A Multi-category Task for Bitrate Interval Prediction with the Target Perceptual Quality

Zhenwei Yang[1], Liquan Shen[2*]
[1] School of Communication and Information Engineering, Shanghai University,
Shanghai, 200072, China
[e-mail: yzw_shcm@163.com]
[2] Shanghai Institute for Advanced Communicationand Data Science, Shanghai University,
Shanghai 200072, China
[e-mail: jsslq@shu.edu.cn]
*Corresponding author: Liquan Shen

*Received June 26, 2021; revised November 14, 2021; accepted November 27, 2021;
published December 31, 2021*

Abstract

Video service providers tend to face user network problems in the process of transmitting video streams. They strive to provide user with superior video quality in a limited bitrate environment. It is necessary to accurately determine the target bitrate range of the video under different quality requirements. Recently, several schemes have been proposed to meet this requirement. However, they do not take the impact of visual influence into account. In this paper, we propose a new multi-category model to accurately predict the target bitrate range with target visual quality by machine learning. Firstly, a dataset is constructed to generate multi-category models by machine learning. The quality score ladders and the corresponding bitrate-interval categories are defined in the dataset. Secondly, several types of spatial-temporal features related to VMAF evaluation metrics and visual factors are extracted and processed statistically for classification. Finally, bitrate prediction models trained on the dataset by RandomForest classifier can be used to accurately predict the target bitrate of the input videos with target video quality. The classification prediction accuracy of the model reaches 0.705 and the encoded video which is compressed by the bitrate predicted by the model can achieve the target perceptual quality.

Keywords: Perceptual coding, Bitrate prediction, Rate control, Machine Learning, Feature Extraction.

1. Introduction

With the rapid development of social media, massive videos with various content and form are uploaded to the video platforms. For video service providers, they must face challenges of various network conditions and play devices on the client sides. It is desirable to know the bitrate demand for different-content videos with different quality, because it is directly related to the streaming cost of the platform as well to a better streaming media service delivered to users.

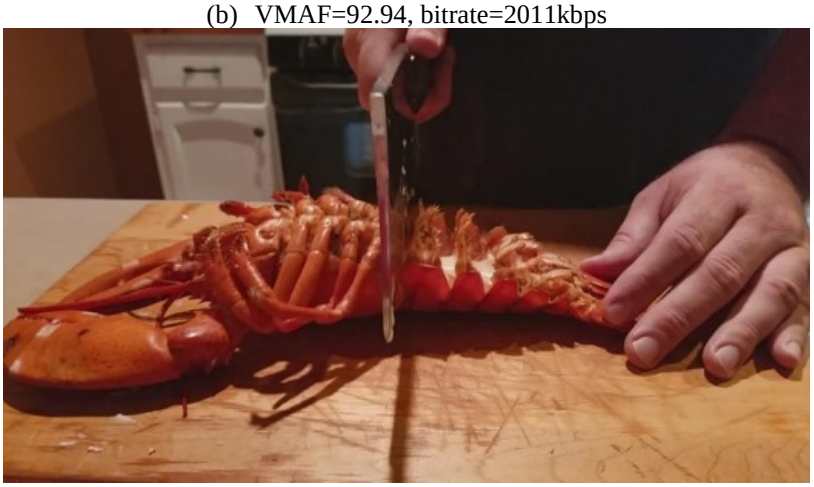
Since the network has made rapid progress, users are willing to pursue greater video quality. Existing research [1] has put forward analysis and improvement on the adaptive bit rate in live broadcast services. However, when encountering special environment such as taking a subway or places in the limited network conditions, adaptively adjusting the video quality to ensure a fluent experience is necessary. Therefore, it is important to accurately predict the video target bitrate based on the visual quality requirement.

