DeViQ – A deep no reference video quality model

Steve Göring    Janto Skowronek    Alexander Raake

Audiovisual Technology Group, Technische Universität Ilmenau, Germany;
Email: [steve.goering, janto.skowronek, alexander.raake]@tu-ilmenau.de,

March 22, 2018
Motivation

▶ most internet traffic generated via video streaming providers [4]

▶ user's expectation: best possible video quality under every condition

▶ trending technologies: 4k/UHD, HDR, 360 degree, encoders, ...

▶ automated monitoring/optimization of perceived video quality

→ a brief look on current pixel based video/image quality models
Motivation

- most internet traffic generated via video streaming providers [4]
- user’s expectation: best possible video quality under every condition
  - trending technologies: 4k/UHD, HDR, 360 degree, encoders, …
  - automated monitoring/optimization of perceived video quality

→ a brief look on current pixel based video/image quality models
Motivation

- most internet traffic generated via video streaming providers [4]
- user’s expectation: best possible video quality under every condition
- trending technologies: 4k/UHD, HDR, 360 degree, encoders, ...
- automated monitoring/optimization of perceived video quality

→ a brief look on current pixel based video/image quality models
Motivation

▶ most internet traffic generated via video streaming providers [4]

▶ user’s expectation: best possible video quality under every condition

▶ trending technologies: 4k/UHD, HDR, 360 degree, encoders, ...

▶ automated monitoring/optimization of perceived video quality

→ a brief look on current pixel based video/image quality models
Motivation

- most internet traffic generated via video streaming providers [4]
- user’s expectation: best possible video quality under every condition
- trending technologies: 4k/UHD, HDR, 360 degree, encoders, ...
- automated monitoring/optimization of perceived video quality

→ a brief look on current pixel based video/image quality models
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/ technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

► full-reference models highly accurate to human perception [18]
  ○ e.g. Netflix’s VMAF [14] → reference video

► hand-crafted features [12, 14]
  ○ new encoders/ technologies → new artefacts → new features

► models using deep neural networks [3, 11, 8, 5, 6, 9]
  ○ patching to reduce input size → losses global connections; many patches for 4K
  ○ training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- **full-reference models highly accurate to human perception** [18]
  - e.g. Netflix’s VMAF [14]  →  reference video

- **hand-crafted features** [12, 14]
  - new encoders/ technologies  →  new artefacts  →  new features

- **models using deep neural networks** [3, 11, 8, 5, 6, 9]
  - patching to reduce input size  →  losses global connections; many patches for 4K
  - training requires per frame quality scores  →  huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/ technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
Video/image quality models

- full-reference models highly accurate to human perception [18]
  - e.g. Netflix’s VMAF [14] → reference video

- hand-crafted features [12, 14]
  - new encoders/technologies → new artefacts → new features

- models using deep neural networks [3, 11, 8, 5, 6, 9]
  - patching to reduce input size → losses global connections; many patches for 4K
  - training requires per frame quality scores → huge database

→ ideas for solving the identified research problems
How to solve the identified problems?

- huge training database for no-reference model:
  - generate ground-truth per frame data from full-reference model: VMAF [14, 10]

- hand-crafted features
  - using a pre-trained DNN for automatic feature extraction: inception-v3 [17]

- patching and global connection; many patches for 4K resolution
  - using hierarchical sub-images with larger block size: 299x299

→ introduce our model DeViQ (Deep Video Quality)
How to solve the identified problems?

- huge training database for no-reference model:
  - generate ground-truth per frame data from full-reference model: VMAF [14, 10]

- hand-crafted features
  - using a pre-trained DNN for automatic feature extraction: inception-v3 [17]

- patching and global connection; many patches for 4K resolution
  - using hierarchical sub-images with larger block size: 299x299

→ introduce our model DeViQ (Deep Video Quality)
How to solve the identified problems?

