Final Presentation

Data Analysis for Quality of Experience Assessments

Mert Kayhan
Burak Safak
Xingwei Qu
Sen Wang

Lehrstuhl für Datenverarbeitung

February 20, 2018

Table of contents

Introduction

- Big Buck Bunny, El Fuente 1, Tennis
- ACR-HR method was used
- 480p resolution, H.264 codec
- 28 people completed the test (5 were eliminated)
- PCR and PLSR are used to predict the MOS

Part A: Experimental Design

- Detailed introduction
- Training phase (Both ends of the scale are presented)
- Video Quality Assessment (Vertical discrete scale, 1 to 5, Next button bottom left hand side)

Experimental Design

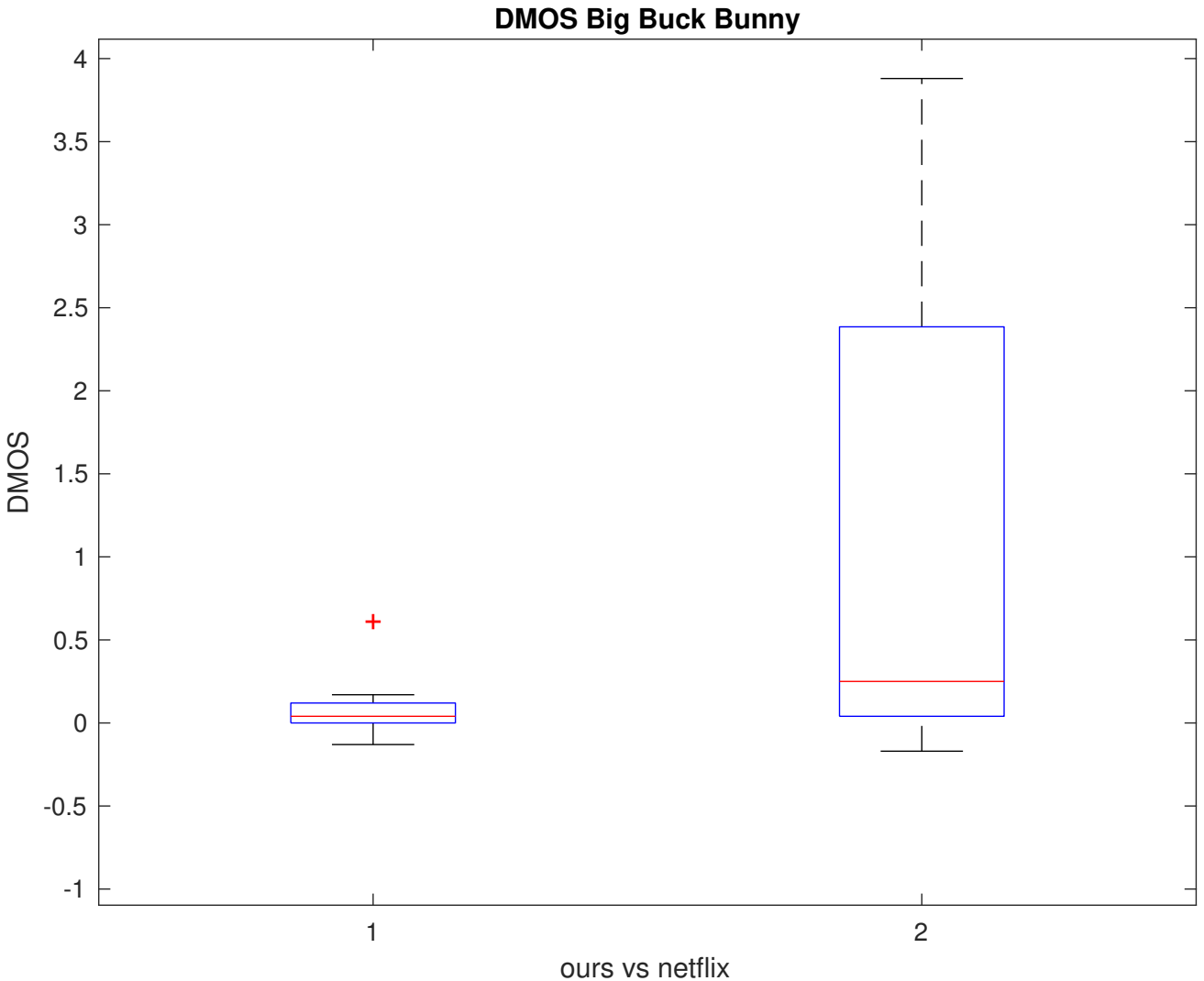
Please rate the "visual" video quality

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

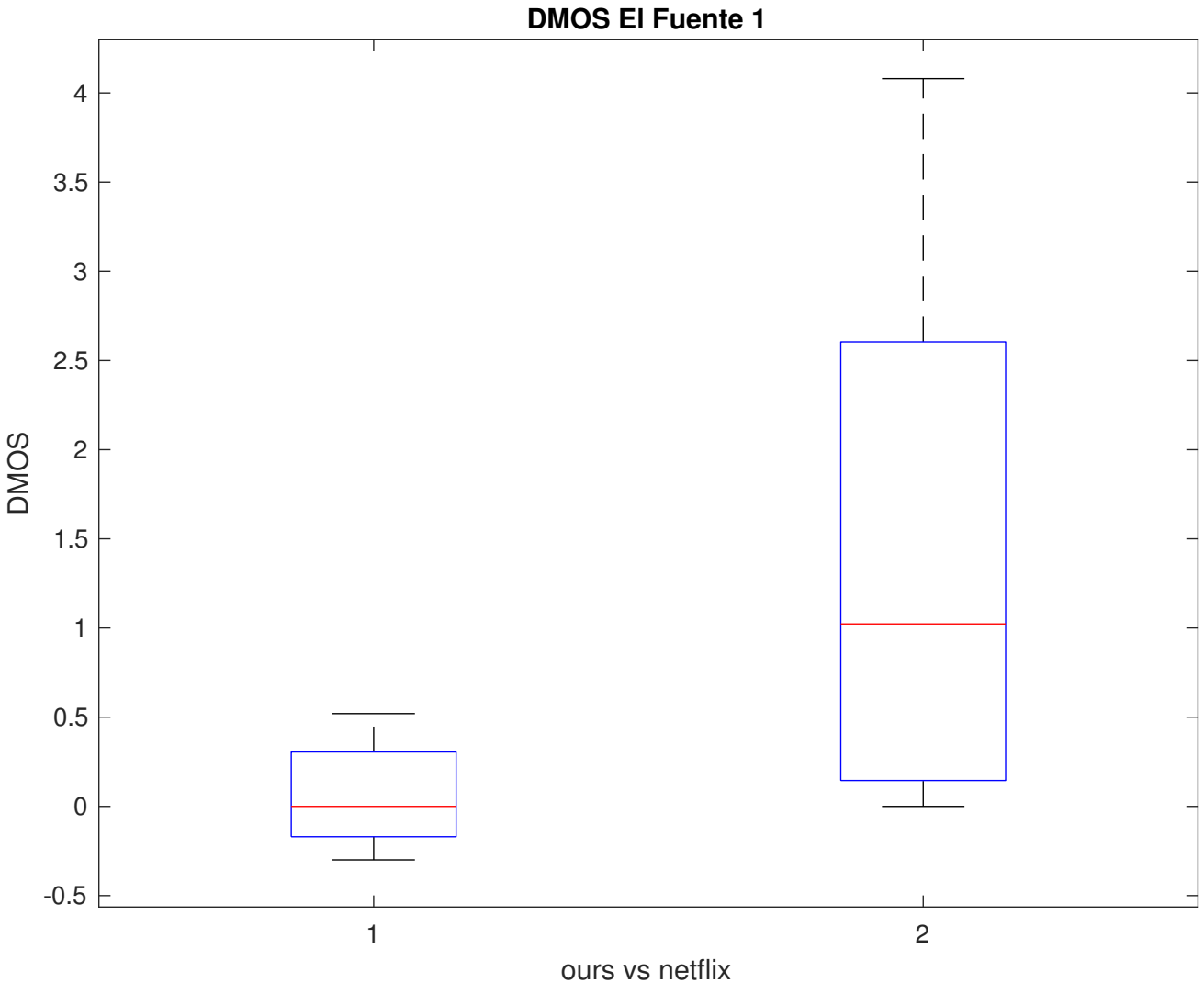
Experimental Design

- No repetitions of the videos were included. (Time concerns)
- Random block included. (To account for possible biases)
- No separate voting time (However, cannot click Next before voting)
- Crowdsourcing (Test environment)

