Deep Video Precoding: Toward A Generalized Deep Perceptual Optimizer

Technical presentation: AOMedia Research Symposium

Aaron Chadha, Russell Anam, Eirina Bourtsoulatze, Ilya Fadeev, Vasileios Giotsas, Yiannis Andreopoulos

October 2019
SUMMARY OF WHAT WE DO

Our solution can be applied either on both sides, or on the encoding side only. For more technical details please visit our website [https://www.isize.co/bitsave/](https://www.isize.co/bitsave/) [https://www.isize.co/upscale/](https://www.isize.co/upscale/)

- Perceptual Quality Metrics (VMAF, VIF, DLM, MS-SSIM...), included in VQEG and used by Netflix, video standardization, etc.


- Any MPEG or Open Encoder → Bitstream → Any MPEG or Open Decoder → Decoded Video Frame

iSIZE + Conventional Codec Pipeline: Deep Video Precoding and Post-Decoding Enhancement

- Any MPEG or Open Encoder → iSIZE Precoder: Deep Perceptual Optimizer → Bitstream → Any MPEG or Open Decoder → Decoded Video Frame

- Perceptual Quality Metrics (VMAF, VIF, DLM, MS-SSIM...) included in VQEG and used by Netflix, video standardization, etc.

- iSIZE Post-Decoder (OPTIONAL)
SUMMARY OF WHY WE DO IT

- Moore’s law (mainly device power+heat dissipation) and cloud-based scaling have both hit the wall
- **We are at an inflection point**: perceptual metrics and DL are mature enough to allow for robust perceptual precoding
- Codecs are amazing SNR/SSIM-to-bitrate machines, but these loss functions have significant limitations

[D. Ronca, Facebook, 2019](https://www.linkedin.com/pulse/encoder-complexity-hits-wall-david-ronca/)
[Sikora, Proc. of the IEEE, 2005,](https://doi.org/10.1109/JPROC.2004.839601)
OUR GOAL: REVERSE-ENGINEER HUMAN VISUAL PERCEPTION

• Deep video precoding is a deep perceptual optimizer using neural networks:
  
  - This escapes the legacy constraints of video encoding standards and is codec independent
  - It enables AI to penetrate video encoding & delivery with backward compatibility and single-pass stream processing
  - It creates an adversarial approach between deep video precoding and advanced perceptual metrics

→ This may allow us to completely reverse-engineer human visual perception of VoD and livestreamed content
ISIZE PRECODER – AVAILABLE AT BITSAVE.TECH

- Our demo shows content that has been precoded by iSize Precoding and encoded by HEVC in 4K resolution at half the bitrate without any loss of VMAF compared with normal 4K HEVC encoded content.

1.) iSIZE precoding
2.) HEVC encoder at 15mbps in 4K
3.) Playing 15mbps 4K stream
4.) VMAF 92.5 achieved

1.) HEVC encoder at 30mbps in 4K
3.) Playing 30mbps 4K
4.) VMAF 76.0 achieved
Select “Precoder” or “R&D” to get the precoder output and use your own encoder to compare against your own recipe.
RESULTS ON FHD CONTENT
HEVC VBR, TuneVMAF, HD resolution, 3Mbps-12Mbps

Notes:
1. For fair comparison, we use the same codec settings.
2. VMAF is a state-of-the-art perceptual quality metric proposed by Netflix and recognized by VQEG, see:
   https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12
3. BD-rate is -47% or 6.9 VMAF points.
4. VMAF of 6 points is the JND threshold.
5. We use arithmetic-mean VMAF (FFmpeg libvmaf model v0.6.1) but similar or higher ΔVMAF was obtained with harmonic mean.
RESULTS ON FHD CONTENT: IN MORE DETAIL
HEVC VBR, TuneVMAF, HD resolution, 3Mbps-12Mbps

Notes:
1. For fair comparison, we use the same codec settings.
2. MS-SSIM is the multi-scale structural similarity index metric.
3. BD-rate is 47% or 0.9% of MS-SSIM.
RESULTS ON FHD CONTENT: IN MORE DETAIL
HEVC VBR, TuneVMAF, HD resolution, 3Mbps-12Mbps

