Joint Effort Group JEG-Hybrid

Development/research of generally applicable hybrid video quality assessment algorithms

Marcus Barkowsky, Enrico Masala, Glenn Van Wallendael

VQEG Meeting, Oct 14-18, 2019, Shenzhen, China

Mission

- To develop a generally applicable no reference Hybrid Perceptual/Bit-Stream model
- With a small set of subjective experiments
  - Limited training possibilities
  - Limited validation
- Currently
  - Large scale DB with 60,000+ PVS (no losses) and 500,000+ PVS with distortion due to packet losses, many full-reference objective quality measures

Current status

- Working on the JEG-Hybrid DB
  - adding frame-level features from VQM (Matlab open implementation)
  - adding FSIM [1]

- Researching on new approaches to quality evaluation when the MOS is not available (e.g., large scale databases)

- Working with Sky Group on characterizing the behavior of objective measures (publicly available and not) on industry grade content
  - More details by Sky Group

[1] Zhang, L., Zhang, L., Mou, X. and Zhang, D., 2011. FSIM: A feature similarity index for image quality assessment. IEEE transactions on Image Processing, 20(8), pp.2378-2386.

Recent publications

- **Computing Quality-of-Experience Ranges for Video Quality Estimation**, QoMEX 2019
  - By Lohic Fotio Tiotsop, Enrico Masala, Ahmed Aldahdooh, Glenn Van Wallendael, Marcus Barkowsky
- **A Neural Network Based Approach for Observer Behavior Investigation in Media Quality Assessment**, submitted to Signal processing: Image communication journal, 2019
  - By Tomas Mizdos, Lohic Fotio Tiotsop, Enrico Masala, Marcus Barkowsky, Peter Pocta
- More details in a minute

Where can I get more information?

- Biweekly meetings

- http://vqegjeg.intec.ugent.be/wiki/
  (notably section resources, constantly updated, volunteers welcome!)

How may I get involved?

- Subscribe to the VQEG-JEG mailing list jeg@lyris.vqeg.org
  http://www.its.bldrdoc.gov/vqeg/email-reflectors.aspx
- Join our biweekly conference call

Research Activity #1

- **Computing Quality-of-Experience Ranges for Video Quality Estimation**, QoMEX 2019
- Key ideas:
  - Predicting MOS ranges instead of single MOS values
  - As a function of a number of objective measures (PSNR, SSIM, MS-SSIM, VIF, VMAF)
  - To estimate MOS distributions

