Viewport-driven Multi-metric Fusion Approach for 360° Video Quality Assessment

EPFL: Roberto Azevedo, Pascal Frossard
YouTube Media Algorithm: Neil Birkbeck, Ivan Janatra, Balu Adsumilli
Background

- Subjective and objective quality for 360° videos still an open problem
  - VR headset -> Increased level of immersion -> Changes the QoE perspective
- Follow up from our previous study with more limited dataset (Azevedo et al., 2020)
  - Individual metrics computed on viewports correlates better with subjective scores than metrics computed on the projection domain...
  - ...but no single metric performs best across all distortion types
- Objective: Build a multi-metric model (e.g. VMAF for 2D videos) for 360-degree VQA

Related work

Error-based metrics

Related work

Deep learning

- MC360I3D (image-only)
- DeepVR-IQA (image-only)
- V-CNN (video, viewport-based CNN)

General Approach
We tried 3 viewport sampling modes x 3 FOV (30°, 40°, 50°) … as with our previous study, Uniform 40° seems to perform best.

Example - Uniform 40°
Objective Metrics

Spatial Activity

\[ S(z) = \sqrt{(G_1 \ast z)^2 + (G_1^T \ast z)^2}, \]

\[ G_1 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \]

\[ s = S(u) - S(v). \]

\[ \text{SA}(v, u) = \sqrt{\frac{1}{MN} \sum_{i,j} |s_{ij}|^2}, \]

Objective Metrics

PSNR-HVS and PSNR-HVS-M

- **PSNR-HVS**
  - Divides image in 8x8 non-overlapping blocks, and
  - Applies weight on the difference based on contrast sensitivity function (CSF)

- **PSNR-HVS-M**
  - Like PSNR-HVS, with additional contrast masking multiplier applied to the DCT coefficients difference

---

Objective Metrics

SSIM and MS-SSIM

- **SSIM**
  - Luminance
    \[ I(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \]
  - Contrast
    \[ C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \]
  - Structure
    \[ S(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \]
  - SSIM
    \[ SSIM(x, y) = [(I(x, y)]^\alpha \cdot [C(x, y)]^\beta \cdot S(x, y)]^\gamma \]

- **MS-SSIM**

  \[ MSSSIM(x, y) = [I_m(x, y)]^{\alpha_m} \prod_{m=1}^{M} [C_m(x, y)]^{\beta_m} \cdot [S_m(x, y)]^{\gamma_m} \]

Objective Metrics

Gradient-magnitude Similarity Deviation (GMSD)

\[ GMS(u, v) = \frac{2 \cdot m(u) \cdot m(v) + c}{m(u)^2 + m(v)^2 + c} , \]

\[ m(z) = \sqrt{(z \ast G_2)^2 + (z \ast G_2^T)^2} , \]

\[ G_2 = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix} . \]

\[ \text{GMSD}(u, v) = \sqrt{\frac{1}{NM} \sum_{ij} \left( GMS(u, v) - \bar{GMS}(u, v) \right)^2} , \]

\[ \bar{GMS}(u, v) = \frac{1}{NM} \sum_{ij} GMS(u, v) . \]


(a) Original. (b) Distorted. (c) GMS map.
Objective Metrics

Relative change in Temporal Information

- Current 360-VQA approaches don’t seem to incorporate temporal effects

\[ TI[F_n] = \text{std}(\Delta F_n), \text{ where } \Delta F_n = F_n - F_{n-1} \]

\[ TI_{rel}[F_n] = \frac{|TI_{ref}[F_n] - TI_{dist}[F_n]|}{TI_{ref}[F_n]} \]
Temporal Pooling

- Metrics computed per frame, then pooled. Why?
  - Smooth effect
  - Asymmetric effect
  - Recency effect

\[ Q_{LP}^n(f) = \begin{cases} 
Q_{LP}^n(f - 1) + \alpha \cdot \Delta Q(f), & \text{if } \Delta Q^n \leq 0 \\
Q_{LP}^n(f - 1) + \beta \cdot \Delta Q(f), & \text{if } \Delta Q^n > 0 
\end{cases} \]

\[ Q_{pool}^n = \frac{1}{F} \sum_{f=1}^{F} (Q_{LP}^n(f) \cdot ln(\gamma \cdot f + 1)) \]

Use $\alpha = 0.03$, $\beta = 0.2$, $\gamma = 1000$.

Regression

- Generate feature vector containing each combination of pooled metric and viewport
- Use these to learn non-linear mapping with subjective scores
- Tested both SVR and RFR, ended up using RFR
- Run the following:
  - Our method (projection, VP collage, and VP domains)
  - PSNR (projection and VP collage domains)
  - S-PSNR
  - WS-PSNR
  - MS-SSIM (projection and VP collage domains)
  - VMAF (projection and VP collage domains)
Experiments

- We ran two experiments:
  - Fixed train-test set: use single fixed 80% train/validation set and 20% test set, prescribed by Dataset.
  - Cross-validation: in each of the 1000 runs, split Dataset to 80% train/validation set and 20% test set, and run as Fixed.
Dataset

VQA-ODV

- Contains 60 ref + 180 impaired equirect sequences.
  - Ref videos have varying resolutions (4k-8k), varying length (10-23s), varying fps (24-30fps)
  - Impaired videos use H.265 encoding with 3 QP levels (27, 32, 42)
- Rating from 221 subjects, divided into 10 groups
  - Use single-stimulus with hidden reference
  - Has MOS and DMOS
- Using HTC Vive as HMD; take HMD resolution into account when sampling viewport

Dataset

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>Group 4</td>
<td>Group 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 5</td>
<td>Group 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 7</td>
<td>Group 8</td>
<td></td>
</tr>
</tbody>
</table>

Fixed train-test sets

- For our method: Run group shuffle cross-validation on training set to find best RF hyper-parameters, train the model on training set and test on the test set.
- For comparison metrics: Fit a 4-parameter logistic function on the training set, and compute its function with the test set.