Notes:
1. For fair comparison, we use the same codec settings.
2. VIF is the visual information fidelity metric.
3. BD-rate is -39% or 5.3 points.
RESULTS ON FHD CONTENT: IN MORE DETAIL
HEVC VBR, TuneVMAF, HD resolution, 3Mbps-12Mbps

Notes:
1. For fair comparison, we use the same codec settings.
2. ADM2 is a variant of the DLM (detail loss metric).
3. BD-rate is -88% or 7.8 points.
FRAME-BY-FRAME COMPARISON

![Graph showing VMAF log rush field cuts for 1080p30 HEVC, 5mbps and iSIZE+HEVC, 5mbps]
FRAME-BY-FRAME COMPARISON

SSIM log rush field cuts 1080p30

- HEVC, 5mbps
- iSIZE+HEVC, 5mbps
HEVC RESULTS ON FHD CONTENT VIA AMAZON MTURK

Instructions

View full instructions

Watch two versions of the same video side-by-side, and select if (i) the visual quality is about the same, (ii) left is better, or (iii) right is better.

You can also pause to inspect left-right sections of frames, or watch sections of the video multiple times if it helps your assessment.

You need to watch the video in full before you rate the visual quality.

Select which side of the video has better visual quality.

The options below will be activated once you have watched the video.

You have watched 4 seconds so far.

Left  About the same  Right
HEVC RESULTS ON FHD CONTENT VIA MTURK

<table>
<thead>
<tr>
<th>Video Title</th>
<th>Bitrate (iSize/Codec)</th>
<th>iSize+HEVC Better (%)</th>
<th>HEVC Better (%)</th>
<th>No Preference (%)</th>
<th>ΔVMAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>tractor</td>
<td>5m/5m</td>
<td>88</td>
<td>07</td>
<td>05</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>44</td>
<td>26</td>
<td>31</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>95</td>
<td>05</td>
<td>00</td>
<td>0.2</td>
</tr>
<tr>
<td>old town</td>
<td>5m/5m</td>
<td>79</td>
<td>14</td>
<td>07</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>56</td>
<td>28</td>
<td>15</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>63</td>
<td>30</td>
<td>08</td>
<td>5.8</td>
</tr>
<tr>
<td>crowd run</td>
<td>5m/5m</td>
<td>74</td>
<td>12</td>
<td>14</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>28</td>
<td>35</td>
<td>38</td>
<td>-3.3</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>56</td>
<td>10</td>
<td>33</td>
<td>0.2</td>
</tr>
<tr>
<td>aspen</td>
<td>5m/5m</td>
<td>67</td>
<td>08</td>
<td>26</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>54</td>
<td>13</td>
<td>33</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>63</td>
<td>05</td>
<td>32</td>
<td>4.6</td>
</tr>
<tr>
<td>controlled burn</td>
<td>5m/5m</td>
<td>63</td>
<td>05</td>
<td>33</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>62</td>
<td>05</td>
<td>32</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>58</td>
<td>10</td>
<td>33</td>
<td>3.9</td>
</tr>
<tr>
<td>rush field</td>
<td>5m/5m</td>
<td>75</td>
<td>00</td>
<td>25</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>63</td>
<td>08</td>
<td>30</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>56</td>
<td>03</td>
<td>41</td>
<td>3.4</td>
</tr>
<tr>
<td>touchdown pass</td>
<td>5m/5m</td>
<td>46</td>
<td>26</td>
<td>28</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>60</td>
<td>28</td>
<td>12</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>62</td>
<td>13</td>
<td>30</td>
<td>1.9</td>
</tr>
<tr>
<td>Average</td>
<td>5m/5m</td>
<td>70.2</td>
<td>10.2</td>
<td>19.6</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>5m/8.5m</td>
<td>52.3</td>
<td>20.4</td>
<td>27.3</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>8.5m/14m</td>
<td>64.4</td>
<td>10.6</td>
<td>25.0</td>
<td>2.9</td>
</tr>
</tbody>
</table>

A|B videos and full bitstreams: [https://www.size.co/portfolio/demo](https://www.size.co/portfolio/demo)
RESULTS ON 8K CONTENT
HEVC VBR, TuneVMAF, 8K resolution

