Quality Assessment of Adaptive Bitrate Videos using Image Metrics and Machine Learning

Søgaard, Jacob; Forchhammer, Søren; Brunnström, Kjell

Published in:
Proceedings of Qomex 2015

Link to article, DOI:
10.1109/qomex.2015.7148105

Publication date:
2015

Document Version
Peer reviewed version

Citation (APA):
Søgaard, J., Forchhammer, S., & Brunnström, K. (2015). Quality Assessment of Adaptive Bitrate Videos using Image Metrics and Machine Learning. In *Proceedings of Qomex 2015* IEEE. https://doi.org/10.1109/qomex.2015.7148105

Quality Assessment of Adaptive Bitrate Videos using Image Metrics and Machine Learning

Jacob Søgaard[1], Søren Forchhammer[1], and Kjell Brunnström[2,3]
[1] Technical University of Denmark, Kgs Lyngby, Denmark
[2] Acreo Swedish ICT AB, Kista, Sweden
[3] Mid Sweden University, Sundsvall, Sweden

***Abstract*—Adaptive bitrate (ABR) streaming is widely used for distribution of videos over the internet. In this work, we investigate how well we can predict the quality of such videos using well-known image metrics, information about the bitrate levels, and a relatively simple machine learning method. Quality assessment of ABR videos is a hard problem, but our initial results are promising. We obtain a Spearman rank order correlation of 0.88 using content-independent cross-validation.**

I. Introduction

In recent years the amount of video traffic over the internet has grown. HTTP Adaptive Streaming (HAS) is a popular method for delivering video over the internet. It adapts the video to the current network conditions, server load and end user device capabilities. Video Quality Assessment (VQA) is an important tool to ensure the Quality of Experience (QoE) for video delivery. QoE for ABR videos is not well understood and by extension VQA of ABR videos is a hard problem [1]. In this work we consider the effects of ABR as reflected in the LIVE mobile database [2], apart from frame freeze. We present some initial results of an early-stage VQA tool for ABR videos. The approach in this paper is to use well-known image quality metrics to assess the frames of a video and using different pooling methods, calculate several features for the video. Additionally we use simple information about the bitrate levels of the video i.e. the number of increasing and decreasing steps in the video. Finally, we use a relatively simple machine learning method to map the features to a quality score for the video. For related work see e.g. [3], [4].

II. Image Metrics and Video Features

We use three different image metrics to assess the quality of the individual frames in a video: PSNR, SSIM [5], and MS-SSIM [6]. For SSIM and MS-SSIM we use the default parameters. To define quality-relevant features for the video we use temporal pooling on the image metrics scores over the frames as described in this section. The objective image metrics are used separately.

As a measure of the average quality and the variation hereof, we calculate the mean μ and standard deviation σ of the objective metric scores. Since the perception of quality also depends on recency [1], we also calculate the mean of the metric scores corresponding to the first and last 2 seconds of the video, respectively. Finally, inspired by [7] we divide the objective scores into clusters depending on the image metric scores using the method from [8], such that we obtain clusters of varying quality levels. Then a weighted average of the objective score from the clusters with lowest μ_L and highest means μ_H are calculated as:

$$\tilde{\mu} = \frac{\sum_{i \in C_L} S_i + w \sum_{i \in C_H} S_i}{|C_L| + w|C_H|} \tag{1}$$

where S_i are the subjective scores indexed by i, C_L and C_H are the set of clusters with lowest and highest mean, respectively. $|\cdot|$ denotes the size of each set. The weight w is defined as:

$$w = \left(\frac{\mu_L}{\mu_H}\right)^2 \tag{2}$$

In this way, the low objective scores will carry more weight in sequences where there is a large difference between good and bad quality. The weight w is also used as a video feature, since it carries information about the difference in quality levels in the video. Thus, a total of 6 features based on the image metric is produced. Note, that in order to ensure the features to be in the interval $[0;1]$ for numerical reasons, we rescale the PSNR-values to this interval using predetermined threshold values for minimum and maximum PSNR. Additionally, 2 features based on the bitrate levels in the video are also produced. They are simply defined as the number of increasing steps and decreasing steps of bitrate levels per second. The proposed Full-Reference (FR) model is thus defined by these 8 features and the mapping outlined in Section III.

III. Machine Learning

We use the method known as the Elastic Net (EN) to map the features to a quality score. The goal is to estimate the coefficients β of a regularized linear regression model:

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}}\ ||y - X\beta||^2 + \lambda_2||\beta||^2 + \lambda_1||\beta||_1 \tag{3}$$

where y is the target quality values, X is a feature matrix with rows of feature vectors, and λ_1 and λ_2 are regularization parameters of the $\mathbf{L}^1$-norm and the $\mathbf{L}^2$-norm, respectively. Due to the $\mathbf{L}^1$-norm in (3) the solution of an EN can generally be considered to be sparse and therefore feature selection is inherently a part of the EN method. For more information about the EN method we refer to [9]. In our experiments we use the implementation of the EN presented in [10].

IV. Results

To test the performance of our methods we use the LIVE mobile database [2] (which consists of 200 distorted videos). Since we do not consider freezing in our assessment, videos with this kind of artifact is removed from the dataset (in

TABLE I. CROSS-VALIDATION PERFORMANCE.

