CHALLENGES IN CLOUD BASED INGEST AND ENCODING FOR HIGH QUALITY STREAMING MEDIA

Anne Aaron [*], *Zhi Li* [*], *Megha Manohara* [*], *Joe Yuchieh Lin* [†], *Eddy Chi-Hao Wu* [†], *and C.-C Jay Kuo* [†]

[*] Netflix Inc.,100 Winchester Circle Los Gatos, CA, United States
[†] University of Southern California, Ming Hsieh Department of Electrical Engineering, 3740 McClintock Avenue, Los Angeles, CA, United States

ABSTRACT

The Netflix ingest and encoding pipeline is a cloud-based platform that generates video encodes for the Netflix streaming service. Due to the large throughput of the system, automated video quality assessment of the source videos and the generated encodes is essential in ensuring the quality of experience of viewers. This paper discusses the motivations for integrating video quality assessment in the production pipeline, outlines currently deployed solutions and presents the technical challenges in improving the system.

Index Terms— Video, streaming, quality, cloud

1. INTRODUCTION

Netflix has grown significantly in global reach, subscribers and device support over the last several years. The streaming service launched in 2007 with a couple thousand titles and an Internet Explorer plugin hosting Windows Media Player [1]. Fast forward to 2015 and Netflix has more than 57 million subscribers [2] and tens of millions of active devices covering smart TV's, game consoles, set-top boxes, computers, tablets, and smart phones. Netflix streaming accounts for more than one third of peak North American download traffic [3]. In 2014 alone, Netflix increased its subscriber base by 13 million new members, started ingesting and streaming 4K UHD videos and was an early adopter of HEVC encoding support in the cloud [1, 2].

With the growth of the service comes a larger influx of titles that require processing by the system and more video stream representations generated per title. The primary challenge is to implement an ingest and encoding pipeline that is highly robust and scalable. The production system should be designed such that it can easily scale and support the demands of the business (i.e., more titles, more encodes, shorter time to deploy), while guaranteeing a high quality of experience for the customer.

For both ingest and encoding, automated video quality assessment plays a vital role in ensuring the quality of the Netflix streams. Due to the high throughput of the pipeline, manual visual inspection of all the sources and encodes is not inherently scalable nor economically feasible. Netflix brings in thousands of titles from content partners each year. For source inspection, only no-reference quality algorithms can be applied, since the system has no knowledge of a "correct" reference. For post-encoding inspection, full-reference or partial-reference video quality assessment can be utilized with the video source as the reference. This paper defines the challenges, outlines current deployed solutions and presents open problems on quality assessment for a cloud-based ingest and encoding pipeline.

2. SYSTEM OVERVIEW

The Netflix video encoding pipeline is a cloud-based processing system that ingests high quality video sources and processes each source into video encodes of various codec profiles, at multiple quality representations per profile. The encodes are packaged then deployed to a content delivery network for streaming. During a streaming session, the client requests its supported encode profile and adaptively switches between quality levels based on the network conditions.

Fig. 1 depicts a high-level diagram of the ingest and encoding pipeline. During ingest the system inspects a video source to 1) detect content that could lead to a bad viewing experience and 2) generate metadata required by the encoding pipeline. If the inspection deems the source unacceptable, the system automatically informs the content vendor about issues and requests a redelivery of the source.

After a source is successfully ingested, it is handed over to the video encoding block together with the metadata generated during inspection. The video is transcoded into various encodes of VC1, H.263, H.264/AVC, and HEVC at bitrates ranging from to 100 kbps to 16 Mbps. Inspection algorithms are then applied to the encode to validate the correctness and quality of the stream.

In Fig. 1, source inspection and video encoding are depicted as single blocks. In practice, each of these operations occur on multiple worker machines in the cloud and segments of the video are processed in parallel. This decreases the end-to-end processing delay, reduces the required local storage and improves error robustness of the system (if a machine is abruptly terminated, only a small portion of the work is lost). One of the first operations performed during source ingest is

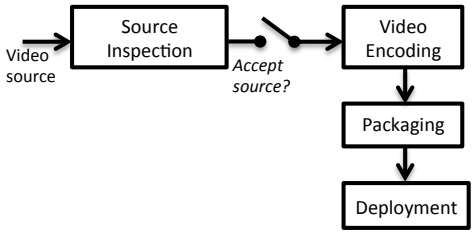

Fig. 1. System overview of Netflix pipeline

generating an index of the source file such that each inspection and encoding worker can use the index to correctly download and process any specified range of video frames. To generate the final results, an inspection aggregator is employed to combine the data from inspection workers. Similarly, multiple encode workers encode different chunks of the source. An assembler concatenates the transcoded segments of video into the complete video stream.