- Clips from red.com
- Downscaling was by factor (4/3,4/3)
FRAME-BY-FRAME COMPARISON

8K content: Typical VMAF plot

- iSIZE+HEVC at 40mbps, mean VMAF=92.0
- HEVC at 80mbps, mean VMAF=86.1
FRAME-BY-FRAME COMPARISON

8K content: Typical VMAF plot

- Green dashed line: iSIZe+HEVC at 40mbps, mean VMAF=92.0
- Blue line: HEVC at 40mbps, mean VMAF=81.5
RESULTS ON FHD CONTENT AND VERSATILE VIDEO CODING

VVC JVET VTM v.6.2rc1, IntraPeriod=64, RateControl=1

Notes:
1. For fair comparison, we use the same codec settings.
2. VMAF is a state-of-the-art perceptual quality metric proposed by Netflix and recognized by VQEG, see:
3. We use arithmetic-mean VMAF (FFmpeg libvmaf model v0.6.1) but similar or higher ∆VMAF was obtained with harmonic mean
RESULTS ON FHD CONTENT AND VVC: IN MORE DETAIL

VVC JVET VTM v.6.2rc1, IntraPeriod=64, RateControl=1

Notes:
1. For fair comparison, we use the same codec settings.
2. MS-SSIM is the multi-scale structural similarity index metric.
RESULTS ON FHD CONTENT AND VVC: IN MORE DETAIL

VVC JVET VTM v.6.2rc1, IntraPeriod=64, RateControl=1

Notes:
1. For fair comparison, we use the same codec settings.
2. VIF is the visual information fidelity metric.
RESULTS ON FHD CONTENT AND VVC: IN MORE DETAIL

VVC JVET VTM v.6.2rc1, IntraPeriod=64, RateControl=1

Notes:
1. For fair comparison, we use the same codec settings.
2. ADM2 is a variant of the DLM (detail loss metric).
FRAME-BY-FRAME COMPARISON

VMAF log rush field cuts 1080p30

- VVC 1.8mbps
- iSIZE+VVC 1.8mbps
FRAME-BY-FRAME COMPARISON

SSIM log rush field cuts 1080p30

- VVC 1.8mbps
- iSIZE+VVC 1.8mbps
# VVC RESULTS ON FHD CONTENT: VIA MTURK

A|B videos and full bitstreams: [https://www.isize.co/portfolio/demo2](https://www.isize.co/portfolio/demo2)

<table>
<thead>
<tr>
<th>Video Title</th>
<th>Bitrate (iSize/VVC)</th>
<th>iSize+Codec Better (%)</th>
<th>VVC Better (%)</th>
<th>No Preference (%)</th>
<th>ΔVMAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>tractor</td>
<td>1.8m/1.8m</td>
<td>73</td>
<td>10</td>
<td>18</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>65</td>
<td>16</td>
<td>19</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>70</td>
<td>14</td>
<td>16</td>
<td>0.3</td>
</tr>
<tr>
<td>old town</td>
<td>1.8m/1.8m</td>
<td>70</td>
<td>19</td>
<td>11</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>65</td>
<td>24</td>
<td>11</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>69</td>
<td>21</td>
<td>10</td>
<td>9.6</td>
</tr>
<tr>
<td>crowd run</td>
<td>1.8m/1.8m</td>
<td>56</td>
<td>18</td>
<td>26</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>34</td>
<td>46</td>
<td>20</td>
<td>-4.2</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>21</td>
<td>71</td>
<td>08</td>
<td>-11.4</td>
</tr>
<tr>
<td>aspen</td>
<td>1.8m/1.8m</td>
<td>61</td>
<td>11</td>
<td>28</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>59</td>
<td>24</td>
<td>18</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>56</td>
<td>16</td>
<td>28</td>
<td>11.7</td>
</tr>
<tr>
<td>controlled burn</td>
<td>1.8m/1.8m</td>
<td>65</td>
<td>14</td>
<td>21</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>58</td>
<td>20</td>
<td>23</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>59</td>
<td>18</td>
<td>24</td>
<td>15.5</td>
</tr>
<tr>
<td>rush field</td>
<td>1.8m/1.8m</td>
<td>71</td>
<td>06</td>
<td>23</td>
<td>12.3</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>68</td>
<td>14</td>
<td>19</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>61</td>
<td>19</td>
<td>20</td>
<td>9.3</td>
</tr>
<tr>
<td>touchdown pass</td>
<td>1.8m/1.8m</td>
<td>70</td>
<td>16</td>
<td>14</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>60</td>
<td>23</td>
<td>18</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>70</td>
<td>20</td>
<td>10</td>
<td>4.2</td>
</tr>
<tr>
<td>Average</td>
<td>1.8m/1.8m</td>
<td>67</td>
<td>13</td>
<td>20</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>1.8m/3.0m</td>
<td>58</td>
<td>24</td>
<td>18</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>iSIZE+HEVC 3.0m/VVC 3.0m</td>
<td>58</td>
<td>26</td>
<td>16</td>
<td>5.6</td>
</tr>
</tbody>
</table>
FAST IMPLEMENTATION