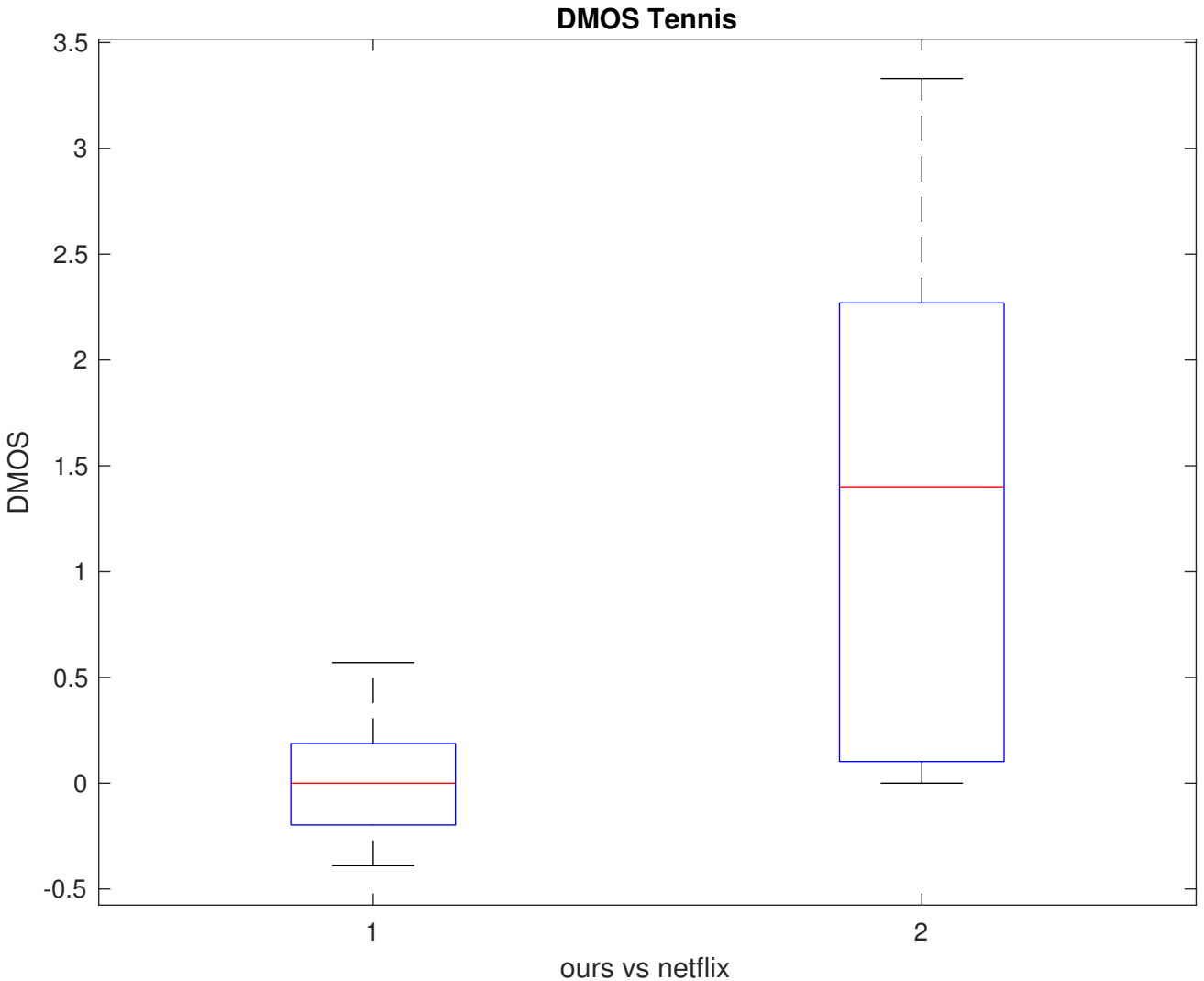
Results


DMOS Big Buck Bunny
DMOS
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
1
2
ours vs netflix