► huge training database for no-reference model:
  ◦ generate ground-truth per frame data from full-reference model: VMAF [14, 10]

► hand-crafted features
  ◦ using a pre-trained DNN for automatic feature extraction: inception-v3 [17]

► patching and global connection; many patches for 4K resolution
  ◦ using hierarchical sub-images with larger block size: 299x299

→ introduce our model DeViQ (Deep Video Quality)
How to solve the identified problems?

- huge training database for no-reference model:
  - generate ground-truth per frame data from full-reference model: VMAF [14, 10]

- hand-crafted features
  - using a pre-trained DNN for automatic feature extraction: inception-v3 [17]

- patching and global connection; many patches for 4K resolution
  - using hierarchical sub-images with larger block size: 299x299

→ introduce our model DeViQ (Deep Video Quality)
DeViQ – General approach

- (1) automatic feature extraction
  - pre-trained classification DNN
  - hierarchical sub-images: full, 1/2 of each dimension, 1/4 and 1/8 = 85 images
  - no-reference features; brisque+niqe [12, 13]

- (3) random forest model with (2) feature selection

- final quality score: mean value of each frame
DeViQ – General approach

1. Automatic feature extraction
   - Pre-trained classification DNN
   - Hierarchical sub-images: full, 1/2 of each dimension, 1/4 and 1/8 = 85 images
   - No-reference features; brisque+nique [12, 13]

2. Random forest model with feature selection

3. Final quality score: mean value of each frame
DeViQ—General approach

(1) automatic feature extraction
- pre-trained classification DNN
- hierarchical sub-images: full, 1/2 of each dimension, 1/4 and 1/8 = 85 images
- no-reference features; brisque+niqe [12, 13]

(3) random forest model with (2) feature selection

final quality score: mean value of each frame
all source videos: UHD-I (3840x2160); 60 fps (except sintel*); 10 s

train


validation

DeViQ – Evaluation – Conditions

- 3 codecs; 5 resolutions; 2/3 bitrates per resolution

→ encoded to 320 videos: train=50%; validation=50%; no overlap

- calculated VMAF scores for \( \approx 200k \) frames

- for validation: subjective test (22 participants; avg. age=26.7)

- comparison to retrained brisque+nique model/ full-reference metrics
DeViQ – Evaluation – Conditions

- 3 codecs; 5 resolutions; 2/3 bitrates per resolution

- → encoded to 320 videos: train=50%; validation=50%; no overlap

- calculated VMAF scores for ≈ 200k frames

- for validation: subjective test (22 participants; avg. age=26.7)

- comparison to retrained brisque+nique model/ full-reference metrics
DeViQ – Evaluation – Conditions

- 3 codecs; 5 resolutions; 2/3 bitrates per resolution

- → encoded to 320 videos: train=50%; validation=50%; no overlap

- calculated VMAF scores for ≈ 200k frames

- for validation: subjective test (22 participants; avg. age=26.7)

- comparison to retrained brisque+niqe model/ full-reference metrics
DeViQ – Evaluation – Conditions

- 3 codecs; 5 resolutions; 2/3 bitrates per resolution
- → encoded to 320 videos: train=50%; validation=50%; no overlap
- calculated VMAF scores for $\approx 200k$ frames
- for validation: subjective test (22 participants; avg. age=26.7)
- comparison to retrained brisque+niqe model/ full-reference metrics
3 codecs; 5 resolutions; 2/3 bitrates per resolution

encoded to 320 videos: train=50%; validation=50%; no overlap

calculated VMAF scores for \( \approx 200k \) frames

for validation: subjective test (22 participants; avg. age=26.7)

comparison to retrained brisque+niqe model/ full-reference metrics
DeViQ – Evaluation – Prediction vs. VMAF

average VMAF-scores with DeViQ and brisque+nique predictions

<table>
<thead>
<tr>
<th>method</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>pearson</th>
<th>kendall</th>
<th>spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>deviq</td>
<td>18.87</td>
<td>0.60</td>
<td>0.84</td>
<td>0.66</td>
<td>0.84</td>
</tr>
<tr>
<td>brisque+nique</td>
<td>19.75</td>
<td>0.56</td>
<td>0.85</td>
<td>0.64</td>
<td>0.83</td>
</tr>
<tr>
<td>vifp</td>
<td>22.28</td>
<td>0.44</td>
<td>0.58</td>
<td>0.46</td>
<td>0.63</td>
</tr>
</tbody>
</table>
DeViQ – Evaluation – Prediction vs. MOS

