Learning Compressible 360° Video Isomers

Yu-Chuan Su and Kristen Grauman
The University of Texas at Austin

1. Preliminary: 360° Video Compression

Common Pipeline for 360° Video Compression

- 360° video compression is under rapid development
- The main challenge is to find a proper sphere-to-plane projection

2. Main Idea: 360° Video Isomers

360° videos are equivalent under rotation, but the encoded bit streams are not.

3. Cubemap Size Analysis

Dataset
- High resolution (4K) 360° videos crawled from YouTube
- 80 videos / 4.2 hours total length
- One cubemap orientation for each GOP (fixed 2s clip)
- Sample the orientation along two rotation axis

4. Approach

- Exhaustive search is computationally infeasible (1.5 hours for a 2 seconds video)
- Our approach – predict optimal orientation from video content

- Learn to predict the video size at each orientation
- Use skip connection to retain information for image details

5. Experiment Results

Prediction Examples

Size Distribution

- Cubemap size distribution is non-uniform w.r.t. orientation
- Encoded bit stream size depends on the video content

Reasons for Size Difference

- Cubemap is not perspective image
- Face boundaries introduce discontinuity
- Content distort near boundaries

Achievable Size Reduction Through Rotation

<table>
<thead>
<tr>
<th></th>
<th>H264</th>
<th>HEVC</th>
<th>VP9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg (%)</td>
<td>10.37±8.79</td>
<td>9.88±8.23</td>
<td>9.76±6.62</td>
</tr>
<tr>
<td>Range (%)</td>
<td>[1.08, 76.93]</td>
<td>[1.40, 74.95]</td>
<td>[1.70, 75.84]</td>
</tr>
</tbody>
</table>

- Size reduction – \( r = 100 \times \frac{S_{max} - S_{min}}{S_{max}} \)
- Reduce the video size 1) by 10% overall, 2) by up to 77% for single video

Quantitative Results

Percentage of reduction by rotation

- Our method achieves 82% the size reduction achieved by rotation
- The absolute size reduction is roughly linear to the original video size
- Our method requires 0.3% the computation compared with exhaustive search

Generalizability

- The learned model can generalize to different encoders
- The model generalizes well to different visual content