A No-Reference Autoencoder Video Quality Metric

Helard B. Martinez\textsuperscript{1}, Andrew Hines\textsuperscript{2}, and Mylène C.Q. Farias\textsuperscript{1}
\textsuperscript{1}University of Brasília, Brazil
\textsuperscript{2}University College Dublin, Ireland
http://www.ene.unb.br/mylene

ICIP, 24 September 2019, Taipei, Taiwan
Motivation and Goals

Audio-visual Content
Motivation and Goals

- Increase of types of MM services
- Quality of Experience (QoE) is an important aspect
- Tools to quantify the quality of MM experience
Design a NR pixel-based video quality metric

- Created a large audio-visual dataset (diverse content and degradations);
- Collected video quality responses using an immersive methodology;
- Extracted spatial-temporal features from the videos;
- Used Autoencoders algorithms to produce select the ‘best’ visual features;
- Mapped these visual features into subjective quality scores.
Subjective Experiments

Immersive Methodology (M. Pinson)

- Increase content diversity;
- Use longer videos, which convey an idea, with their audio;
- Keep the experiment interesting and reduce fatigue.
Apparatus and Physical Conditions

- Experiments divided into 3 sessions: Display, Training, Main;
- Scores collected (ACR scale, 5 points): $MQS_{HRC}$ - Mean Quality Score (HRC)
- Recording Studio: @University of Brasilia
- Desktop computer, LCD monitor, set of earphones, Sound card Asus Xonar DGX 5.1
- Viewing conditions: ITU Rec. BT.500
Dataset

- Distortions: Bitrate compression, Packet-Loss, and Frame-Freezing;
  - Video coding: H.264 and H.265 video codecs (200 to 16,000 kbs);
  - Packet Loss: loss rates 0.01 to 0.10
  - Frame freezing: pauses (# pauses 1 to 3 - Length 1s to 3s)

- HRC: Hypothetical Reference Circuit

- Packet-loss and frame-freezing did not happen simultaneously
### Dataset: Experiment 1

<table>
<thead>
<tr>
<th>Subjective Experiment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants:</strong></td>
<td>60</td>
</tr>
<tr>
<td><strong>Age Range:</strong></td>
<td>19 - 36</td>
</tr>
<tr>
<td><strong>Source Stimuli:</strong></td>
<td>60</td>
</tr>
<tr>
<td><strong>Test Conditions (HRC):</strong></td>
<td>12</td>
</tr>
<tr>
<td><strong>Test Sequences:</strong></td>
<td>720</td>
</tr>
<tr>
<td><strong>Average length:</strong></td>
<td>37 sec</td>
</tr>
<tr>
<td><strong>Temporal resolution:</strong></td>
<td>1280x720 (720p)</td>
</tr>
<tr>
<td><strong>Spatial resolution:</strong></td>
<td>30 fps</td>
</tr>
<tr>
<td><strong>Color space format:</strong></td>
<td>4:2:0</td>
</tr>
<tr>
<td><strong>Bit-depth:</strong></td>
<td>16 bits</td>
</tr>
<tr>
<td><strong>Sample frequency:</strong></td>
<td>48 kHz</td>
</tr>
<tr>
<td><strong>Audio codec:</strong></td>
<td>PCM</td>
</tr>
</tbody>
</table>
Deep Autoencoder Model

- Deep Autoencoders: select a visual features set with lower dimension and good description capacity;
- Training (k-fold):
  - **Global Feature Matrix**: dataset spatial and temporal features
  - **Global Target Matrix**: dataset quality scores target
- NR Video Quality Metric.
Feature Extraction

Visual Features (90 features)

- (88) Natural Scene Statistics features - (2014, Zhang)
- (2) Spatial and Temporal features - (2007, Ostaszewska)
- 90-by-n matrix (n: number of frames)
Model Training

Target Set

- Built using subjective scores, for training
- 4-by-n matrix: 4 quality groups (ACR) and n video frames
Model Training

Autoencoder Layer

- Input:
  - Global feature matrix
- Two autoencoders:
  - Features 1, AE1
  - Features 2, AE2
Model Training

Classification Layer
- Input:
  - Features 2
  - Global target matrix
- Softmax layer
  - Soft Net
Model Training

Deep Network Model

- Encoder1 (AE1)
  - Encoder2 (AE2)
  - SoftNet

- Trained with:
  - Global feature matrix
  - Global target matrix

DeepNet Model
Autoencoder 1

Autoencoder 2
Autoencoder 1

Autoencoder 2

DeepNet Model
### Model Training

#### Training Parameters

<table>
<thead>
<tr>
<th>Layer</th>
<th>Parameters</th>
<th>Audiovisual Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoencoder Layer</td>
<td>Input</td>
<td>90-by-N matrix</td>
</tr>
<tr>
<td></td>
<td>Layer size #1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Layer size #2</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Decoder transfer function</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td>L2 weighth regularization</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Sparsity Regularization</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Sparsity Proportion</td>
<td>0.05</td>
</tr>
<tr>
<td>Classification Layer</td>
<td>Input</td>
<td>20-by-N matrix</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4-by-N matrix</td>
</tr>
<tr>
<td></td>
<td>Loss Function</td>
<td>Cross Entropy</td>
</tr>
<tr>
<td>Additional Info</td>
<td>Training Set</td>
<td>Exp. 1 of Dataset</td>
</tr>
<tr>
<td></td>
<td># sequences</td>
<td>720</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>10-fold CV</td>
</tr>
</tbody>
</table>
Model Performance

Output Processing

<table>
<thead>
<tr>
<th>Index</th>
<th>0.8</th>
<th>0.9</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Max value →

<table>
<thead>
<tr>
<th>0.8</th>
<th>0.9</th>
<th>0.7</th>
<th>0.8</th>
<th>0.8</th>
<th>0.9</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
</table>

Index →

| 1 | 2 | 2 | 2 | 2 | 1 | 3 | 3 | 2 | 2 | 2 |

Max + Index →

| 1.8 | 2.9 | 2.7 | 2.8 | 2.8 | 1.9 | 3.6 | 3.7 | 2.8 | 2.8 | 2.9 |

Quality Score → 2.79
Model Performance

Metrics for comparison

- **Video:**
  - FR: SSIM, PSNR
  - NR: DIIVINE, VIIDEO, BIQI, NIQE, BRISQUE
NR Video Quality Metric

(PCC)

(SCC)
Model Performance

External Database: LiveNetflix-II

- 420 sequences AV Full-HD
- 15 source sequences, 7 network conditions, 4 bitrate adaptation strategies
NR Video Quality Metric - LiveNetflix-II

(PCC)

(SCC)
Conclusions:

- Used visual features to design a NR-VQ;
- Trained on a diverse content dataset;
  - Compression and transmission degradations;
  - Immersive subjective experiment;
- A Deep Autoencoder Model for quality assessment;
- Performed well, when compared to state-of-the-art metrics;

Future Work:

- Refine training parameters;
- Test additional descriptive features;
- Training on different databases.
Questions?
mylene@ene.unb.br,
http://www.ene.unb.br/mylene
http://www.ene.unb.br/mylene/databases.html