  - NB: It is not MOS + confidence interval

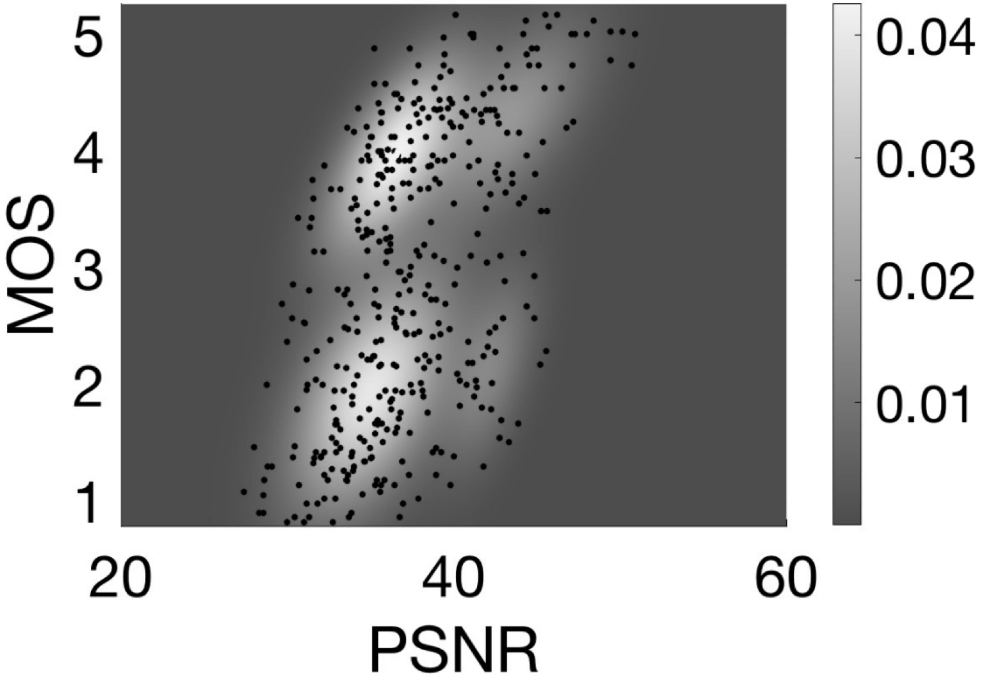
MOS Ranges

- Many works focus on predicting a single MOS value for each PVS

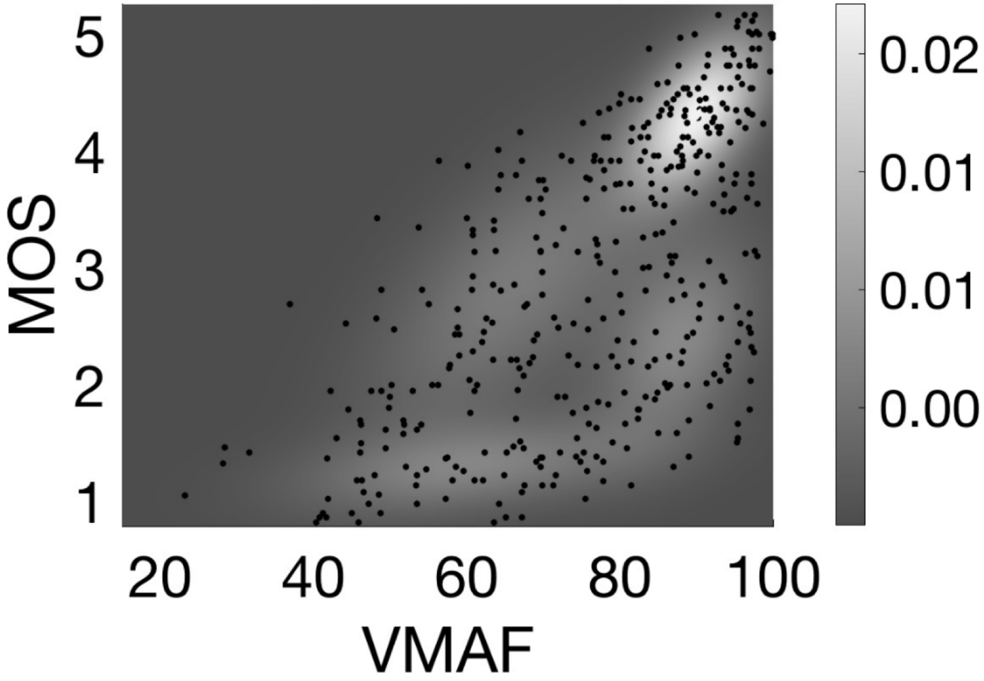
- Radically different approach
  - Take a reasonably varied subjectively annotated dataset (e.g., VQEG-HDTV)
    - Expected to represent a wide range of conditions (content, distortion)
  - Model the behavior of the MOS as function of some parameter (e.g., objective measures M)
    - Proposal: Gaussian Mixture Model (GMM) to estimate the joint distribution f(M, MOS)
  - Compute the desired values from the estimated distribution
    - E.g., MOS_{min} and MOS_{max} so that $P(MOS_{min} < MOS < MOS_{max})$ = desired probability $1-\alpha$

