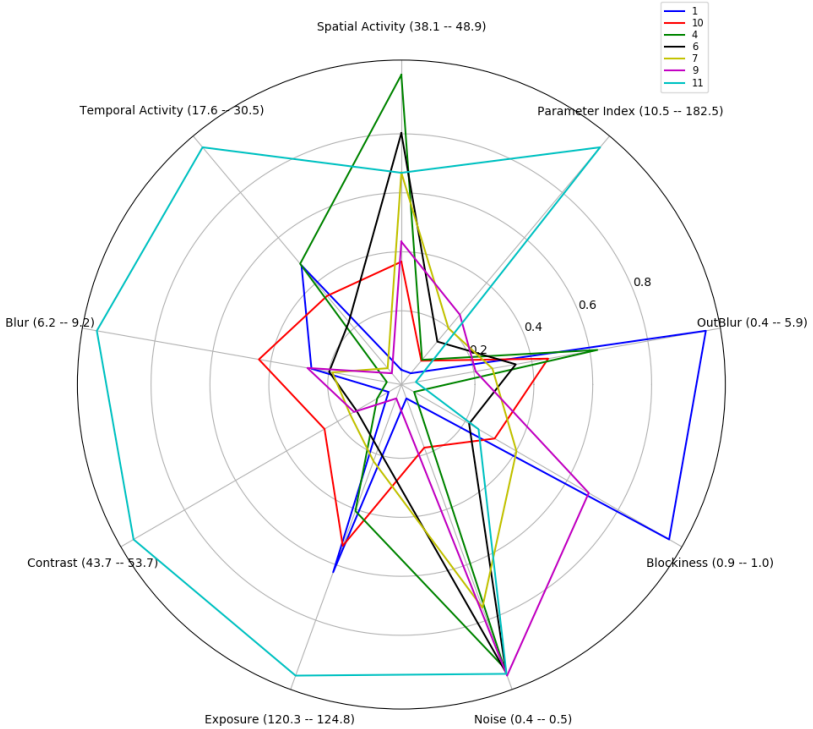
- Cluster #2 - huge cluster (20% of corpus), various sequences with mean exposure slightly below average, compressed with high bitrates and resolutions
- Cluster #3 - mostly consists of videos showing couple dancing on dark background, compressed with the lowest bitrates
- Cluster #5 - family eating and talking in dark room, average bitrates
- Cluster #8 - content similar to cluster #3, much higher bitrates and resolutions

Dark clusters - observations

- VMAF gives highest quality prediction, VIF - the lowest
- Subjective analysis of clusters #3 and #8 shows that VMAF is right there
- Cluster #2 - VIF results are closer to reality than in 3 and 8, still worse than VMAF

Bright clusters





Bright clusters - cluster details

- Most common videos in cluster #1 are bright and have high value of temporal activity: rollercoaster ride and basketball game in the daylight, very low bitrates and resolutions. Cluster #4 has similar content but much higher bitrates and resolutions.
- Clusters #6, #7 and #9 have a wide range of sequence types and compression parameters.
- Cluster #10 consists of videos with high spatial activity and blur
- Sequences in cluster #11 have the highest values of brightness, temporal activity, blur and compression parameters

Bright clusters - observations

- VMAF is very sensitive to compression parameters changes
- VIF is reluctant to give highest scores, for average videos it’s close to reality
- Cannot prove VMAF is wrong

Results

- No big difference between PSNR and SSIM
- There is a correlation between content type and metric accuracy
- All metrics behave similarly for sequences compressed with H264 and H265 codecs
- VIF is too strict for dark videos without many details compressed with higher bitrates

Thank you