3. ASSESSMENT OF SOURCE VIDEO QUALITY

Netflix ingests source videos from content partners after a licensing contract is set-up. In some cases, the delivered source video contains distortion or artifacts which would result in bad quality video encodes – garbage in means garbage out. These artifacts may have been introduced by multiple processing and transcoding steps before delivery, data corruption during transmission or storage, or human errors during content production. Instead of trying to fix the source video issues after ingest (for example, apply error concealment to corrupted frames or re-edit sources which contain extra content), Netflix rejects the problematic source video and requests for redelivery. Rejecting problematic sources ensures that:

- The best source video available is ingested into the system. In many cases, the error mitigation techniques only partially fix the problem.
- Unnecessary complex algorithms (which could have been avoided by better processes upstream) do not burden the Netflix ingest pipeline.
- Content partners are motivated to triage their production pipeline and address the root causes of the problems. This will lead to improved video source deliveries in the future.

3.1. Source Video Impairments

The following are examples of impairments in the video source. If a problematic source is ingested into the pipeline, these artifacts would be passed on to the encodes, and potentially intensified by the encoding process.

Scaling artifacts. Downsampling to a lower resolution or upsampling to a higher resolution can lead to scaling artifacts in the source video. The degree of impairment in the source would depend on the scaling factors employed as well as the level of sophistication of the scaling algorithm. Scaling artifacts include blurring and ringing around edges.

Compression artifacts. A source video delivered to Netflix may have undergone transcoding to a bitrate resulting in compression artifacts such as blocking, ringing around edges, contouring and loss of detail because of high quantization. Compression artifacts are not only annoying to the viewer, but also waste bits when the encoder tries to preserve these artifacts during encoding.

Corrupted frames. The source video may be delivered with corrupted frames. The data corruption could be in the bitstream of the source video, i.e., a decoder would detect a non-compliant bitstream. In some cases, the source video bitstream is perfectly valid, but the pixel-domain frames exhibit corruption. For example, a video file is corrupted during transmission, the file is decoded and errors are concealed, and the video is re-encoded into the final file delivered to Netflix. Corrupted frames have varying levels of severity. The corruption may affect a few pixels, a single macroblock, multiple blocks or multiple successive frames.

Non-native frame rate. The Netflix delivery specifications require that source videos are delivered in their native frame rate in order to preserve original artistic content and avoid temporal conversion artifacts. One commonly observed frame rate conversion is 3:2 pulldown. 24 frames per second (fps) film content is converted to 29.97 fps video by duplicating fields of the original source at a regular interval. Visually, this leads to jerky motions in slow smooth scenes, also referred to as telecine judder.

Insertion of extra content. Extra frames in the encoded video, such as color bars, advertisements, slates, placards, and commercial blacks, can negatively affect customer experience. Netflix specifies that they be removed from the source video but human or process errors in the content production could lead to deliveries that do not meet this requirement.

3.2. Current Inspections and Open Problems

The main goal of the source inspection stage is to detect the video source impairments, examples of which are enumerated in Section 3.1. The challenge is to tune the algorithms to improve the detection rate while keeping false positives at a minimum. If an inspection has a high false positive rate, the source video rejection cannot be routed automatically to the content partner, and manual verification is required. This contradicts the goal of reducing human quality control in the ingest pipeline.

Corrupted frames are detected in the ingest pipeline by simply decoding the source video and monitoring errors coming from the decoder. This is low-complexity with zero false positive rate unless there are decoder bugs. However, this method cannot detect pixel-domain corruption nor does it provide a good indication of the severity of the corruption.

To address pixel-domain corruption, as well as detect unacceptable scaling or compression artifacts in the source video, no-reference video quality metrics are under investigation [4, 5]. These techniques utilize spatio-temporal

natural scene statistic models or knowledge of the possible distortions in the video, or a combination of the two. In integrating no-reference video quality metrics into the pipeline, the following aspects are being considered - applicability to a diverse set of content, complexity of implementation and reliability in measuring severity of perceptual distortion.

To detect frame rate conversion based on frame or field duplication, the source inspection calculates the difference of a pixel from the co-located pixel in the previous frame and averages the values for a given field. By analyzing the field differences of adjacent fields, duplicated fields are detected. If a consistent cadence (for example, one duplicated field every five fields in the case of 3:2 pulldown) is observed for a significant percent of the video, the video source is rejected.

Erroneously added segments of black frames, either at the start or tail of the program, or within the program to denote insertion of commercials, can easily be detected by inspecting the histogram of the pixel values of a frame. In some cases, actual content (for example, long dark scenes) can trigger a false positive. The image data is combined with information from the audio track to reduce false positives.