Gaussian Mixture Model

- GMM estimated from VQEG-HDTV



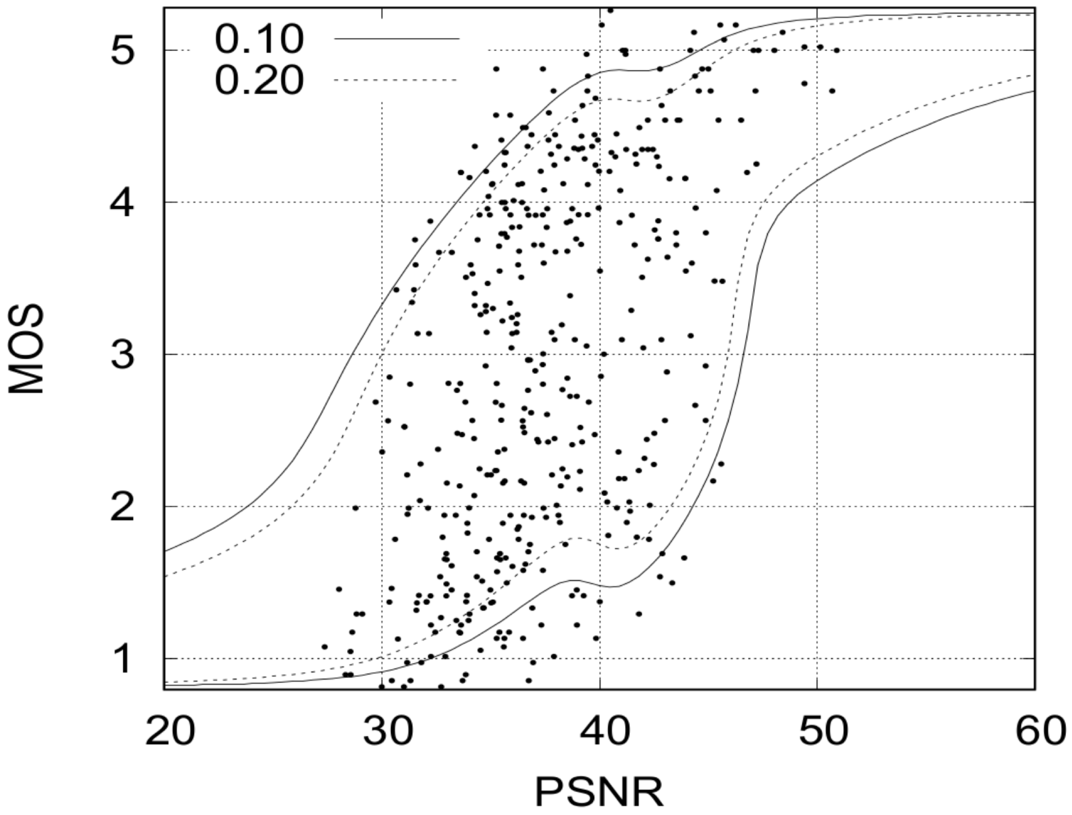
(a) $f(PSNR, MOS)$

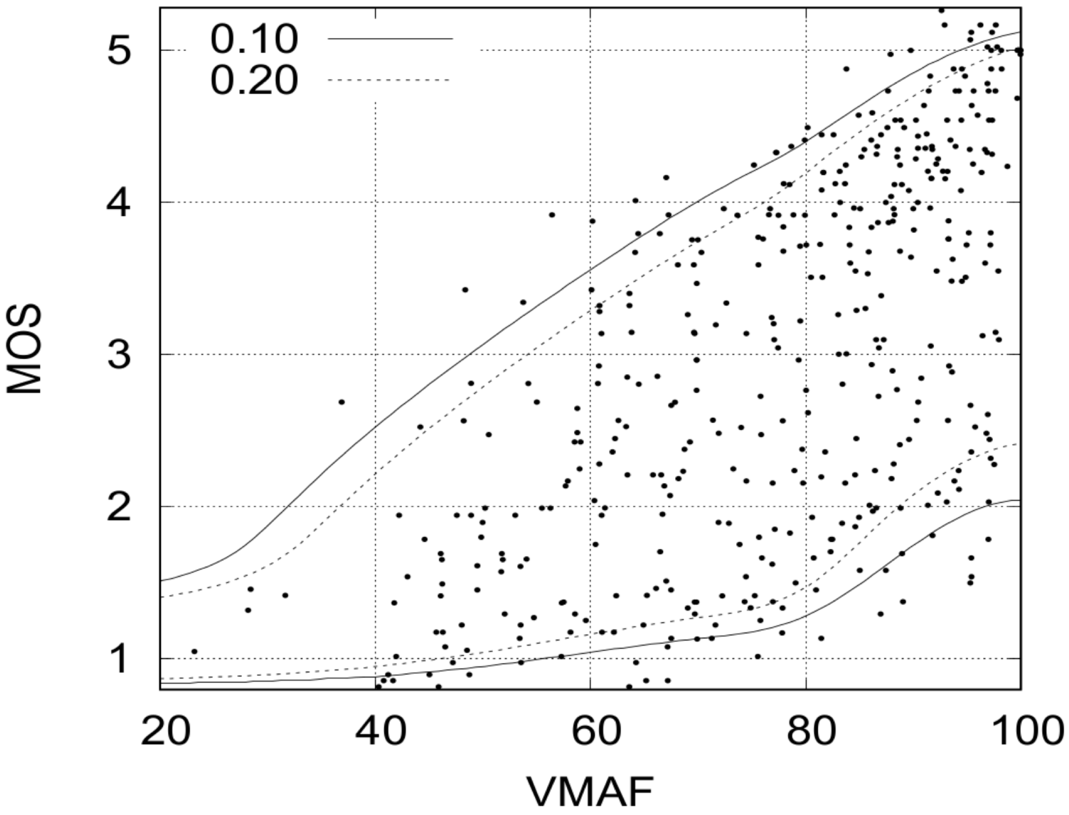


(b) $f(VMAF, MOS)$

Results

- P(MOS in range) = 90% or 80%





Test on Other Datasets

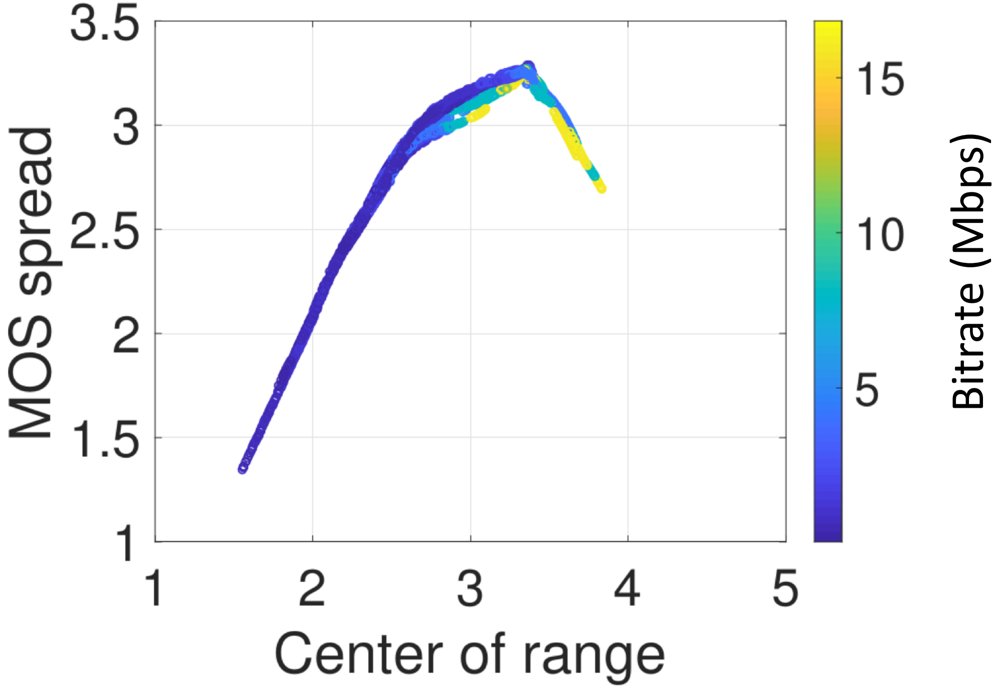
- GMM fitted to VQEG-HD; MOS_{max} (and min) are the average for the PSNR SSIM MSSSIM VIF VMAF0.6.2
- tested on Netflix Public dataset
  α : requested probability of MOS outside range

α	VQEG-HD		Netflix Public	
	Expected	Actual	Expected	Actual
0.01	4/415	0/415	1/70	0/70
0.05	21/415	21/415	4/70	4/70
0.10	42/415	44/415	7/70	13/70
0.15	63/415	70/415	11/70	19/70
0.20	84/415	85/415	14/70	23/70