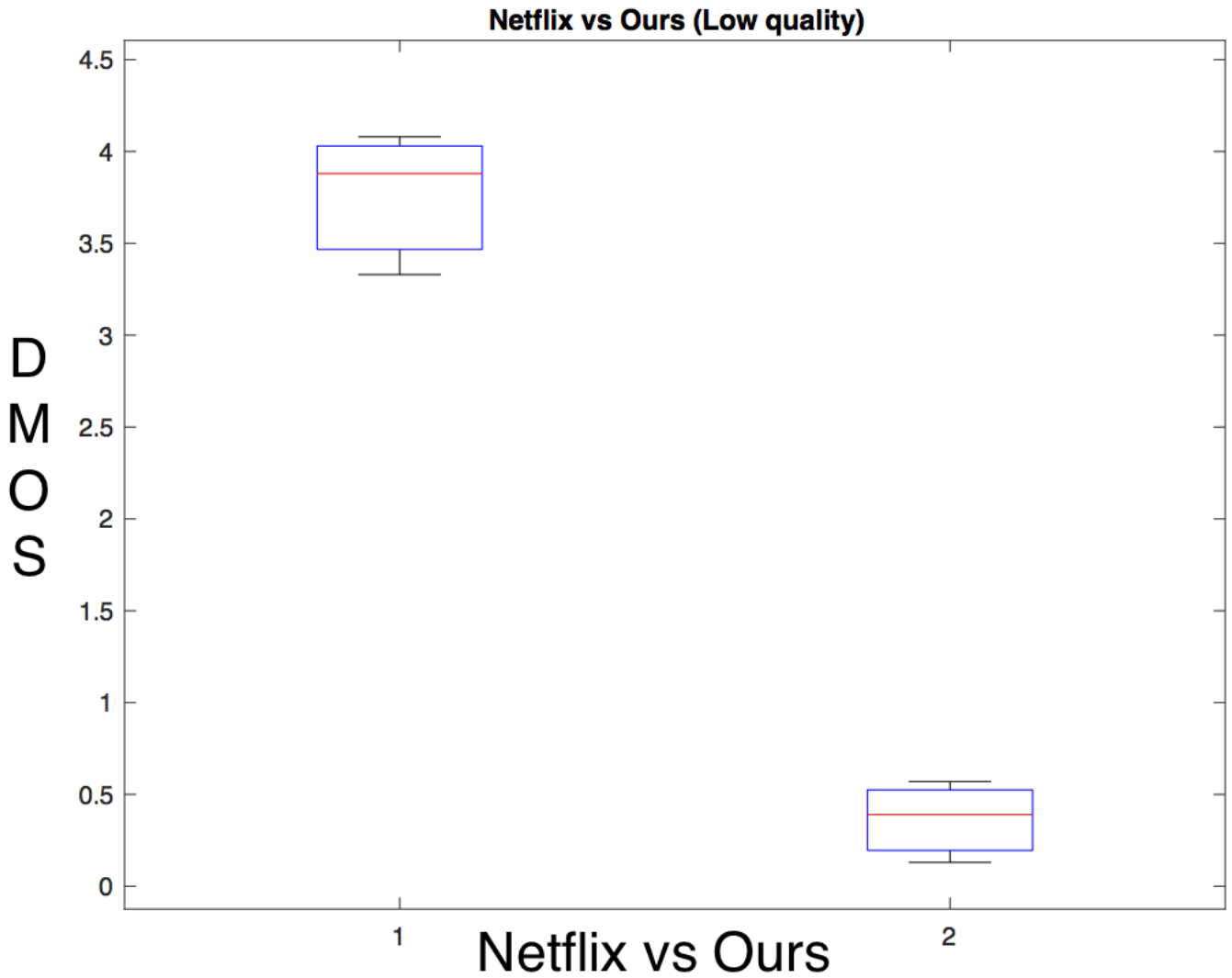
Results



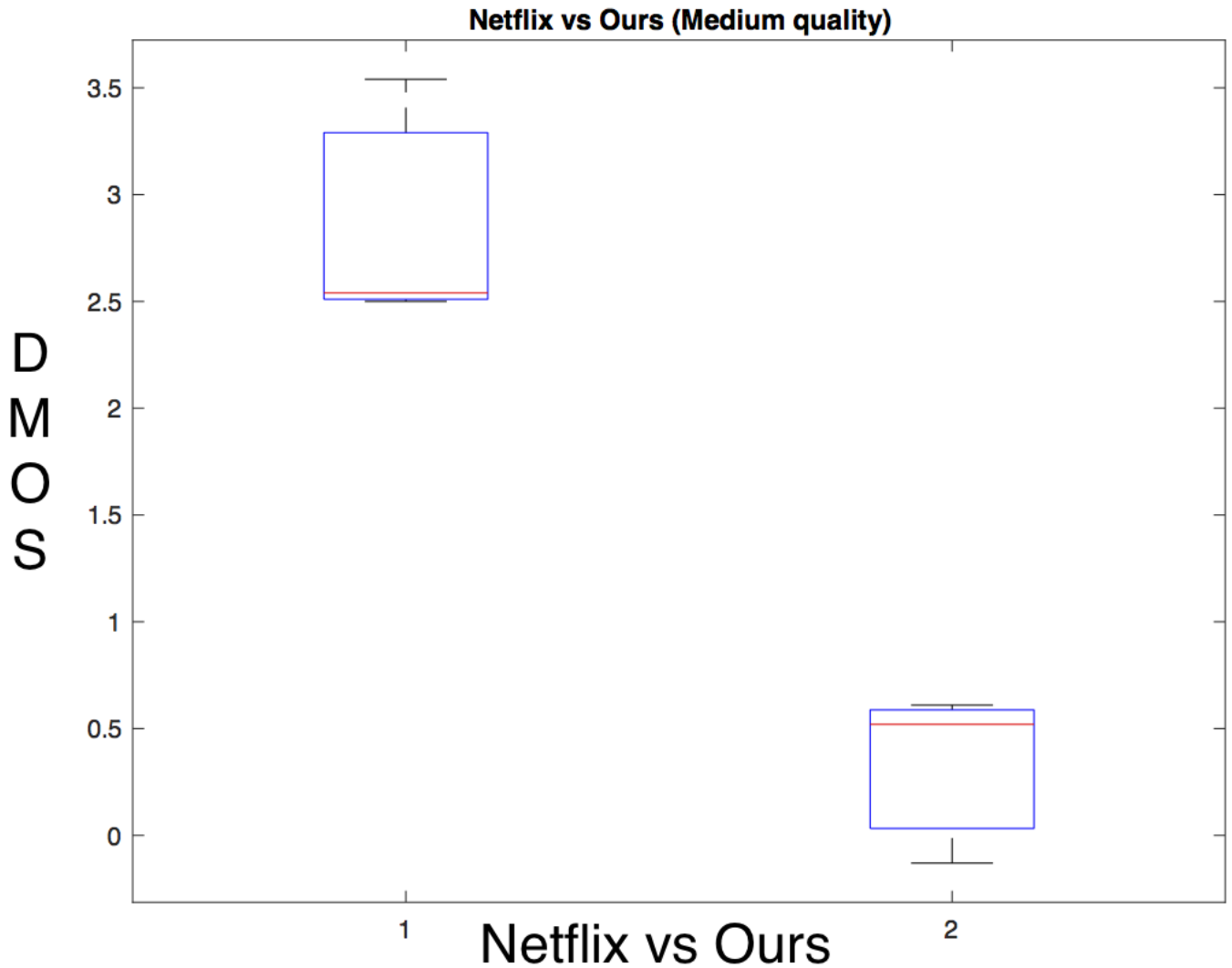
Results



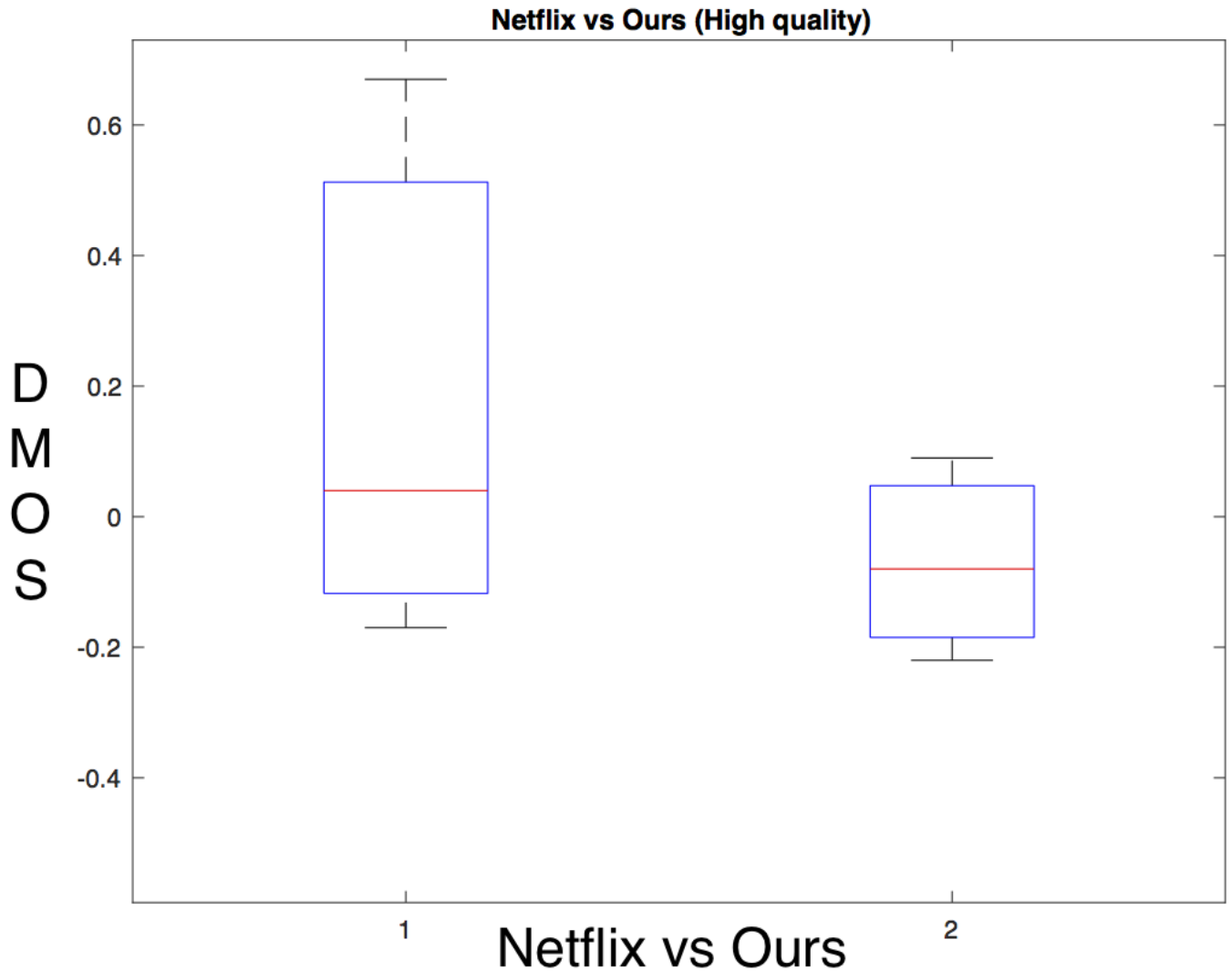
Results


Netflix vs Ours (Low quality)
DMOS
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
2
Netflix vs Ours

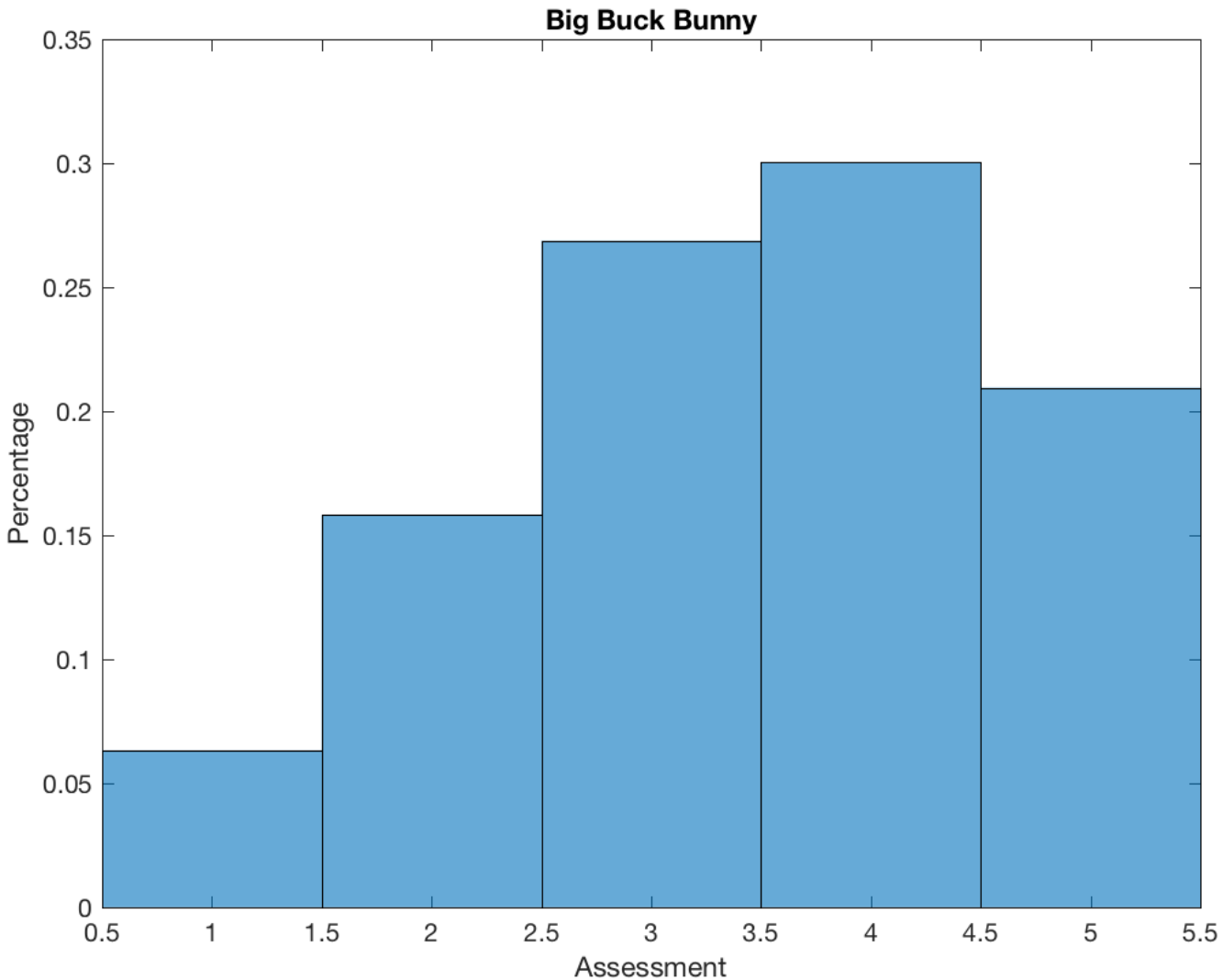
Results


Netflix vs Ours (Medium quality)
DMOS
3.5
3
2.5
2
1.5
1
0.5
0
1
2
Netflix vs Ours