4. ASSESSMENT OF ENCODING QUALITY

While source inspection at the ingest aims to reject bad video sources, encoding quality assessment aims to predict the perceptual quality of video representations generated by the encoding pipeline. This section discusses the requirements on ideal assessment algorithms that can be deployed in Netflix's production system, and their potential use cases.

4.1. Encoding Impairments

As Netflix video streams are delivered over the top of Transport Control Protocol (TCP), Internet packet losses are shielded from causing impairments in the video decoded on subscriber devices. The main source of quality impairments thus comes from video encoding, whose main goal is to reduce the video data size. In video encoding, there are mainly two types of artifacts that causes quality degradation:

Scaling artifacts. Similar to the scaling artifacts presented in the source video, but only that they are now generated by Netflix's native encoding pipeline.

Compression artifacts. Similar to the compression artifacts in the source video, but may be much more prominent because the encoded videos to be passed to the streaming pipeline have much lower bitrates.

Typically, the encoding pipeline combines both scaling and lossy compression to reduce the video data size. For a target bitrate, tweaking the parameters of both operations can yield perceptual quality optimization.

System bugs. Besides the two types of artifacts above, quality degradation in the decoded video could also be the result of bugs in the encoding pipeline. These cases must be detected before the defective videos leak to the streaming pipeline, and the bugs must be corrected.

Type	JND	PSNR (dB)
TV Drama	0	43.26
Action Movie	0.2	41.91
Animation 1	0	46.86
Animation 2	0	37.55

Table 1. Encode vs. source video

Type	Bitrates (Kbps)	JND	ΔPSNR (dB)
Animation	5800 vs. 2350	0.1	13.25
TV Drama	5800 vs. 2350	1.2	0.21

Table 2. Comparison of encode pairs

4.2. Implementation Considerations

Video source characteristics. Netflix carries a large diverse collection of movie and TV show titles. An ideal quality assessment algorithm must predict the perceptual quality for a wide variety of source characteristics. To illustrate the challenge, Table 1 and Table 2 show examples where a traditional quality metric, peak signal-to-noise ratio (PSNR), yields poor prediction of the perceptual quality. For several titles PSNR scores are compared with just noticeable difference (JND) scores measured through subjective tests. Both tables show that PSNR correlates badly with the subjective JND scores.

Availability of reference video. While full-reference quality assessment is more reliable, partial-reference assessment can sometimes be used to simplify the system design.

Computational complexity. Owing to its design for scalability, Netflix's encoding pipeline is capable of accommodating quite substantial computations performed on each of the video chunks. Thus, while ideally an encoding quality assessment algorithm should have as low complexity as possible, an algorithm with moderate complexity is also acceptable.

4.3. Use Cases

The following highlights several use cases for an encoding quality assessment algorithm in the Netflix media pipeline.

Quality assurance. Similar to source inspection, quality assurance aims to reject bad encodes caused by bugs or bad parameter choice. For each new encode, two types of tests are performed to detect quality drop: 1) regression test, which compares the new encode with a previous encode; 2) post-encode test, which compares the new encode with the original source. If it is detected that the quality drop is beyond a threshold, the new encode must be rejected. In the past, this helped us identify bugs in our production platform.

Quality monitoring. For a Netflix title, each of its many encoded video representations can be associated with scores generated by the quality assessment algorithm. When a Netflix subscriber streams a title, the bitrate selection and the associated quality scores are automatically recorded over time. The average score can be used to gauge the general satisfaction of subscribers.

Optimizing encoding parameters. For a target bitrate, there are typically several encoding parameters, if tweaked, can yield optimized perceptual quality. Compression quantization parameter (QP) and downsampling resolution are example parameters that can be refined. The challenge is to

accurately model the perceptual quality as a function of these parameters, before numerical optimization can be applied. Besides, other encoding choices such as the number of representations and their bitrate spacing, can also be determined based on the model.

Optimizing streaming bitrate selection. Netflix's streaming algorithms match the streaming bitrate to a viewer's network speed. But streaming bitrate may not always correlate with perceptual quality. By examining quality scores within a future horizon and re-allocating bits among the video segments, it is possible to improve the video quality perceived by viewers. For example, a static scene may not need as many bits as the subscriber's network speed could supply. This surplus could instead be used in a dynamic scene a minute away from the static scene to yield improved viewing experience.

Codec and processing technology evaluation. An encoding quality assessment algorithm can evaluate current and new codecs and processing technologies, thus help decide if to incorporate these technologies into the production pipeline. It can also drive decisions for future codec design.