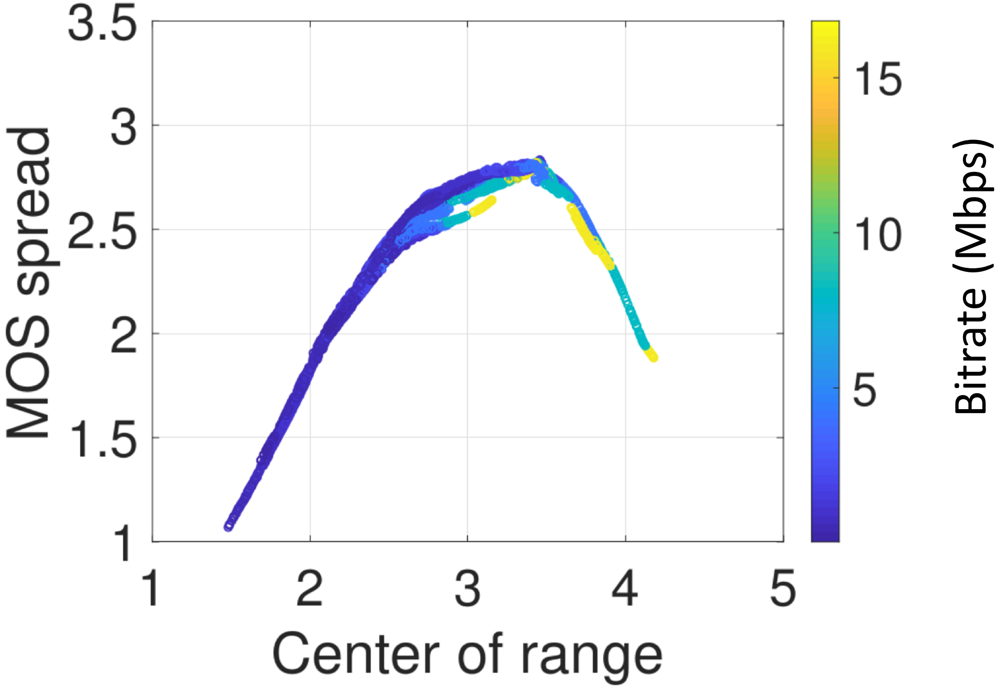
Test on JEG-Hybrid Large Scale DB

- Not subjectively annotated; spread = range = MOS_{max} - MOS_{min}

P(MOS in range) = 90%



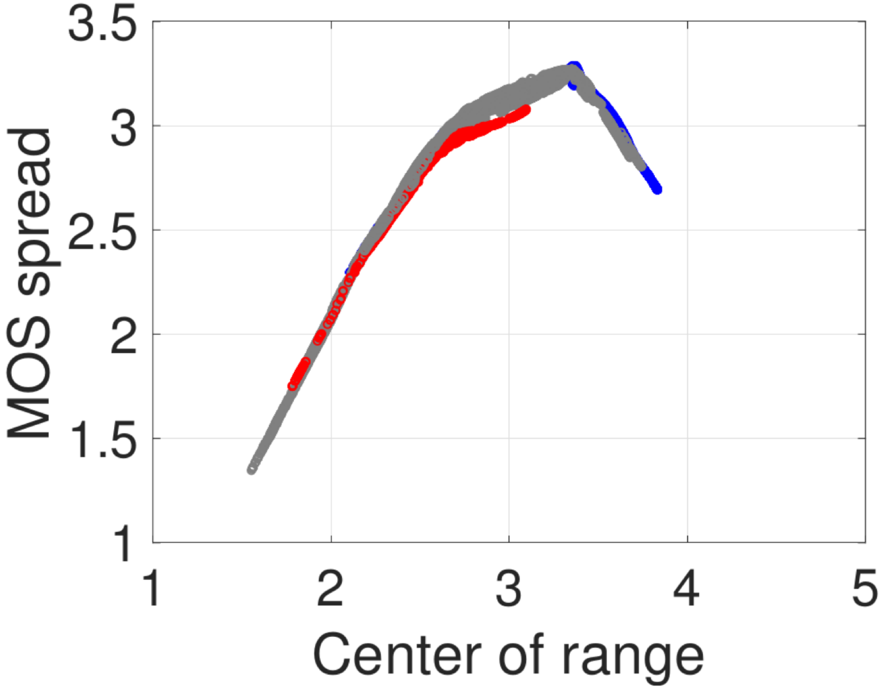
P(MOS in range) = 80%



Test on JEG-Hybrid Large Scale DB

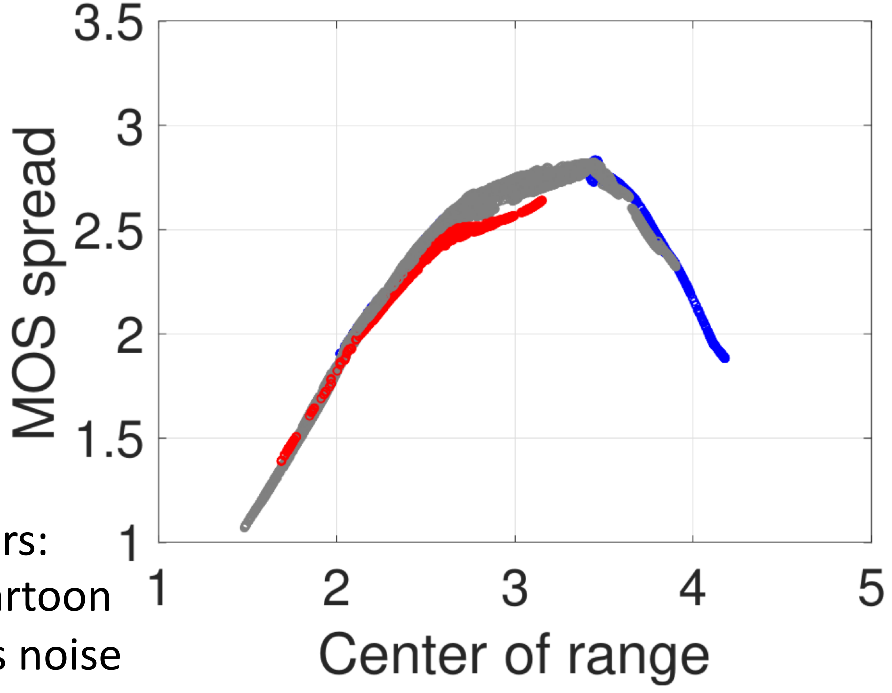
- <u>Not</u> subjectively annotated

P(MOS in range) = 90%



Colors:
Blue: cartoon
Red: has noise

P(MOS in range) = 80%



Future Work

- Consider joint distribution across other objective measures
  - Currently, only one at a time (e.g., PSNR or VMAF), then fusion through mean for MOS_{min} and MOS_{max}
- Better characterization of sequences in VQEG-HDTV
  - Some are heavily distorted, most of objective measures not designed for this purpose
- Ideas?

Research Activity #2

- **A Neural Network Based Approach for Observer Behavior Investigation in Media Quality Assessment**, submitted to Signal Processing: Image Communication (Aug 2019)
- Key ideas:
  - Modeling the behavior of single observers using NN and objective measures as input (as opposed to focusing on predicting MOS directly)
  - Extending existing annotated datasets by creating new content expected to have similar MOS to the annotated one