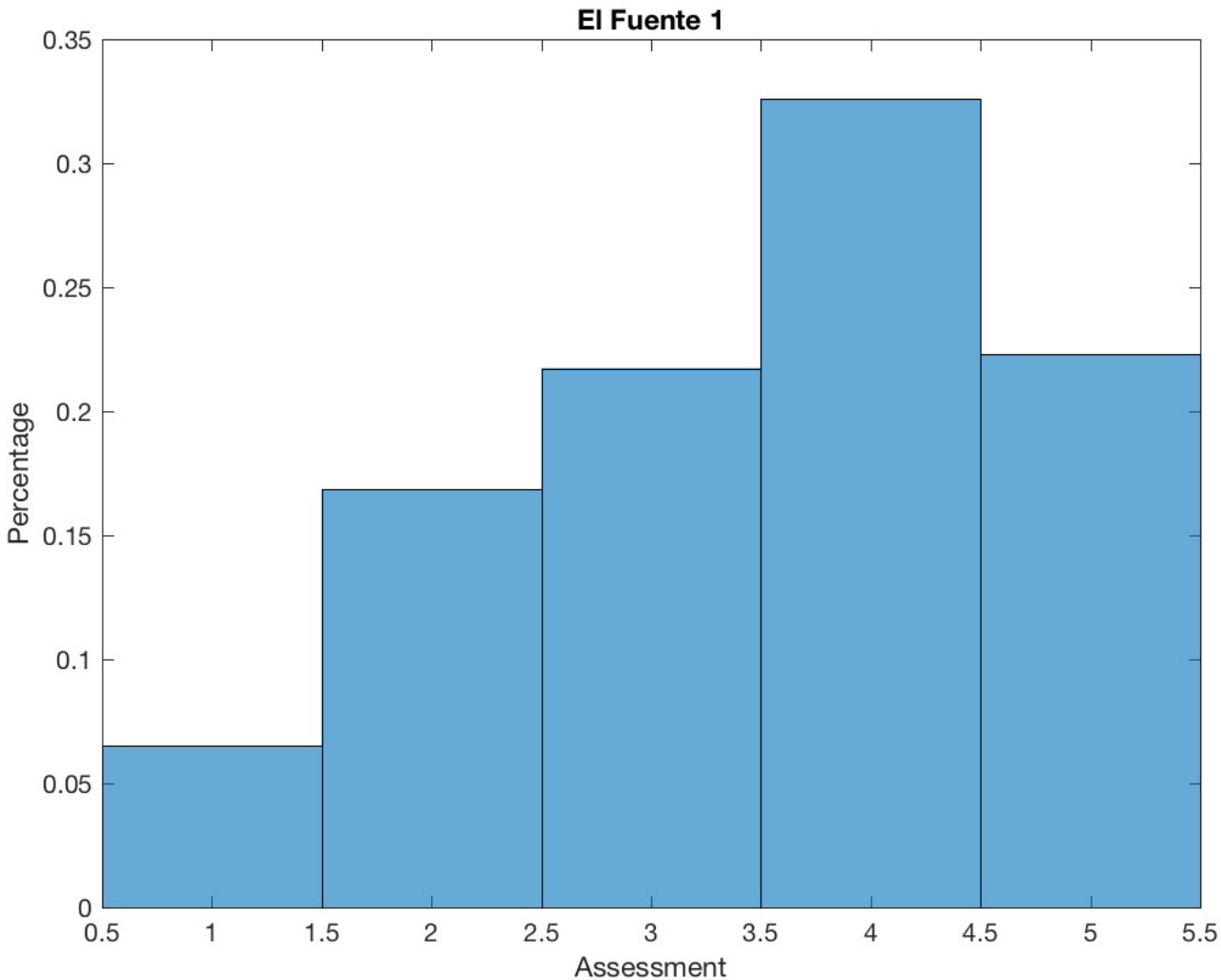
Results


Netflix vs Ours (High quality)
DMOS
0.6
0.4
0.2
0
-0.2
-0.4
1
2
Netflix vs Ours

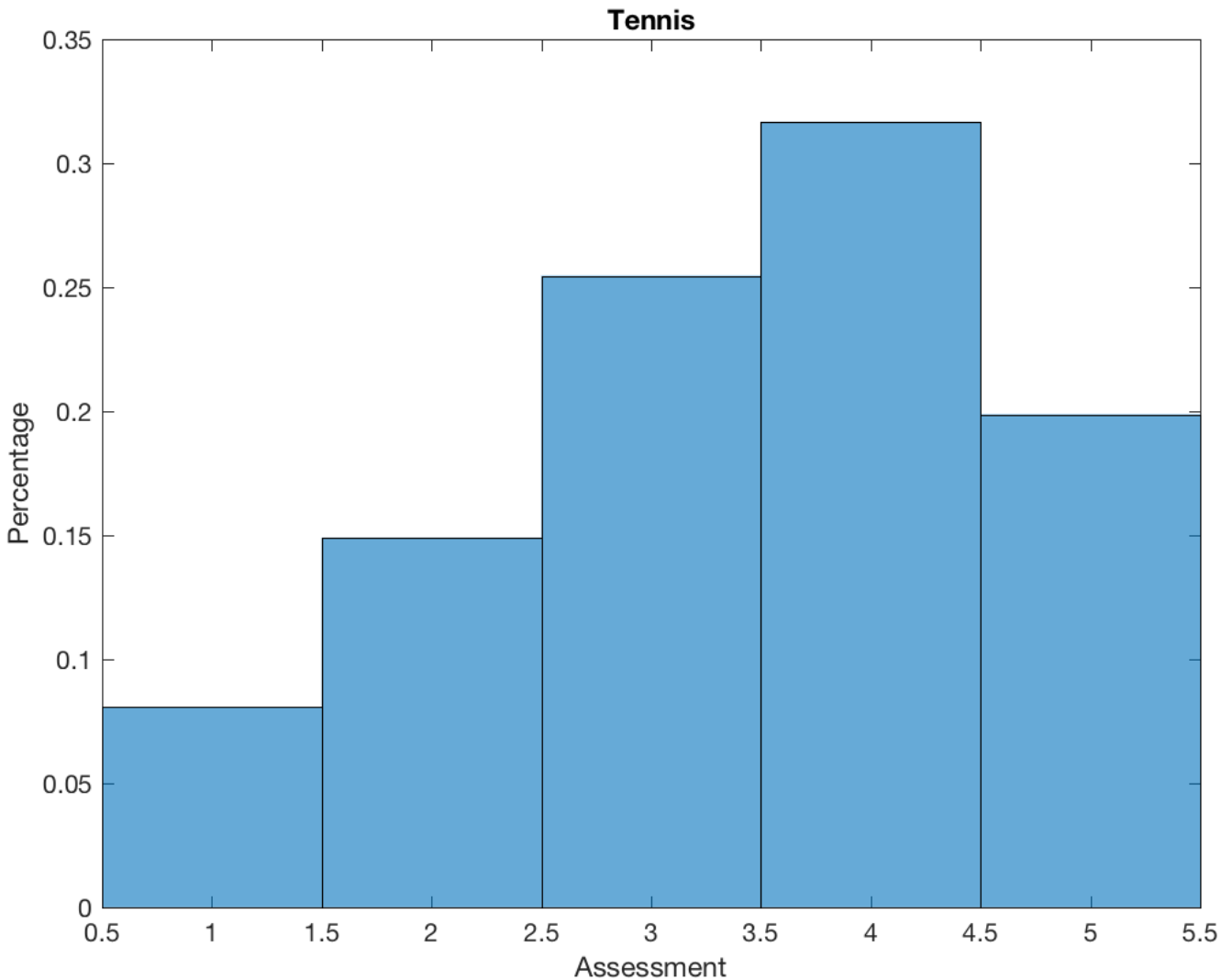
Results


Big Buck Bunny
Percentage
Assessment
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

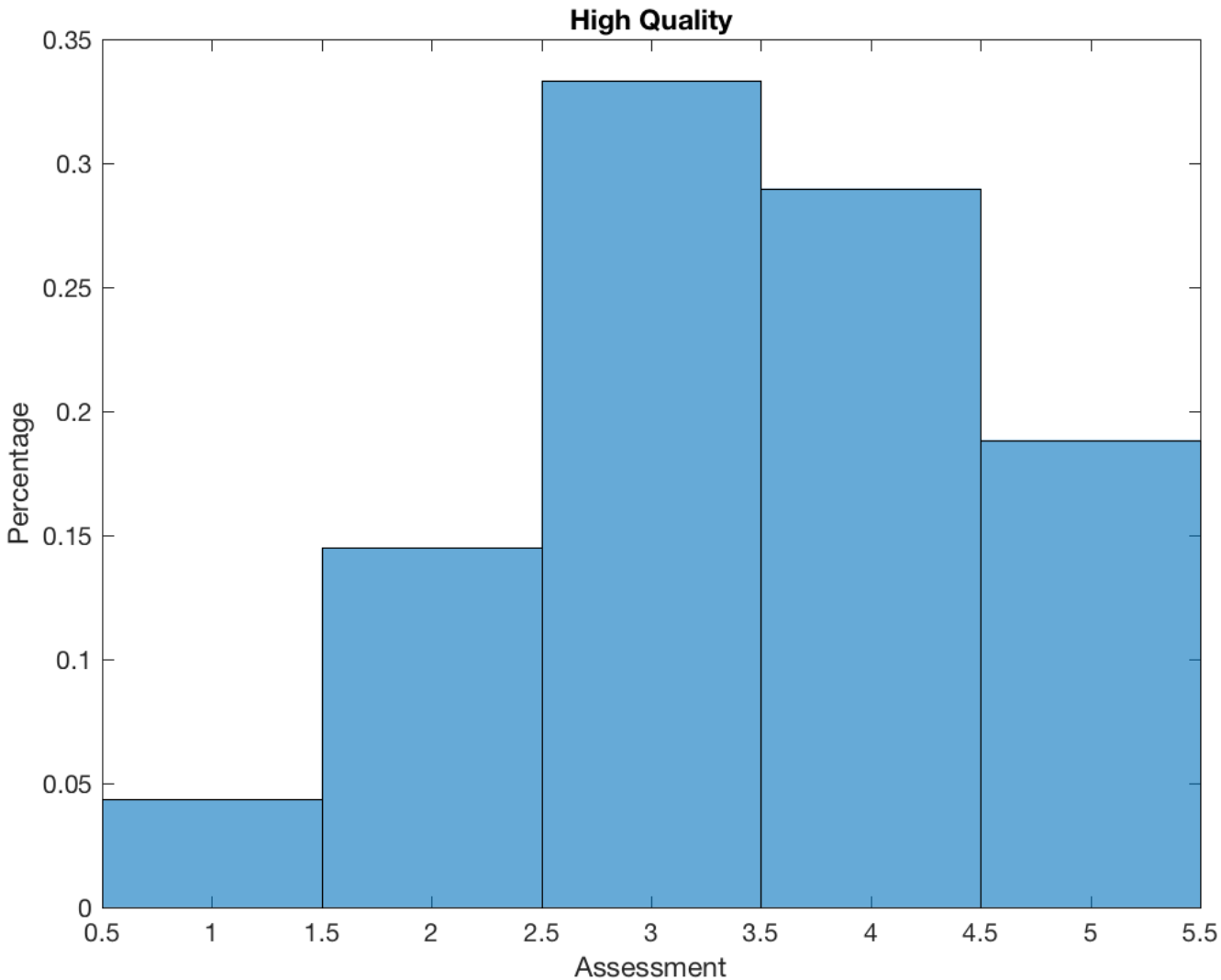
Results


El Fuente 1
Percentage
Assessment
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