- 4Core/4Thread 8Gb Ubuntu 18.04 VM (Intel Xeon CPU)
- Under downscaling, we encode/decode a lower-res (downscaled) video.

- Runtime and MAC (both per frame) for perceptual precoding for 1080p input resolution:

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Time (ms)</th>
<th>Ratio</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3840x2160</td>
<td>284.07</td>
<td>x2</td>
<td>514M</td>
</tr>
<tr>
<td>3072x1728</td>
<td>256.74</td>
<td>x3/2</td>
<td>520M</td>
</tr>
<tr>
<td>1920x1080</td>
<td>175.17</td>
<td>x1</td>
<td>1150M</td>
</tr>
<tr>
<td>1280x720</td>
<td>60.44</td>
<td>x3/4</td>
<td>182M</td>
</tr>
<tr>
<td>960x540</td>
<td>64.95</td>
<td>x1/2</td>
<td>122M</td>
</tr>
<tr>
<td>768x432</td>
<td>58.09</td>
<td>x2/5</td>
<td>174M</td>
</tr>
<tr>
<td>640x360</td>
<td>49.38</td>
<td>x1/3</td>
<td>144M</td>
</tr>
</tbody>
</table>
OUR ROADMAP

iSIZE BitSave

1. Mobile & Wireless: **Save bandwidth**
   - Able to stream at lower signal strength (signal to noise ratio - SNR)
   - Improve user experience by decreasing fluctuation in video quality when wireless SNR changes
   - Improve data usage on capped connections
   - Improve device battery life by more efficient encoding/decoding

2. **Reduce CDN cost** for content providers
   - 30% less per media asset

3. **Improved live stream experience** and **longer battery life**

iSIZE UpScale

1. Mobile & Wireless: **Save bandwidth**

2. **Enhancing content** on the client side (HD→4K, 4K→8K)

3. **Enhancing camera capture** (HD camera upgraded to 4K via software)

4. **Enhancing digital zoom** with deep learning

iSIZE BitMind (Under development)

1. **Intelligent cropping of bitstreams** for extreme bandwidth and complexity savings when the “viewer” is a DeepNet

2. **1000x speedup** against conventional DeepNets in video

3. **20x reduction in bitrate** (video streams down to 3kbps!)

4. **Suitable for AI** on wearables, robotics, self-driving vehicles and UAVs

Backwards compatible / advanced machine learning
BEYOND BITSAVE AND THE CURRENT SOLUTIONS

FUTURE POTENTIAL AND PAPERS IN THE PUBLIC DOMAIN


Use compressed-domain info from the codec for superfast video tagging, action recognition, or fast event detection, e.g., see: https://arxiv.org/abs/1710.05112

Make further rate saving by determining the bitstream subsets needed by machine learning systems, e.g., stream only what a deep neural network needs from the video in order to carry out reliable analysis/classification, see our recent preprint: https://arxiv.org/abs/1810.03964

Use the speed and bandwidth efficiency of these solutions, to allow for real-time analytics on video for wearable cameras, robotic applications, etc.

Build compact video signatures based on codec bitstream data for retrieval of similar videos for recommendation services and linked data analytics
“QUALITY MEANS DOING IT RIGHT WHEN NO ONE IS LOOKING”

Henry Ford