5. VMAF

This section describes a newly developed video quality assessment algorithm, motivated by the needs of the Netflix media pipeline, and discusses initial results. Video Multimethod Assessment Fusion (VMAF) is a full-reference perceptual video quality metric that aims to approximate human perception of video quality. This metric is intended to be useful as an absolute score across all types of content, and focused on quality degradation due to rescaling and compression.

VMAF estimates the best perceived quality score by computing scores from multiple quality assessment algorithms, and fusing them using a support vector machine (SVM) [6, 7]. Currently, three image fidelity metrics and one temporal signal have been chosen as features to the SVM.

Anti-noise SNR (ANSNR). ANSNR mitigates some drawbacks of simple SNR for film-grained content. A weak low-pass filter is applied to the source and a stronger low-pass filter is applied to the encode before the SNR calculation. It is good for detecting compression and strong scaling artifacts, but is not sensitive to quality changes for high quality videos.

Detail loss measure (DLM) [8]. DLM estimates the blurriness component in the distortion signal using wavelet decomposition. It uses contrast sensitivity function (CSF) to model the human visual system (HVS), and the wavelet coefficients are weighted based on CSF thresholds. It can detect blurriness in mid quality ranges well, but not that well for discriminating higher quality ranges.

Visual information fidelity (VIF) [9]. VIF quantifies the Shannon information shared between the source and the distortion relative to the information contained in the source itself. A gaussian scale mixture in the wavelet domain is used to model the source, and signal gain and additive noise in wavelet domain is used to model the distortion. HVS is modeled as a dual to the source, along with additive white gaussian noise to model internal neural noise. VIF is good for detecting blurring artifacts, but is insensitive to blocking.

Metric	Pearson	Spearman
VMAF	0.926	0.927
PSNR	0.623	0.710

Table 3. Performance of VMAF and PSNR in Pearson and Spearman correlation

Motion information. Motion in videos is chosen as the temporal signal, since the HVS is less sensitive to quality degradation in high motion frames. The global motion value of a frame is the mean co-located pixel difference of a frame with respect to the previous frame. Since noise in the video can be misinterpreted as motion, a low-pass filter is applied before the difference calculation. It is used in the SVM mapping as well as for adjustments after the mapping if the motion value exceeds a threshold.

Currently, the SVM model is trained using NFLX-V training set with 18 1920x1080 source clips of 6 second length. The training set was chosen to cover a wide range of high level features (animation, sports, indoor, camera motion, face close-up, people, water, obvious salience, object number) and low level characteristics (film grain noise, brightness, contrast, texture, motion, color variance, color richness, sharpness). From the 18 source clips, 152 encodes were generated using the x264 encoder. The lowest quality clips were encoded at 384x288, 150 kbps and the highest quality clips were encoded at the original resolution with bit rate 2 - 20 Mbps depending on the source clip. Each encode was played side-by-side with the source clip and scored from 0 to 100 (with 100 being of the highest quality) by 4 video experts.

As a simple validation test for the VMAF metric, the quality scores of 25% of the NFLX-V Training Set were predicted by training the SVM using the remaining 75%. Table 3 lists the Pearson and Spearman correlation of the VMAF scores with respect to the human scores and compares the results with the correlation values for PSNR.

6. CONCLUSION

This paper presented an overview of the Netflix cloud-based video ingest and encoding pipeline, defined the challenges and outlined current deployed solutions for incorporating automated video quality assessment into the production system. Initial results for the VMAF full-reference quality metric were discussed but more comprehensive testing of the algorithm is underway. Other promising no-reference and full-reference quality assessment schemes are under evaluation. Given the diversity of the Netflix content, this system is a rich and challenging platform for validating and improving state-of-the-art video quality assessment algorithms, with the ultimate goal of ensuring the best of quality of experience for subscribers.

7. REFERENCES

[1] D. Ronca, "Netflix's video workflow: Transcoding, codec's and 4k streaming," Invited talk at Streaming Media East 2014, New York, http://www.streamingmedia.com/ConferenceVideos.

[2] Netflix, "Q4 2014 letter to shareholders," January 2015, http://ir.netflix.com/.

[3] Sandvine, "Global internet phenomena report: 2H 2014," November 2014, https://www.sandvine.com/trends/global-internet-phenomena/.

[4] M.A. Saad, A.C. Bovik, and C. Charrier, "A DCT statistics based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun 2010.

[5] S. Gabarda and G. Cristobal, "Blind image quality assessment through anisotropy," *J Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 24, no. 12, pp. B42–B51, Dec 2007.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, May 2011.

[8] S. Li, F. Zhang, L. Ma, and K.N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 935–949, Oct 2011.

[9] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, Feb 2006.