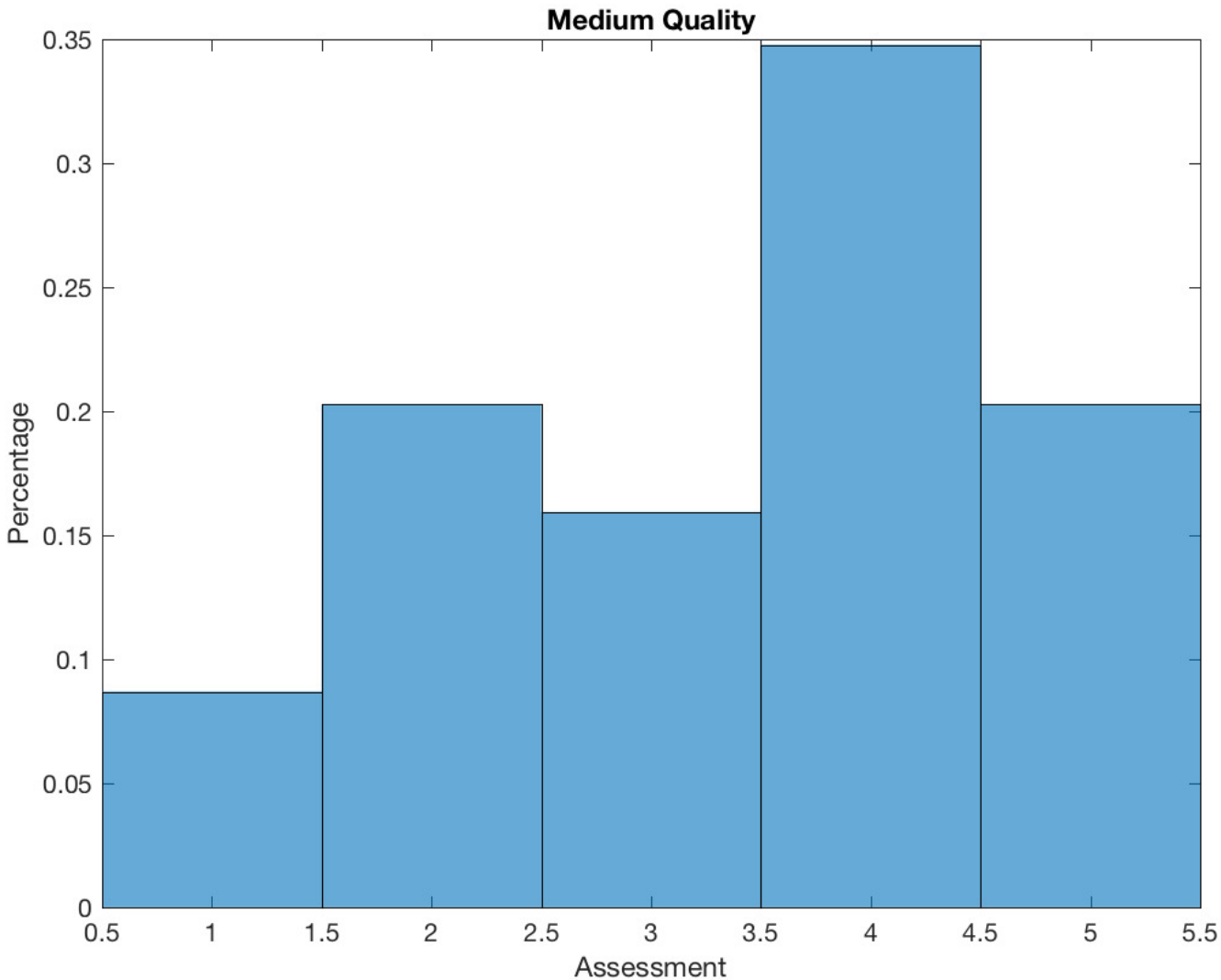
Results


Tennis
Percentage
Assessment
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

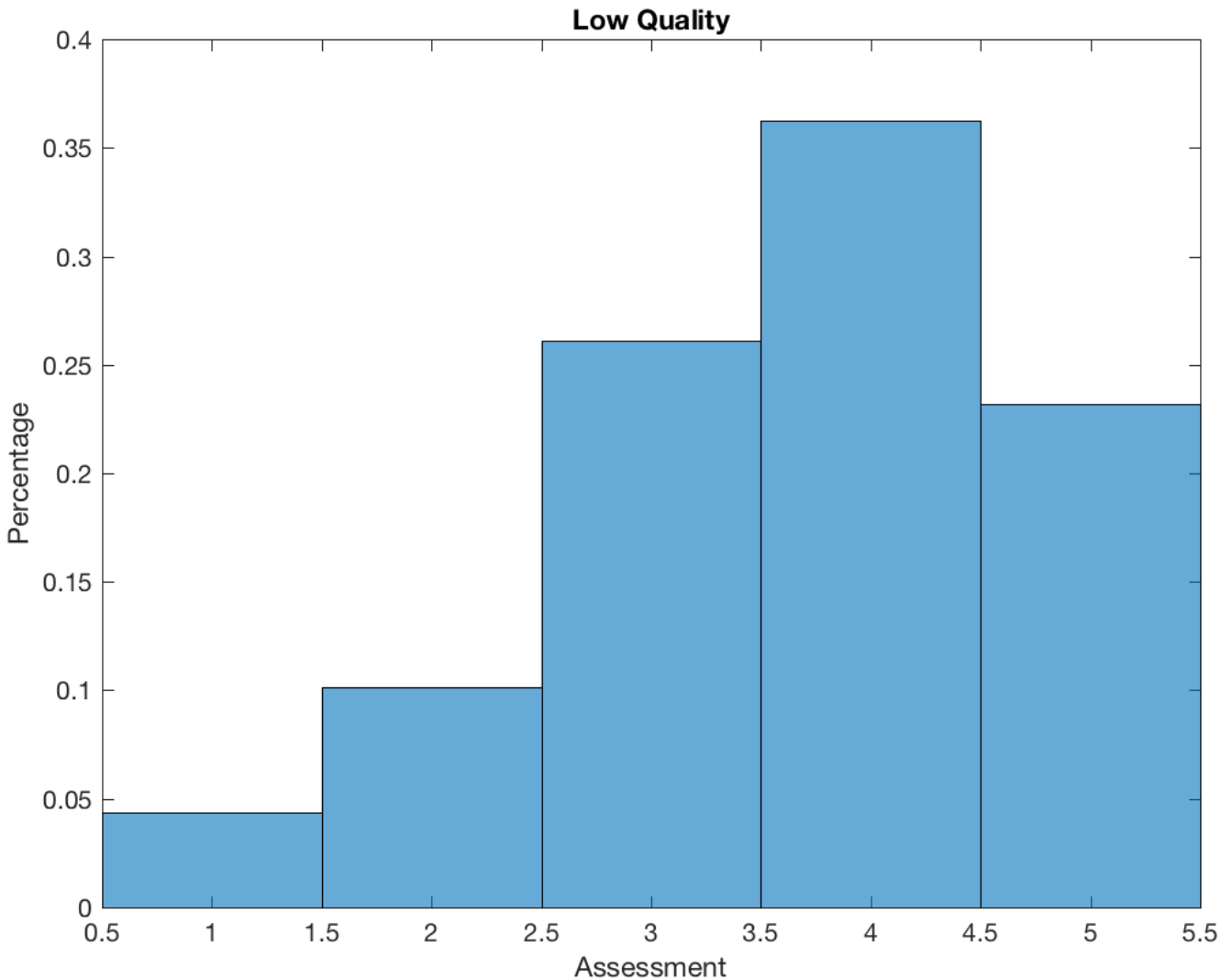
Results


High Quality
Percentage
Assessment
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

Results


Medium Quality
Percentage
Assessment
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

Results


Low Quality
Percentage
Assessment
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

Subjects

- From many countries and continents (Germany, Turkey, Italy, etc.)
- Mostly friends and family
- Average age around 20-25
- Mostly men

Discussion and Conclusion

- Random block was not very necessary
- More subjects to take the test
- Many subjects did not finish the test
- Two ends of the scale could have been presented in a better way
- Variation within subjects
- Better outlier detection

Differences with Netflix

- Double stimulus impairment scale
- 34 source clips / 300 distorted clips
- Consumer grade TV, controlled ambient lighting, living room-like environment
- No crowdsourcing!
- Larger budget!