comparison of VMAF, DeViQ, brisque+nique to MOS values

<table>
<thead>
<tr>
<th>method</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>kendall</th>
<th>pearson</th>
<th>spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmaf</td>
<td>0.55</td>
<td>0.76</td>
<td>0.72</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>deviq</td>
<td>0.70</td>
<td>0.61</td>
<td>0.61</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>brisque+nique</td>
<td>0.81</td>
<td>0.47</td>
<td>0.53</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>vifp</td>
<td>0.86</td>
<td>0.41</td>
<td>0.52</td>
<td>0.70</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database
  → DeViQ (Deep Video Quality)

  - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database
  → DeViQ (Deep Video Quality)

  - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database
  → DeViQ (Deep Video Quality)

  - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database
  → DeViQ (Deep Video Quality)

  - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database → DeViQ (Deep Video Quality)
  - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Conclusion and Future Work

- identified main problems: hand-crafted features; patching; huge database
  - DeViQ (Deep Video Quality)
    - performs good compared to full-reference, no-reference models

- open points:
  - frame and sub-image selection
  - average for overall video quality

- DeViQ’s core idea: train a no-reference model based on a full-reference model using a pre-trained image DNN
Thank you for your attention

...... are there any questions?


References II


References III


### deviq, brisque+nique vs. vmaf

<table>
<thead>
<tr>
<th>method</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>pearson</th>
<th>kendall</th>
<th>spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>deviq</td>
<td>18.87</td>
<td>0.60</td>
<td>0.84</td>
<td>0.66</td>
<td>0.84</td>
</tr>
<tr>
<td>brisque+nique</td>
<td>19.75</td>
<td>0.56</td>
<td>0.85</td>
<td>0.64</td>
<td>0.83</td>
</tr>
<tr>
<td>vifp</td>
<td>22.28</td>
<td>0.44</td>
<td>0.58</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>msssim</td>
<td>48.99</td>
<td>-1.70</td>
<td>0.54</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>ssim</td>
<td>49.88</td>
<td>-1.80</td>
<td>0.48</td>
<td>0.44</td>
<td>0.60</td>
</tr>
<tr>
<td>psnrhvs</td>
<td>56.09</td>
<td>-2.55</td>
<td>0.33</td>
<td>0.52</td>
<td>0.72</td>
</tr>
</tbody>
</table>
### deviq, brisque+nique, vmaf vs. mos

<table>
<thead>
<tr>
<th>method</th>
<th>$RMSE$</th>
<th>$R^2$</th>
<th>cohen_d</th>
<th>kendall</th>
<th>pearson</th>
<th>spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmaf</td>
<td>0.55</td>
<td>0.76</td>
<td>0.24</td>
<td>0.72</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>deviq</td>
<td>0.70</td>
<td>0.61</td>
<td>0.19</td>
<td>0.61</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>brisque+nique</td>
<td>0.81</td>
<td>0.47</td>
<td>0.34</td>
<td>0.53</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>vifp</td>
<td>0.86</td>
<td>0.41</td>
<td>-0.34</td>
<td>0.52</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>msssim</td>
<td>1.70</td>
<td>-1.32</td>
<td>-1.72</td>
<td>0.46</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>ssim</td>
<td>1.74</td>
<td>-1.42</td>
<td>-1.76</td>
<td>0.45</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td>psnrhvs</td>
<td>2.27</td>
<td>-3.15</td>
<td>0.30</td>
<td>0.60</td>
<td>0.34</td>
<td>0.76</td>
</tr>
</tbody>
</table>
feature importance

for each 1000 feature values we summed the feature importance of our model; subimage 85=no-reference features