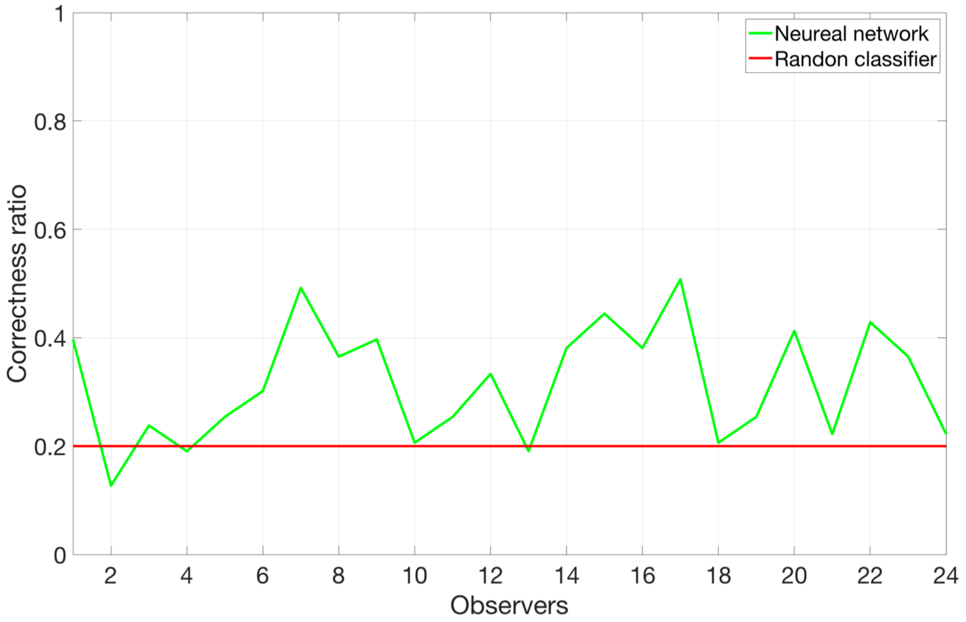
Single Observers Modeling using NN

- Modeling the behavior of single observers using NN and objective measures as input
  - As opposed to focusing on predicting MOS directly
- Obtain sort of "Artificial observers" (AI-based observers)
- Validate their performance
- Use for any purpose where individual scores are useful
  - "Virtual" experiments on non-annotated dataset, to determine if there are potential anomalies in the construction of the dataset
  - Investigate similarities in the behavior of observers