Discussion and Conclusion

- Content makes a big difference in assessment (Big Buck Bunny)
- Subjects did not use the full scale
- Subjects were generally content with the video quality
- High quality videos as expected
- Surprising results especially in low quality videos
- It is difficult to motivate people without incentive
- Single stimulus methods are fast, but not reliable

Part B: Video Quality Metric

- Features Extraction
- Models selection
- Performance
- Discussion and Conclusion

Features Extraction

- **Features Extracted by Netflix**
  - $Vif - scale0, 1, 2, 3...$
  - Adm2 (*DLMandAIM*)
  - Motion2
- **Features Extracted by ourselves**
  - SSIM IW-SSIM MS-SSIM
  - PSNR

SSIM and PSNR

- **Structural similarity(SSIM)**
  - Luminance Comparision
  - Contrast Comparision
  - Structure Comparision
  - $SSIM = l(S, S')c(S, S')s(S, S')$
- **Peak signal-to-noise ratio(PSNR)**
  - $PSNR = 20log_{10}(MAX_I) - 10log_{10}(MSE)$
- For the YUV video SSIM are:
- $SSIM_{ij} = W_Y SSIM_{ij}^Y + W_U SSIM_{ij}^U + W_V SSIM_{ij}^V$
- $W_Y = 0.8 W_U = 0.1 W_V = 0.1$

Improved SSIM

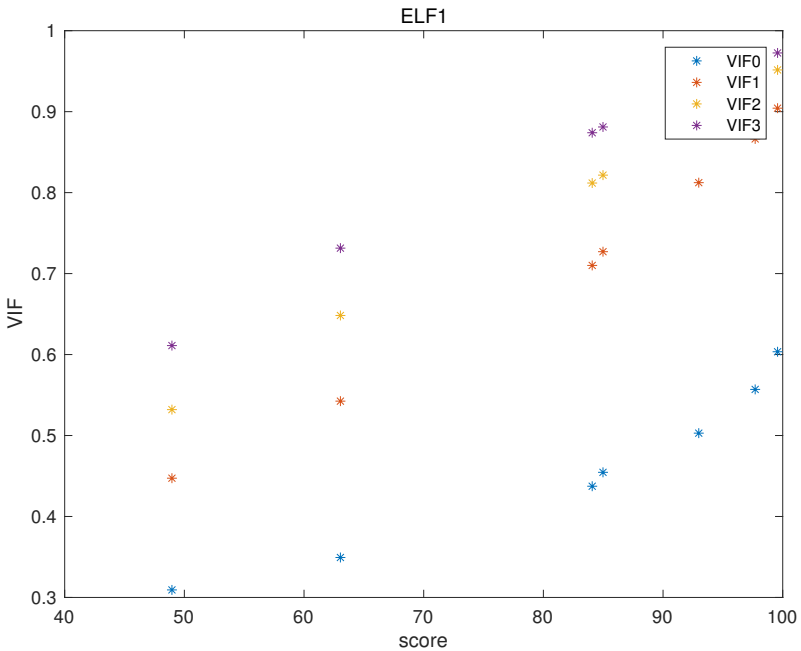
- **Information content weighted structural similarity(IW-SSIM)**
- incorporating the idea of information content weighted pooling.
- time costed

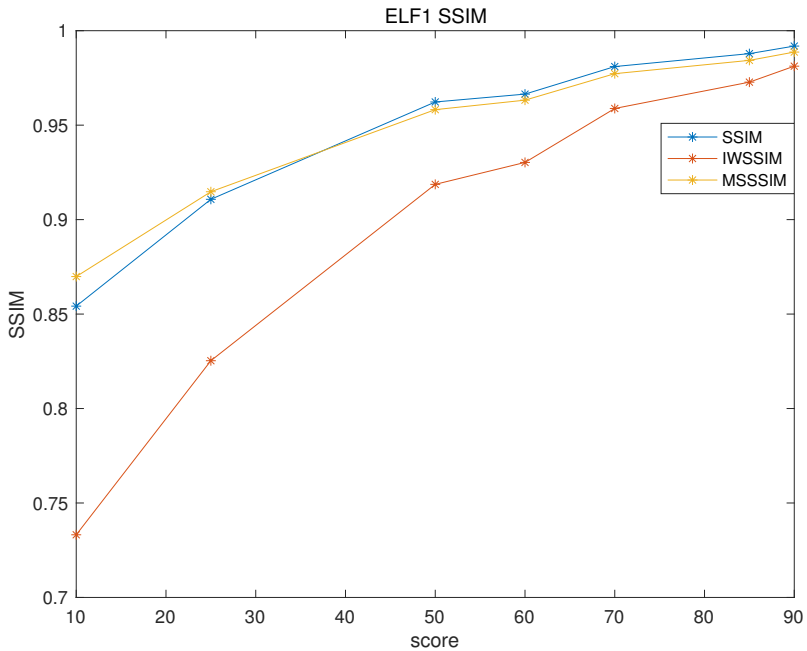
- **Multi-scale Structural Similarity(MS-SSIM)**
- supply more flexibility than single-scale methods in incorporating the variations of image resolution and viewing condition.
- Results is similar to the SSIM

Temporal pooling

- Pooling can be done using averaging over all frames
- In this section Mean pooling is better
- For other video pooling is a big challenge

Results


ELF1
VIF
score
VIF0
VIF1
VIF2
VIF3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
40
50
60
70
80
90
100