	VQ - PSNR			VQ - SSIM			VQ - MS-SSIM		
	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ
SROCC	**0.88**	0.83	0.11	0.86	**0.85**	0.10	**0.87**	**0.85**	**0.099**
LCC	**0.87**	**0.85**	**0.075**	0.83	0.82	0.11	0.83	0.82	0.10
RMSE	**0.57**	**0.61**	0.14	0.67	0.66	**0.11**	0.66	0.67	0.14
OR	**0.41**	**0.40**	**0.13**	0.44	0.47	**0.13**	0.47	0.48	0.14

TABLE II. TRAINING PERFORMANCE.

	VQ - PSNR	VQ - SSIM	VQ - MS-SSIM
SROCC	**0.82**	0.80	0.80
LCC	**0.81**	0.76	0.76
RMSE	**0.54**	0.60	0.60
OR	**0.33**	0.43	0.44

total 40 videos). Therefore, the dataset used consist of 160 encoded videos, made from 10 original source videos and with the following type of degradations: compression, wireless packet-loss, rate-adaption (switching bit-rates), and temporal dynamics (stepwise increasing/decreasing bit-rates). For more information on the sequences we refer to [2].

We use the following measures to report the performance: the Spearman Rank Order Correlation Coefficient (SROCC), the Linear Correlation Coefficient (LCC), the Root Mean Square Error (RMSE), and the Outlier Ratio (OR) [11]. We use content-independent cross-validation to find the optimal parameters of the EN and to measure the cross-validation performance. The cross-validation is performed by leaving out 2 contents for validation and repeat this for every possible content-independent split of training and validation. The performance of our approach (Secs. II and III) measured by the median $\tilde{x}$, mean μ, and standard deviation σ is reported in Table I. We denote our Video Quality model as VQ followed by the objective image metric the features are based on. For further validation we use the estimated optimal parameters found in the cross-validation to train an EN model on all of the data. The training performance of this model is reported in Table II. Since we are using the parameters from the cross-validation and the EN is a sparse model, the risk of overfitting is low, even when measuring the training performance. A scatter plot of the VQ training predictions based on PSNR is shown in Fig. 1 (contents marked by labels in [2]).

The correlation values are slightly higher for the cross-validation than for the training performance, which might indicate that our cross-validation correlation performance is too optimistic in regards to what can be expected from a test performance measured on a similar and independent dataset. However, if e.g. we only use the mean PSNR, a mean SROCC performance of 0.76 over the same splits as in the cross-validation is obtained, while the SROCC over the whole dataset is only 0.69 compared to 0.83 and 0.82 with the proposed method based on PSNR features. The EN is a sparse model and features might therefore be discarded during the training process. When based on PSNR, no features were discarded, while for MS-SSIM the mean objective score for the first 2 seconds, the weighted cluster mean, and the number of increasing bitrate steps were excluded. For SSIM the standard deviation of the image metric scores were additionally discarded. Thus, even though the performance of the models based on PSNR has slightly higher performance, the other models might be preferable since they contain a lower number of features.

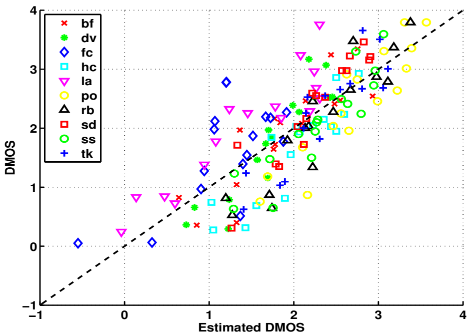

Fig. 1. Scatter plot of the VQ predictions (ML with features based on PSNR).

V. CONCLUSION

In this paper we have developed a FR model for predicting the quality of video in an ABR setting. Though the model is still in early-stage, we get promising results with a cross-validation correlation up to 0.88. There is still room for improvement such as adding more sophisticated video features and taking freezing events into account. Using our machine learning approach, we can also evaluate some features as less relevant, such as the mean objective quality score of the first 2 seconds of a video. It would be interesting to consider how the knowledge gained from this work could be used in building a NR model for the ABR scenario. Lastly, as future work our method should be tested on an independent dataset, to measure the test error performance in this case.

REFERENCES

[1] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, “A survey on quality of experience of HTTP adaptive streaming,” *IEEE Commun. Surveys Tuts.*, 2015.

[2] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, “Video quality assessment on mobile devices: Subjective, behavioral and objective studies,” *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, 2012.

[3] J. Lievens, A. Munteanu, D. De Vleeschauwer, and W. Van Leekwijck, “Perceptual video quality assessment in HTTP adaptive streaming,” in *IEEE Int'l Conf. on Consumer Electronics (ICCE)*, 2015, pp. 72–73.

[4] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, “To pool or not to pool: A comparison of temporal pooling methods for HTTP adaptive video streaming,” in *Fifth Int'l Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 52–57.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *IEEE Asilomar Conf. on Signals, Systems and Computers*, 2003.

[7] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, “Video quality pooling adaptive to perceptual distortion severity,” *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, 2013.

[8] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure.” ACM Press, 1999, pp. 49–60.

[9] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Stat. Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[10] K. Sjöstrand, L. H. Clemmensen, R. Larsen, and B. Ersbøll, “SpaSM: A matlab toolbox for sparse statistical modeling,” *Journal of Stat. Software*, 2012.

[11] *Recommendation ITU-T P.1401: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, Int'l Telecom. Union Std., 2012.