OBJECTIVE MEASURES ON THE ITS4S DATABASE

Enrico Masala
Politecnico di Torino, Italy
enrico.masala@polito.it

VQEG JEG-Hybrid session in Mountain View, Nov 2018
JEG-Hybrid Context

- **Objectively-annotated Large Scale Database**
  - 59,520 HEVC-encoded video sequences (1,920 HRCs)
    - 10 sources, 250 frames each, 25 fps
    - 3 resolutions: 1920x1080, 1280x720, 960x544
      (details in references, already presented in previous meetings)
    - PSNR, SSIM, MS-SSIM, VIF, VQM, VMAF (0.6.0, 0.6.1), PVQM
    - Distortion due to encoding
    - Distortion due to encoding + data (packet) loss (~500,000 samples)
Subjective Annotation?

- **ITS4S database**
  - 4-second length sequences, 1280x720, 24 fps
  - 813 sequences (from 35 video footages)
    - 8 themes + 1 miscellaneous (Broadcast, Everglades, MusicMexico, Nature, Ocean, PublicSafety, Sports, Training)
  - Original purpose: No-Reference Study
  - Subjectively annotated (ACR) by 30+ subjects
  - Original footage is available for all sequences

- **Idea**
  - Run objective quality measures as in the JEG-Hybrid large-scale DB, but having subjective annotation
  - Suitable for this purpose:
    - 518 sequences, 5 HRCs (compression artifacts due to different coding bitrate)
    - A minor share are deemed “bad quality” from the start
    - Note: PVS have to be temporally aligned
Sample of Sequences

• Broadcast
• Chance (miscellaneous)
• Everglades
• Music&Mexico
• Nature
• Ocean
• PublicSafety
• Sports
• Training
Chance (Miscellaneous)
Everglades
Nature
Ocean
Training
Objective Measures

- PSNR, SSIM, MS-SSIM, VIF (from vqmt by EPFL)
- VMAF 0.6.0, 0.6.1
- VMAF 0.6.2 traditional and resulting from bootstrap aggregation ("bagging"), with 95% confidence intervals
- PSNR, SSIM, MS-SSIM (from vmaf by Netflix)
- VQM
  - In some cases, sequences have a few repeated frames to reach 4 sec length needed to run the executable-only vqm version
- PVQM

- Some results in the following
VMAF 0.6.2 (bagging vs “traditional”)
Results: VMAF 0.6.2 vs MOS

![Graph showing the correlation between VMAF 0.6.2 and MOS for different bitrates: 0512K, 0951K, 1256K, 1732K, and 2340K. The graph includes scatter plots with different colors and markers for each bitrate, indicating a trend towards higher MOS values with increasing VMAF values.]
Results: VMAF 0.6.2 vs MOS

![Graph showing the correlation between VMAF 0.6.2 and MOS for various content types including Broadcast, Chance, Everglades, Music&Mexico, Nature, Ocean, PublicSafety, Sports, and Training. The graph illustrates the relationship between VMAF scores and perceived quality as measured by MOS.]
Results: VMAF 0.6.2 vs MOS

![VMAF vs MOS diagram]

- VMAF 0.6.2 vs MOS
- Quantitative analysis of video quality metrics.
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS < predicted)
Visual Comparison (actual MOS > predicted)
Visual Comparison (actual MOS > predicted)
Visual Comparison (actual MOS > predicted)
Visual Comparison (actual MOS > predicted)
Visual Comparison (actual MOS > predicted)

NASA captured this imagery of an M6.7 solar flare
Visual Comparison (actual MOS > predicted)