ELF1 SSIM
SSIM
score
SSIM
IWSSIM
MSSSIM
1
0.95
0.9
0.85
0.8
0.75
0.7
10
20
30
40
50
60
70
80
90

Regression Models

Principal Components Regression PCR

Only creates components to explain the observed variability in the features

Partial Least Squares Regression PLSR

Also takes the response variable into account, namely the MOS

Data preprocessing

- Removal of unreliable subjects:VQEG

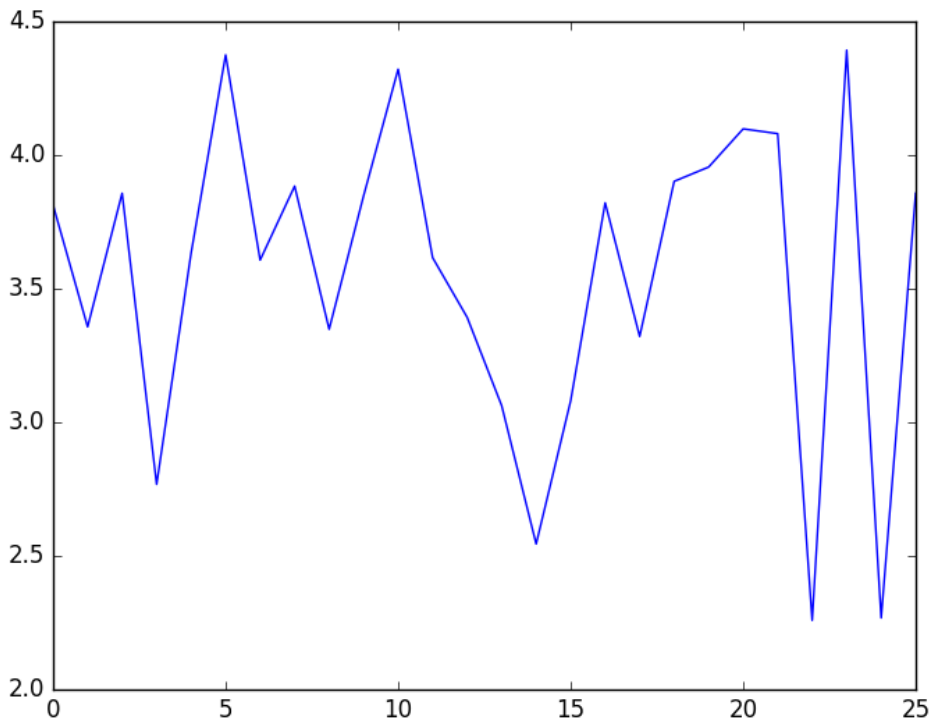


Figure: MOS from Crowdsourcing after 3 Iterations

Data preprocessing

- Normalization

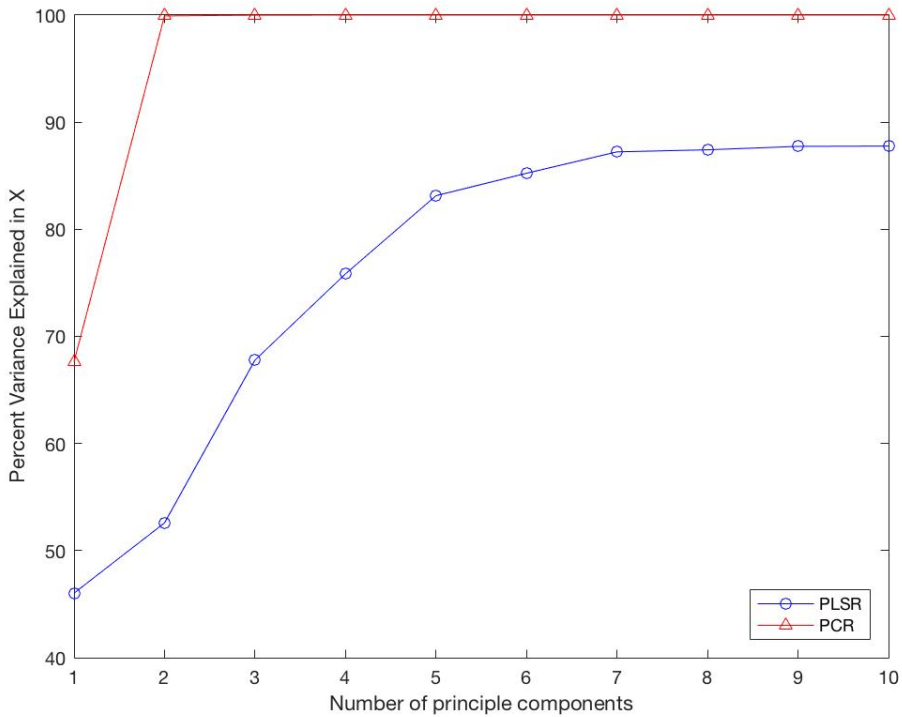


Figure: Percent Variance without normalization

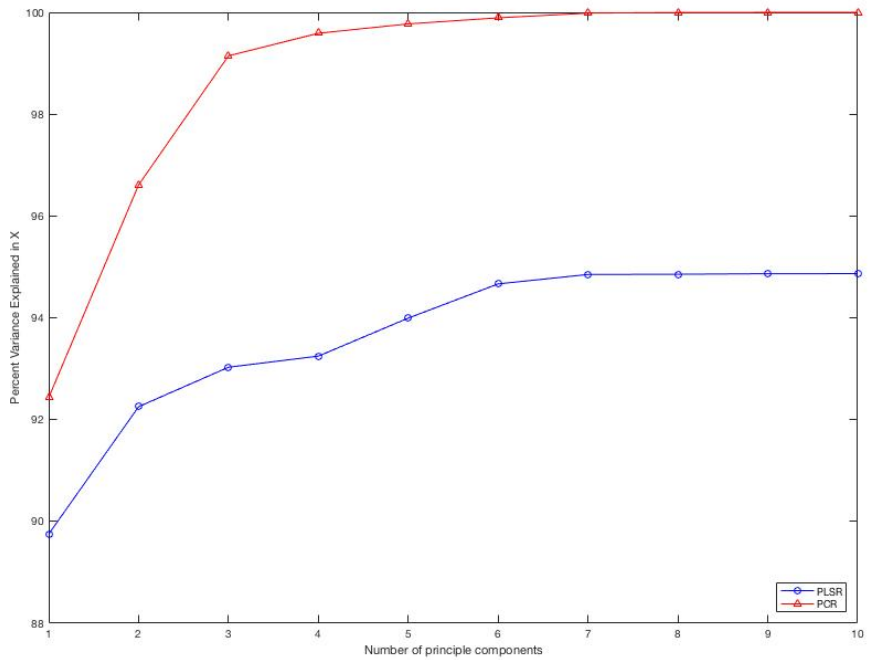


Figure: Percent Variance + psnr

Performance

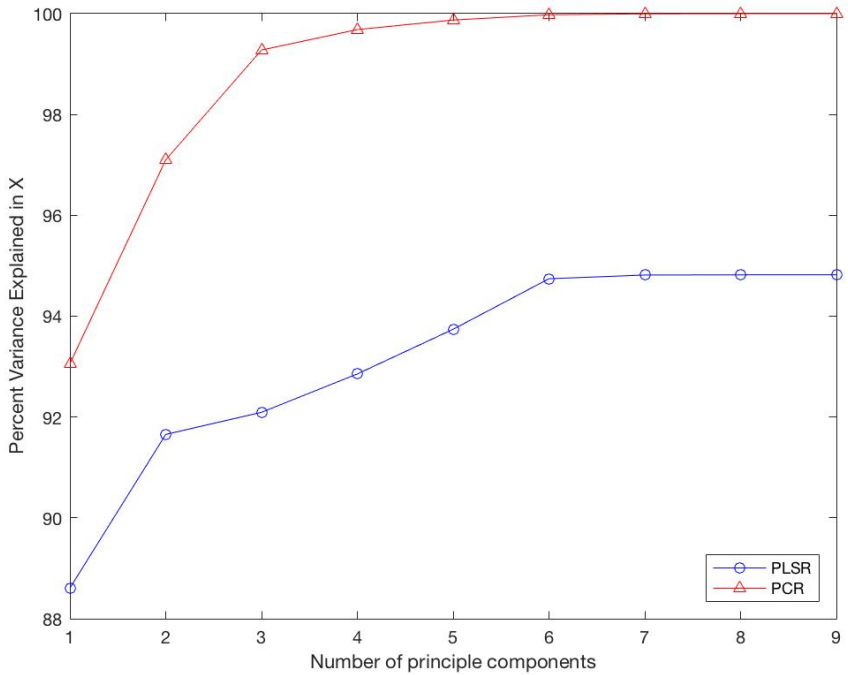


Figure: Percent Variance

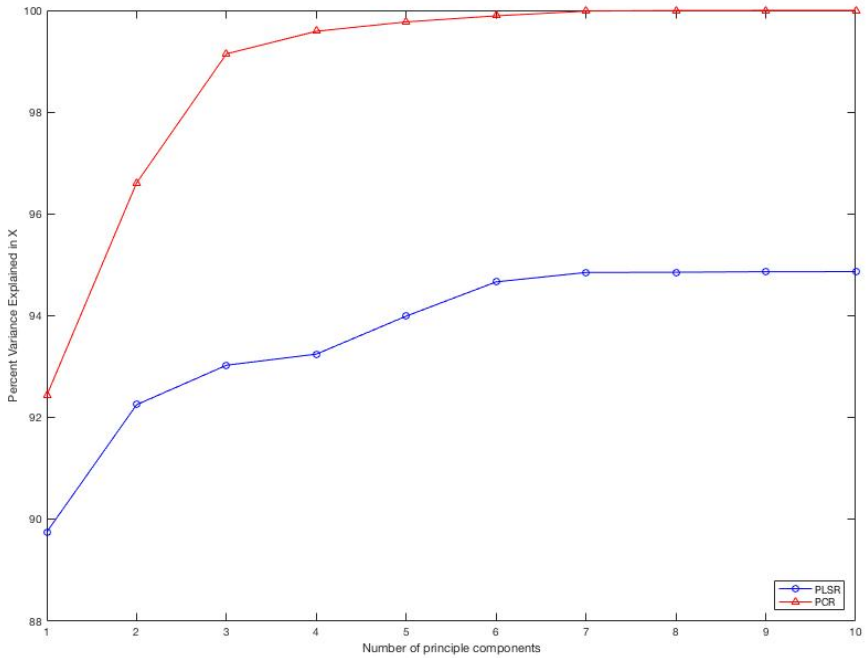


