YouVQ: A new no-reference metric for UGC

Media Algorithms Team

Balu Adsumilli
Yilin Wang

VQEG, Dec. 2021
Agenda

01 What is UGC?

02 YouTube UGC Dataset

03 Introducing YouVQ
What is UGC?
YouTube video traffic

- 500 hours of video shared every minute
- >2B daily active users in 100+ countries in 80+ languages
- 70% of YouTube is watched on mobile devices
- ~1400 combinations of codecs, containers, resolutions, and formats
- most of the videos uploaded are user generated content
User Generated Content (UGC)

Content and emotion > narrative and quality
- Artifact-ridden: shaky cam, low light, portrait, overlays, heavily compressed
- Variability due to content creator, network, unusual viewing environment
Current Video Quality Metrics

Subjective
- Mean Opinion Score (MOS)
- Differential Mean Opinion Score (DMOS)

Objective
- Reference-based metrics: PSNR, SSIM, VMAF
  - Assumes a pristine original that the target should “get close to”
- No-reference metrics: banding, noise, NIQE
  - Does not depend on the original, pristine or otherwise

Are any of these good for UGC?
Trouble with existing notions

High Pixel Difference ≠ Low Perceptual Quality

Left image: greater MSE. Right image: much lower spatial frequencies. Human vision system has a stronger response to the lower spatial frequencies.
Need for accurate no-ref metric for UGC

Growing need for a reliable no-reference fidelity metric (not artifact)
- Original video is either not available or not a reference (not pristine)
  - same relative quality deltas map differently for varying original video quality

Original video: PSNR= 43.77, SSIM=0.969, VMAF=89.34
Transcoded video: PSNR= 43.77, SSIM=0.969, VMAF=89.34

Similar perceptual quality (DMOS~0)
UGC Video Quality Assessment

Foundational question: Will we need to rethink video quality metrics in the presence of non-pristine originals?

We start with a dataset
- Distributed across variations in content, complexity, resolutions, frame rates, formats
- Universal availability
- Ground truth subjective data
YouTube UGC Dataset (YT-UGC): media.withyoutube.com

- **1500** Uploaded videos
  - Sourced from 1.5 million uploads
  - 15 content categories
  - Each video in multiple resolutions, fps

- Ground truth (MOS) for all videos

- Added DMOS for popular categories

- Added 600+ content labels

Balu Adsumilli et al., "Launching a YouTube dataset of user-generated content", YouTube tech blog
Yilin Wang et al., "YouTube UGC Dataset for Video Compression Research", MMSP 2019
Joong Yim et al., "Subjective Quality Assessment for YouTube UGC Dataset", ICIP 2020
Yilin Wang et al., "Rich features for perceptual quality assessment of UGC videos", CVPR 2021
## Perceptual Quality Assessment Aspects

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Low quality</th>
<th>High quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Video Content</strong></td>
<td>MOS=2.052 (no meaningful content)</td>
<td>MOS=4.457 (intense games)</td>
</tr>
<tr>
<td><strong>Distortions</strong></td>
<td>MOS=1.242 (heavy blur)</td>
<td>MOS=4.522 (high contrast, sharp edges)</td>
</tr>
<tr>
<td>(Introduced during video production phase)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Video Compression</strong></td>
<td>MOS=2.372 (heavy compression)</td>
<td>MOS=4.646 (no compression artifacts)</td>
</tr>
<tr>
<td>(introduced by compression or transmission)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
UGC Video Quality Human Evaluation

Real-time strategy game (interesting content)

Blurred texture

Heavily compressed text

Conclusion
Medium low quality (MOS=2.761)

Explanation
Poor text and texture quality lead to bad game watching experience
How do we scale UGC evaluation?

Auto-evaluation from multiple aspects:
- Content
- Distortion
- Compression

Report quality beyond a single score - folding in multiple high level interpretable indicators
Requirements for UGC metric

Comprehensively map to human evaluations accurately, folding in all the nuances of UGC

Target UGC centric no-reference, while still perform reliably with reference

Introducing YouVQ - a VQ metric for UGC
YouVQ Framework
UGC Video Quality Assessment (UGC-VQA)

- Existing handcrafted feature approaches (SSIM, VMAF, etc)
  - Difficult and time-consuming
  - Insufficient feature set (summarized from limited samples)

- Current Machine Learning approaches
  - Automatic feature learning
  - Suitable for large scale UGC data

Direct training on UGC dataset:

UGC Dataset → ML model → Quality score
Training data for UGC video quality assessment

- UGC datasets with quality labels
  - YT-UGC (1.5K), Patch-VQ (40K)

- Compare with non-quality datasets
  - Kinetics-600 (500K videos), YT8M (8M videos), ImageNet (14M images)