NASA captured this imagery of an M6.7 solar flare
### Objective Measures

<table>
<thead>
<tr>
<th>name</th>
<th>PSNR</th>
<th>SSIM</th>
<th>MS-SSIM</th>
<th>VIF</th>
<th>VMAF</th>
<th>VMAF b.aggreg</th>
<th>MOS</th>
<th>MOS</th>
<th>bitrate</th>
<th>session</th>
</tr>
</thead>
<tbody>
<tr>
<td>055-Nsrc33O_2340K</td>
<td>46.02</td>
<td>0.994</td>
<td>0.995</td>
<td>0.731</td>
<td>91.878</td>
<td>90.411</td>
<td>5.045</td>
<td>2.111</td>
<td>0.556</td>
<td>2340000</td>
</tr>
<tr>
<td>083-Osrc19K_0512K</td>
<td>43.84</td>
<td>0.994</td>
<td>0.995</td>
<td>0.661</td>
<td>83.154</td>
<td>84.063</td>
<td>6.687</td>
<td>1.852</td>
<td>0.606</td>
<td>512000</td>
</tr>
<tr>
<td>012-Bsrc11A_0512K</td>
<td><strong>30.13</strong></td>
<td>0.962</td>
<td>0.969</td>
<td><strong>0.466</strong></td>
<td>69.916</td>
<td>69.522</td>
<td>6.636</td>
<td>2.037</td>
<td>0.435</td>
<td>512000</td>
</tr>
<tr>
<td>054-Nsrc33O_0512K</td>
<td>41.81</td>
<td>0.993</td>
<td>0.994</td>
<td>0.568</td>
<td>79.990</td>
<td>81.507</td>
<td>6.076</td>
<td>2.259</td>
<td>0.602</td>
<td>512000</td>
</tr>
<tr>
<td>085-Osrc19K_1256K</td>
<td>40.28</td>
<td>0.990</td>
<td>0.992</td>
<td>0.602</td>
<td>80.817</td>
<td>81.035</td>
<td>4.930</td>
<td>2.370</td>
<td>0.586</td>
<td>1256000</td>
</tr>
<tr>
<td>056-Osrc17E_0951K</td>
<td>31.80</td>
<td>0.979</td>
<td>0.980</td>
<td>0.570</td>
<td>66.432</td>
<td>67.084</td>
<td>3.948</td>
<td>4.185</td>
<td>0.504</td>
<td>951000</td>
</tr>
<tr>
<td>005-Osrc16A_0951K</td>
<td>32.58</td>
<td>0.979</td>
<td>0.981</td>
<td><strong>0.354</strong></td>
<td>64.728</td>
<td>65.924</td>
<td>3.223</td>
<td>4.148</td>
<td>0.490</td>
<td>951000</td>
</tr>
<tr>
<td>075-Csrc29U_0951K</td>
<td>34.78</td>
<td>0.973</td>
<td>0.976</td>
<td>0.514</td>
<td>65.197</td>
<td>65.366</td>
<td>3.692</td>
<td>4.074</td>
<td>0.540</td>
<td>951000</td>
</tr>
<tr>
<td>009-Tsrc35B_0951K</td>
<td>31.58</td>
<td>0.974</td>
<td>0.976</td>
<td>0.494</td>
<td>56.333</td>
<td>58.435</td>
<td>3.922</td>
<td>4.000</td>
<td>0.587</td>
<td>951000</td>
</tr>
<tr>
<td>018-Nsrc33F_1732K</td>
<td>29.89</td>
<td>0.937</td>
<td>0.944</td>
<td><strong>0.386</strong></td>
<td>63.703</td>
<td>61.466</td>
<td>6.095</td>
<td>3.889</td>
<td>0.660</td>
<td>1732000</td>
</tr>
</tbody>
</table>
Results: VMAF 0.6.2 (CI) vs MOS
Content Dependency

![Graph showing content dependency with MOS on the x-axis and VMAF 062 (bagging) on the y-axis. The graph includes various categories such as Broadcast, Chance, Everglades, Music&Mexico, Nature, Ocean, PublicSafety, Sports, and Training. There are distinct clusters and trends represented by different colored markers.]
VMAF & MOS CIs (95%) 

- 35 -
Confidence Intervals: MOS & VMAF

- As a function of the encoding bitrate
- MOS: about [0.4-0.8]
- VMAF: \([3-12]/100 = [0.15-0.6]/5\), higher for lower bitrates
Confidence Intervals: MOS & VMAF

• As a function of the session type
Confidence Intervals: MOS & VMAF

- Scatter plot: no particular behavior
Other Measures
Other Measures

- VQM, PVQM (not a simple arithmetic mean of frame-based measures)


SSIM: Netflix vs vqmt

- Same C1 and C2 constants in the formula
- SSIM vs downscaled SSIM?
MS-SSIM: Netflix vs vqmt

- Only small differences in case of MS-SSIM
Results

• Data available here:
  • CSV file: http://media.polito.it/downloads/jeg/its4s/
  • On Google Datastudio: https://bit.ly/its4s_2018
    https://datastudio.google.com/open/1MUqys1gsnEdKJHFK5q_112oEYqB9e4xZ
Conclusions

- Large content variety helps in characterizing objective measures and find areas of improvement
- Confidence intervals by VMAF might be useful, but bootstrapping aggregation value should be used
- Size of confidence intervals in subjective experiments is mostly between $[0.4-0.8]$ MOS
- Size of VMAF confidence interval is mostly between $[3-12]/100 = [0.15-0.6]/5$, with higher values for lower bitrates
Future Plans

• Better characterize observations that can be done with objective measures, trying to understand how much they can predict potentially difficult situations in subjective evaluation

• Compute the new objective measures on the Large-Scale JEG-Hybrid database of ~60,000 sequences
  • VMAF 0.6.2 with confidence intervals
  • SSIM and MS-SSIM by Netflix (and compare them with the vqmt)
References

• M. Pinson, “ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes”, TM-18-532, 2018
• ftp://iotnas001.th-deg.de/VQEG/JEG/HYBRID/hevc_database/
• http://media.polito.it/downloads/jeg/
• Backup slides
Content Dependency

![Graph showing content dependency with MOS on the x-axis and VMAF 062 (bagging) on the y-axis. The graph includes a scatter plot with various categories such as Broadcast, Chance, Everglades, Music & Mexico, Nature, Ocean, Public Safety, Sports, and Training.](image-url)