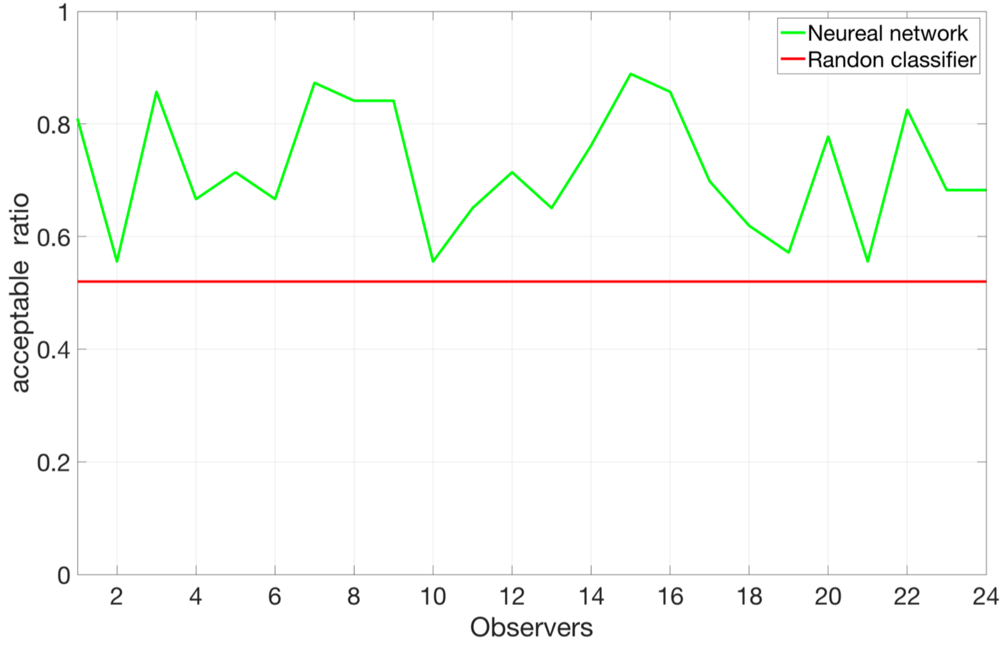
Single Observers Modeling using NN

- Results: 24 AI observers trained on VQEG-HDTV

Predicted = target



| predicted – target | <= 1



Extending Existing Annotated Datasets

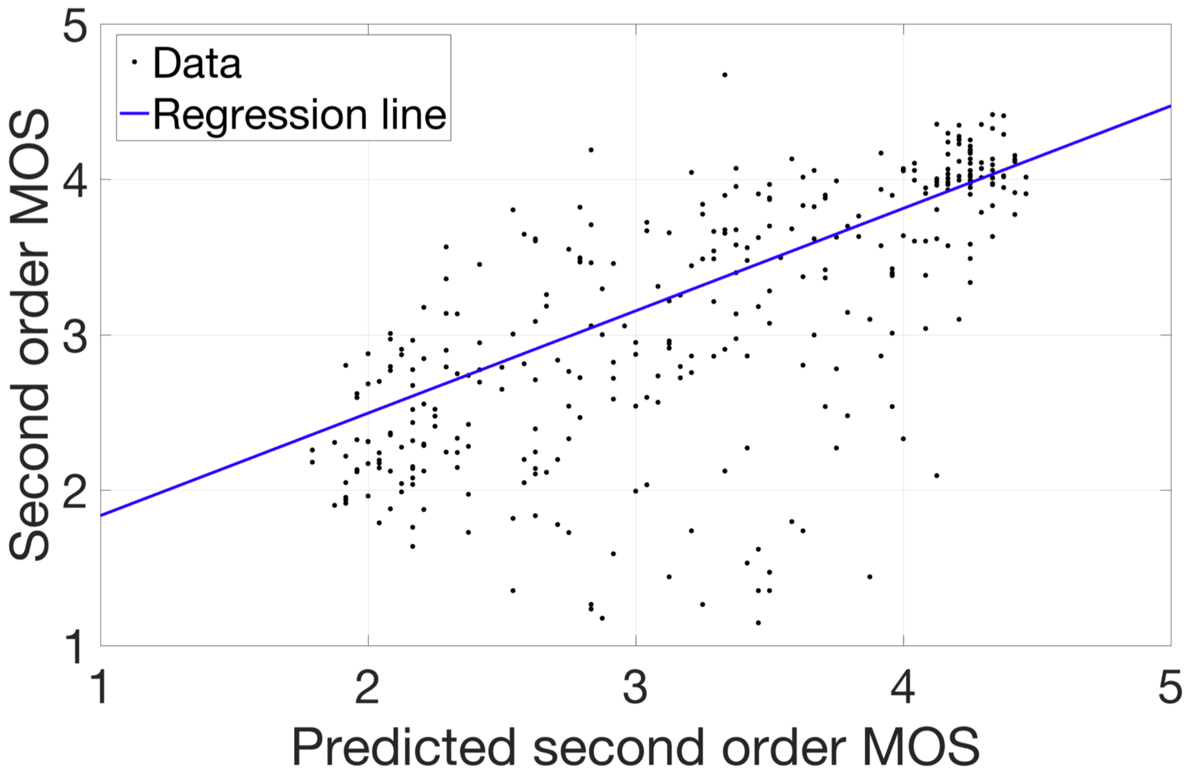
- By creating new content expected to have similar MOS to the annotated dataset
  - Useful for machine learning algorithms: the more the data, the better
- Done on image datasets (e.g., JPEG-distorted) by
  - Taking originals and compressing them with HEVC Intra with various QP
  - Looking at objective measures of annotated and non-annotated (new) content
  - Decide which new content yield the same MOS as the annotated one on the basis of the distance of several objective measures (PSNR, SSIM, VIF)
    - We named the MOS assigned to the extended dataset “second order MOS”

  - Other ideas welcome!

Extending Existing Annotated Datasets

- Results (as first validation step)

Performance of our 24 AI observers when used to predict the second order MOS on the newly created ILT-HEVC dataset. PLCC is 0.68 and SROCC is 0.72.



Conclusions

- We believe that machine learning technique can be useful in the context of quality assessment
- We are trying to develop techniques to extract the most from machine learning
  - Working on MOS ranges, to be able to work on non-annotated data
  - Modeling single observers through NNs to better investigate their behavior
  - Extending annotated datasets to have more input data for NNs