- Transfer Learning - preferred
Direct Transfer Learning

Non-UGC Quality Related Pretraining

Backbones

UGC Dataset Fine-Tuning

Embeddings

Quality score
Direct Transfer Learning

Non-UGC Quality Related Pretraining

Backbones

Embeddings

UGC Dataset Fine-Tuning

Quality score

For recognition: similar

For video quality: very different
Retraining on quality related data

Non-UGC Quality Related Pretraining

UGC Quality Related Retraining

UGC Dataset Fine-Tuning

Quality score
### Effectiveness of UGC quality related retraining

Evaluated on YT-UGC MOS

<table>
<thead>
<tr>
<th>Backbone (EfficientNet-b0)</th>
<th>PLCC</th>
<th>SRCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw (ImageNet, frozen weights)</td>
<td>0.624</td>
<td>0.612</td>
</tr>
<tr>
<td>Raw (ImageNet, trainable weights)</td>
<td>0.671</td>
<td>0.690</td>
</tr>
<tr>
<td>Retrained (KADIS-700K, frozen weights)</td>
<td>0.732</td>
<td>0.735</td>
</tr>
<tr>
<td>Retrained (KADIS-700K, trainable weights)</td>
<td>0.732</td>
<td>0.738</td>
</tr>
</tbody>
</table>

PLCC, SRCC: correlation coefficients in [0, 1], the higher the better.

**Direct transfer learning**

**With quality related retraining**
YouVQ: YouTube Video Quality Assessment Framework

Inputs:
- ContentNet
- DistortionNet
- CompressionNet

Outputs:
- ContentNet
- DistortionNet
- CompressionNet

Chunk Features

AggregationNet

Video Quality Indicators:
- content labels
- distortion types
- compression level

+ Quality score

Yilin Wang et al., "Rich features for perceptual quality assessment of UGC videos", CVPR 2021
YouVQ: YouTube Video Quality Assessment Framework

Benefits of YouVQ framework:
- Self-supervised learning on raw UGC videos, no longer restricted by labeled MOS.
- Complementary features learned from different quality related aspects.
- Works on native resolutions, and sensitive to local details.
YouVQ: YouTube Video Quality Assessment Framework

Benefits of YouVQ framework:
- Self-supervised learning on raw UGC videos, no longer restricted by labeled MOS.
- Complementary features learned from different quality related aspects.
- Works on native resolutions, and sensitive to local details.
## YouVQ Features: ContentNet (CT)

- Multi-label classification
- Model trained on 100k YT8M videos
  - Inputs: single image
  - Outputs: 3862 UGC content labels
  - Loss: cross-entropy
- Backbone: EfficientNet-b0 (pre-trained on ImageNet)

### Backbone Models

<table>
<thead>
<tr>
<th>Backbone model</th>
<th>#Params</th>
<th>#FLOPS</th>
<th>YT8M Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-10</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>23.5M</td>
<td>3.8B</td>
<td>0.325 0.554 0.659</td>
</tr>
<tr>
<td>EfficientNet-b0</td>
<td>5.3M</td>
<td>0.39B</td>
<td>0.463 0.721 0.792</td>
</tr>
<tr>
<td>EfficientNet-b7</td>
<td>66M</td>
<td>37B</td>
<td>0.460 0.723 0.788</td>
</tr>
</tbody>
</table>

### Correlation on YT-UGC Quality Scores

<table>
<thead>
<tr>
<th></th>
<th>MOS</th>
<th>DMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content features only</td>
<td>0.628 0.615</td>
<td>0.584 0.357</td>
</tr>
<tr>
<td>Content+Compression</td>
<td>0.787 0.774</td>
<td>0.672 0.652</td>
</tr>
<tr>
<td>Content+Distortion</td>
<td>0.750 0.752</td>
<td>0.390 0.334</td>
</tr>
<tr>
<td>All three features</td>
<td>0.802 0.796</td>
<td>0.539 0.497</td>
</tr>
</tbody>
</table>

No gain when using EfficientNet-b7 feature for quality assessment.
YouVQ Features: DistortionNet (DT)

- **Synthetic distortions**
  - 23 types, e.g. Gaussian noise and motion blur
  - Distorted variants in 5 levels per type

- **Model trained on KADIS-700K images**
  - Inputs: two images with the same distortion type
  - Outputs: distortion type and level
  - Loss: cross-entropy + pairwise hinge loss

- **Backbone: EfficientNet-b0 (pre-trained on ImageNet)**

![Diagram of DistortionNet](https://via.placeholder.com/150)

**Diagram:**
- Image A → DistortionNet → Distortion type for A, Distortion level for A
- Image A' → DistortionNet → Distortion type for A', Distortion level for A'
- Cross-entropy loss
- Pairwise hinge loss (Assuming A should always have higher level than A')
YouVQ Features: CompressionNet (CP)

- Self-supervised learning
- Compressing original videos with recommended VP9 settings for VOD and Live
- Model trained on YT8M 1080p videos
  - Inputs: original and its VOD and Live versions
  - Outputs: compression level in [0, 1] + compression feature (last layer outputs)
  - Loss: pairwise loss + contrastive loss
- Backbone: D3D (pre-trained on Kinetics-600)

Pairwise loss:
\[(\text{Orig}>\text{VOD}, \text{Orig}>\text{Live}, \text{VOD}>\text{Live})\]