Figure: Percent Variance + psnr

Performance

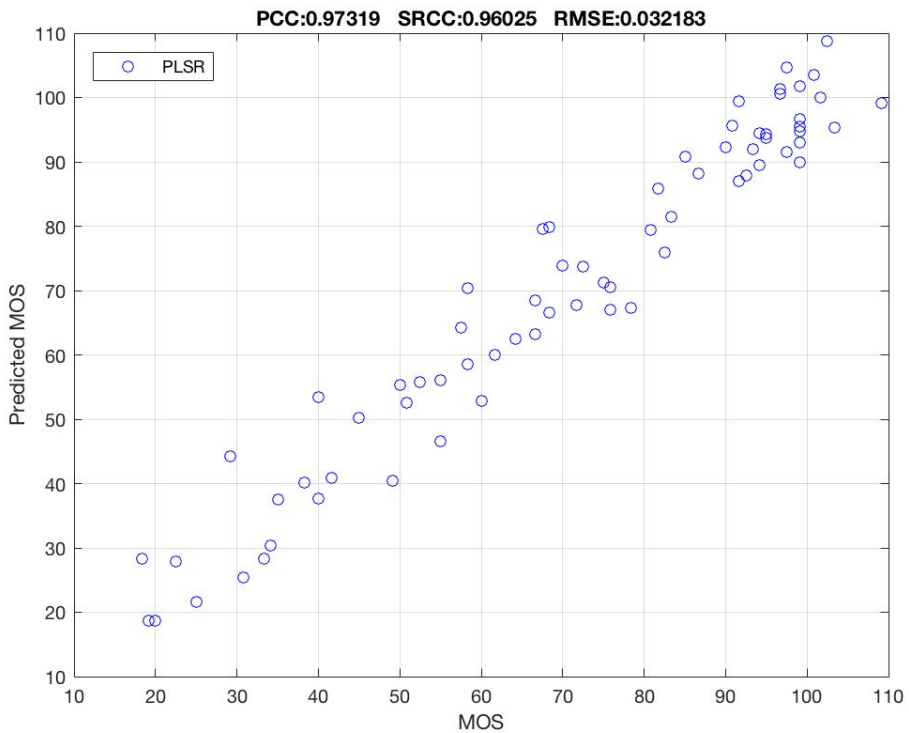


Figure: PLS

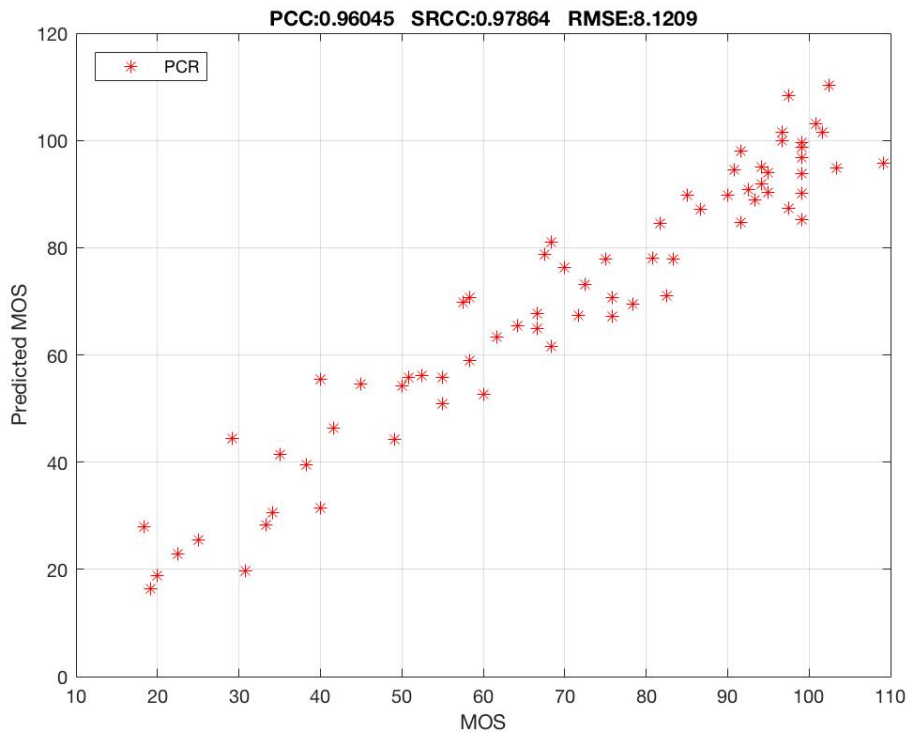


Figure: PCR

Performance

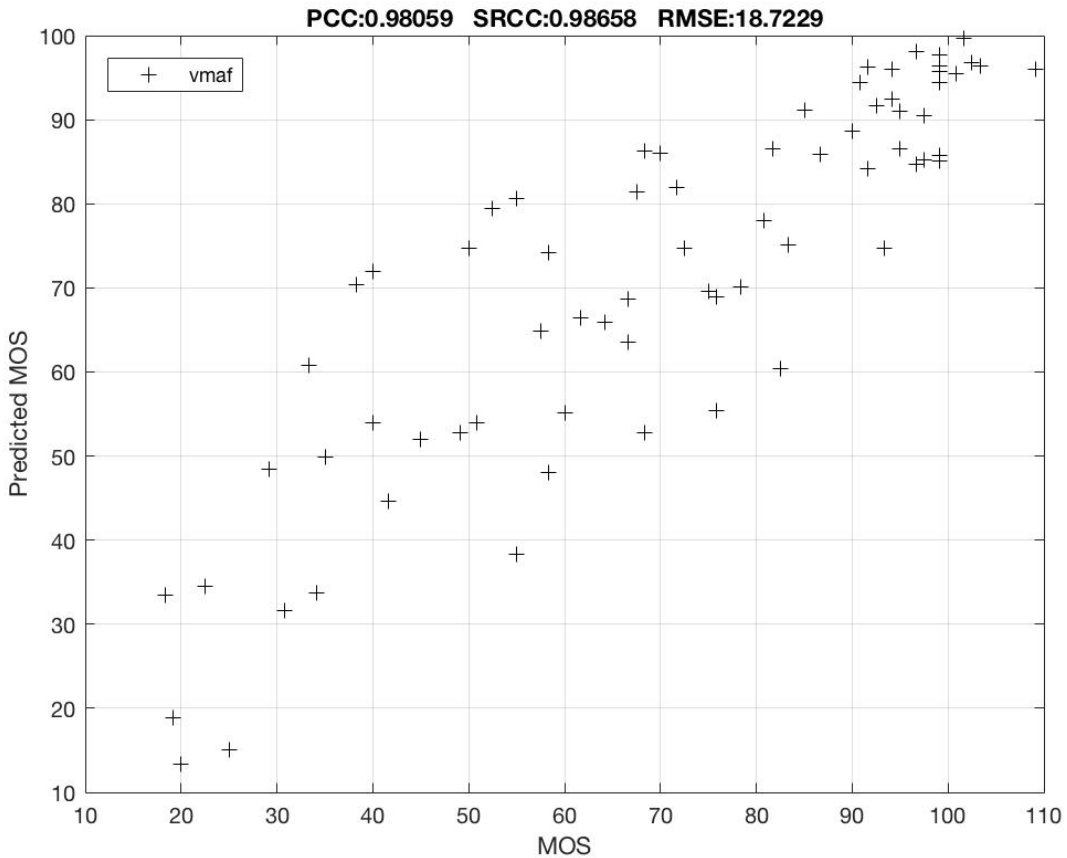


Figure: vmaf

Performance

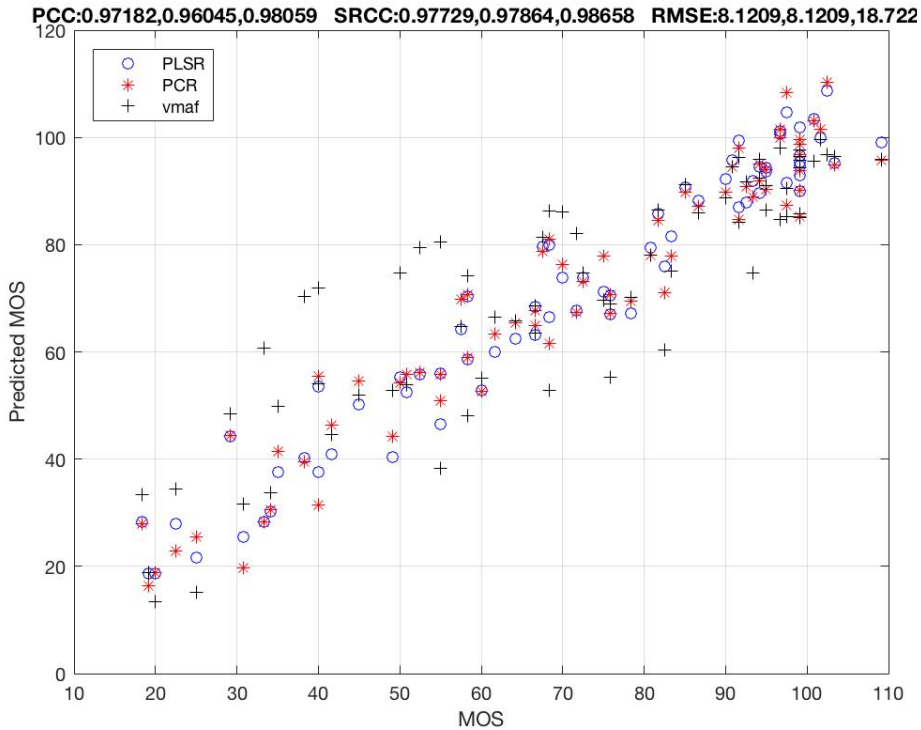


Figure: all

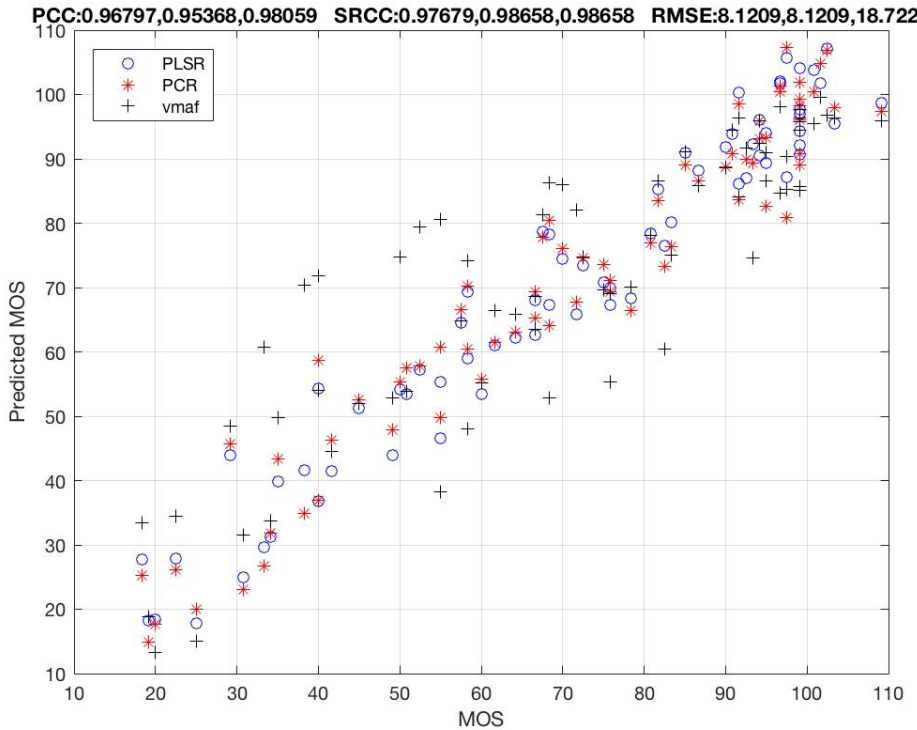


Figure: all+psnr

Performance

Table: Performance Metrics Of Different Models

Models	PCC	SRCC	RMSE
PCR(PC=3)	0.9550	0.9866	8.1209
PCR(PC=4)	0.9599	0.9765	8.1209
PCR(PC=5)	0.9554	0.9765	8.1209
PCR(PC=6)	0.9605	0.9786	8.1209
PLSR	**0.9718**	0.9773	8.1209
PCR+psnr	0.9605	0.9707	8.1209
PLSR+psnr	0.9717	0.9821	8.1209
VMAF	0.9806	0.9866	18.7229

Discussion and Conclusion

- Normalization
- PLSR performs the best comparing to PCR
- Quality of ratings
- Feature selection