### Fixed train-test sets

#### Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>PLCC</th>
<th>SROCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR (Proj.)</td>
<td>0.72495</td>
<td>0.73797</td>
<td>8.176</td>
</tr>
<tr>
<td>PSNR (VP-Collage)</td>
<td>0.76222</td>
<td>0.76345</td>
<td>7.5824</td>
</tr>
<tr>
<td>S-PSNR</td>
<td>0.75138</td>
<td>0.7704</td>
<td>7.7557</td>
</tr>
<tr>
<td>WS-PSNR</td>
<td>0.74328</td>
<td>0.56056</td>
<td>7.9501</td>
</tr>
<tr>
<td>MS-SSIM (Proj.)</td>
<td>0.76005</td>
<td>0.78867</td>
<td>7.8741</td>
</tr>
<tr>
<td>MS-SSIM (VP-Collage)</td>
<td>0.81719</td>
<td>0.84144</td>
<td>7.0024</td>
</tr>
<tr>
<td>VMAF (Proj.)</td>
<td>0.79657</td>
<td>0.79382</td>
<td>7.2481</td>
</tr>
<tr>
<td>VMAF (VP-Collage)</td>
<td>0.84483</td>
<td>0.85637</td>
<td>6.271</td>
</tr>
<tr>
<td>Ours (Proj.)</td>
<td>0.85629</td>
<td>0.86873</td>
<td>6.3588</td>
</tr>
<tr>
<td>Ours (VP-Collage)</td>
<td>0.89867</td>
<td>0.87439</td>
<td>5.7256</td>
</tr>
<tr>
<td>Ours (VP)</td>
<td>0.92575</td>
<td>0.91712</td>
<td>4.9954</td>
</tr>
</tbody>
</table>

VP-Collage domain generally outperforms projection domain.

Our method outperforms VMAF due to selection of individual metrics and improved temporal pooling.

*best*
Fixed train-test sets

Results

Average viewport features importance in our viewport method (for VQA-ODV)
Cross-Validation

Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>PLCC</th>
<th>SROCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR (Proj.)</td>
<td>0.5716</td>
<td>0.6187</td>
<td>9.8249</td>
</tr>
<tr>
<td>PSNR (VP-Collage)</td>
<td>0.6475</td>
<td>0.6858</td>
<td>9.1224</td>
</tr>
<tr>
<td>S-PSNR</td>
<td>0.6246</td>
<td>0.6673</td>
<td>9.3461</td>
</tr>
<tr>
<td>WS-PSNR</td>
<td>0.5980</td>
<td>0.6450</td>
<td>9.5983</td>
</tr>
<tr>
<td>MS-SSIM (Proj.)</td>
<td>0.7500</td>
<td>0.7753</td>
<td>7.9351</td>
</tr>
<tr>
<td>MS-SSIM (VP-Collage)</td>
<td>0.7640</td>
<td>0.7911</td>
<td>7.7578</td>
</tr>
<tr>
<td>VMAF (Proj.)</td>
<td>0.7469</td>
<td>0.7667</td>
<td>7.9631</td>
</tr>
<tr>
<td>VMAF (VP-Collage)</td>
<td>0.7809</td>
<td>0.7980</td>
<td>7.5147</td>
</tr>
<tr>
<td>Ours (Proj.)</td>
<td>0.8173</td>
<td>0.8290</td>
<td>6.8716</td>
</tr>
<tr>
<td>Ours (VP-Collage)</td>
<td>0.8268</td>
<td>0.8264</td>
<td>6.7376</td>
</tr>
<tr>
<td>Ours (VP)</td>
<td>0.8677</td>
<td>0.8677</td>
<td>5.9367</td>
</tr>
</tbody>
</table>

Our method outperforms VMAF due to selection of individual metrics and improved temporal pooling.

VP-Collage domain generally outperforms projection domain.

Best results are highlighted in red.
Cross-Validation

Results

Our method (VP) has:

- smallest range of value
- best average
- higher density
Conclusion

- Viewport-based MMF achieves very good results compared to other objective metrics
  - Even just MMF (without viewport) outperforms single metrics
  - Metrics of separate viewports outperforms metrics of collaged viewports
  - Not as training-data-hungry as deep learning techniques
- Using viewport means it should also work for other projections
- Using multimetric means other individual metric can be added if the type of distortion in the dataset is known
Future work

- Verify our method on multiple datasets
- Verify our method on different projections
- Consider visual attention data (available on VQA-ODV dataset)
Questions / Discussion