Contrastive loss:
\[
\text{sim} (\text{Orig}, \text{VOD}) / (\text{sim} (\text{Orig}, \text{Live}) + \text{sim} (\text{VOD}, \text{Live}) + \text{sim} (\text{Orig}, \text{VOD}))
\]

Note: sim() means feature similarity
YouVQ feature aggregation

- AggregationNet
  - Training with YouVQ features on YT-UGC original MOS
  - Three candidate aggregation models
    - AvgPool, LSTM, ConvLSTM
  - AvgPool performs best
    - most UGC videos have relatively consistent quality among frames

<table>
<thead>
<tr>
<th>Feature</th>
<th>AvgPool</th>
<th>LSTM</th>
<th>ConvLSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
<td>SRCC</td>
<td>RMSE</td>
</tr>
<tr>
<td>CP+CT+DT</td>
<td>0.802</td>
<td>0.816</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Correlations on YT-UGC MOS
### Accuracy

#### Correlations with YT-UGC MOS

<table>
<thead>
<tr>
<th>YouVQ Features</th>
<th>PLCC</th>
<th>SRCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP (Compression)</td>
<td>0.770</td>
<td>0.785</td>
</tr>
<tr>
<td>CT (Content)</td>
<td>0.628</td>
<td>0.628</td>
</tr>
<tr>
<td>DT (Distortion)</td>
<td>0.726</td>
<td>0.744</td>
</tr>
<tr>
<td>CP+CT</td>
<td>0.787</td>
<td>0.801</td>
</tr>
<tr>
<td>CP+DT</td>
<td>0.790</td>
<td>0.802</td>
</tr>
<tr>
<td>CT+DT</td>
<td>0.750</td>
<td>0.767</td>
</tr>
<tr>
<td><strong>CP+CT+DT</strong></td>
<td><strong>0.802</strong></td>
<td><strong>0.816</strong></td>
</tr>
</tbody>
</table>
Generalizability on MOS Prediction for UGC

<table>
<thead>
<tr>
<th></th>
<th>Model fine-tuned on YT-UGC MOS</th>
<th>Directly predicting on KoNViD-1k MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLCC (YouVQ)</td>
<td>0.802</td>
<td>0.670</td>
</tr>
<tr>
<td>PLCC (best of other evaluated metrics)</td>
<td>0.761 (from VSFA)</td>
<td>0.602 (from VSFA)</td>
</tr>
<tr>
<td>Metrics compared</td>
<td>BRISQUE, NIQE, VIIDEO, TLVQM(SVR), TLVQM(RFR), VSFA</td>
<td>TLVQM(SVR), TLVQM(RFR), VSFA</td>
</tr>
</tbody>
</table>
Generalizability on DMOS Prediction for UGC

Evaluated on YT-UGC DMOS (not re-trained)
- 189 originals + three VP9 variants

Pred DMOS = YouVQ(ref) - YouVQ(target)
- Sensitive to compression
- Good correlations without retraining

<table>
<thead>
<tr>
<th>Metric</th>
<th>PLCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.402</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.493</td>
</tr>
<tr>
<td>VMAF</td>
<td>0.401</td>
</tr>
<tr>
<td>LPIPS</td>
<td>0.524</td>
</tr>
<tr>
<td>TLVQM</td>
<td>0.276</td>
</tr>
<tr>
<td>VSFA</td>
<td>0.403</td>
</tr>
<tr>
<td>YouVQ</td>
<td>0.660</td>
</tr>
</tbody>
</table>
Comprehensive Quality Indicators

ContentNet
- top-10 label accuracy
  - 0.792 on YT8M
  - 0.53 on YT-UGC

DistortionNet
- evaluated on KADID-10K
- distortion classification accuracy 0.97

CompressionNet
- self-supervised learning
- high accuracy on predicting pairwise order of compression level

Content labels: Car (0.58), Vehicle (0.42), Sports Car (0.32), Motorsports (0.18), Racing (0.11)

Distortion types: Jitter (0.112), Color quantization (0.111), Lens blur (0.108), Denoise (0.107)

Compression level: 0.892 (high)
Locating Local Quality Issues

YouVQ provides patchwise quality assessment

Patch at time $t = 1$
compression level = 0.000

Patch at time $t = 2$
compression level = 0.904
How YouVQ works in practice

YouVQ diagnosis report

From ContentNet (CT)
Video game, Strategy video game, World of Warcraft, etc

From DistortionNet (DT)
Multiplicative noise, Gaussian blur, Color saturation, Pixelate, etc

From CompressionNet (CP)
0.559 (medium high compression)

Predicted quality score in [1, 5]
(CP, CT, DT): (3.151, 3.901, 3.216)
(CP+CT+DT): 3.149 (medium low quality)
Summary

We introduced YouVQ for UGC video quality assessment

- It is a comprehensive framework to analyze UGC video quality and makes the VQ score more interpretable
- Maps very well to ground truth human evaluations
- Performs consistently and reliably for no-reference, works equally well when reference is present (pristine or otherwise)

Videos and subjective data are available on media.withyoutube.com
Thank you!