In daily experiments, we find that although the encoded videos have similar visual quality scores, the bitrate costs are various. **Fig. 1** shows that the bitrate costs of encoded videos in the same resolution vary greatly when the visual quality scores evaluated by Video Multimethod Assessment Fusion (VMAF) [2] are similar. Several researches [3] [4] adaptively optimize the encoder according to the video content. When the target bitrate can be accurately predicted by the bitrate prediction algorithm based on the video content, instead of repeated encoding, coding time and bitrate will be saved greatly.

Recently, several methods have been proposed to predict the encoding parameters. One type of method is to select the optimal encoding parameters as the video resolution changes. The method used by Netflix [5] [6] consists of performing multiple encodings at different quantization levels to get Rate-Quality (R-Q) curves at several resolutions. The obtained R-Q curves are integrated to construct an optimal R-Q curve across resolutions to obtain the optimal bitrate parameters. Katsenou et al. [7] propose a prediction for quantization parameters. They extract the spatial-temporal characteristics and statistical data of the sequences at different resolutions. This method combines machine learning to predict the quantization level of RQ curves of different resolutions when they intersect. Chen et al. [8] propose a parameter prediction method based on probability distribution.

(a) VMAF=93.35, bitrate=1023kbps

(b) VMAF=92.94, bitrate=2011kbps



(c) VMAF=93.05, bitrate=4003kbps

(d) VMAF=93.49, bitrate=6070kbps

Fig. 1. Videos with different content need to be compressed at different bitrates to get the same perceptual quality.

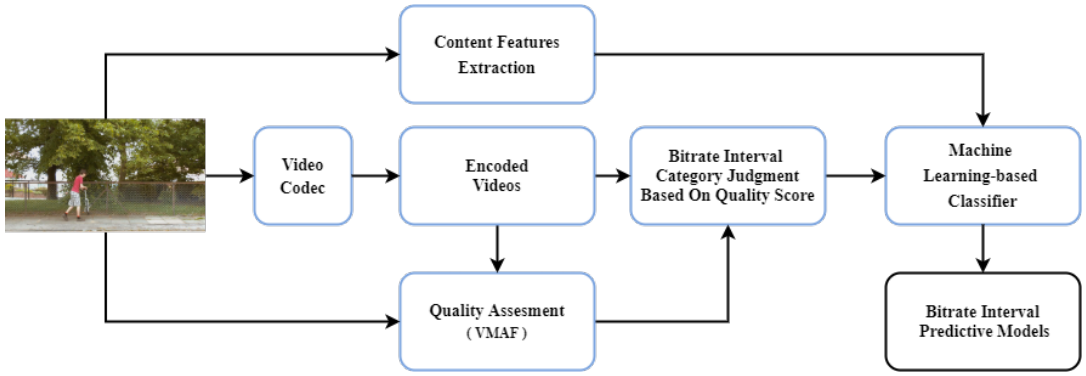
The author measures the usage of bandwidth and viewport size of millions of videos to get their probability distribution. This information is used for parameter optimization to ensure that the video quality remains while the target bitrate reduces.

Another way is to predict the coding parameters through the Rate-Distortion (R-D) curve. John et al. [9] propose a scheme to predict target bitrate of the UGC (User Generated Content) videos by R-D curve. K-means clustering algorithm is used to judge the centroid R-D curve of each category. Then the coding features given by the AV1 encoder of videos are extracted and used for Support Vector Machine (SVM) training to obtain the classification model. The video is classified by the model to obtain the target bitrate. Ling et al. [10] raise a framework for R-D curve prediction based on the characteristic related to the R-D curve. Bjontegaard's Delta-Rate/Quality between R-D curves is used as the basis for clustering. Some spatial-temporal features selected through a hierarchical feature selection scheme are extracted and these features are used in different machine learning classification algorithms to obtain the optimal classification model. The video to be encoded is classified by the model to obtain the target bitrate. Meng et al. [11] propose a binary classification model to determine the encoding parameters of short video. They collect user-generated short videos from Douyin and build a dataset for perceptual coding optimization. Each video is judged by human eyes whether the video encoded with large CRF coefficients can be replaced by the video encoded with small CRF coefficients. The trained binary classification model is used to judge the previous process.

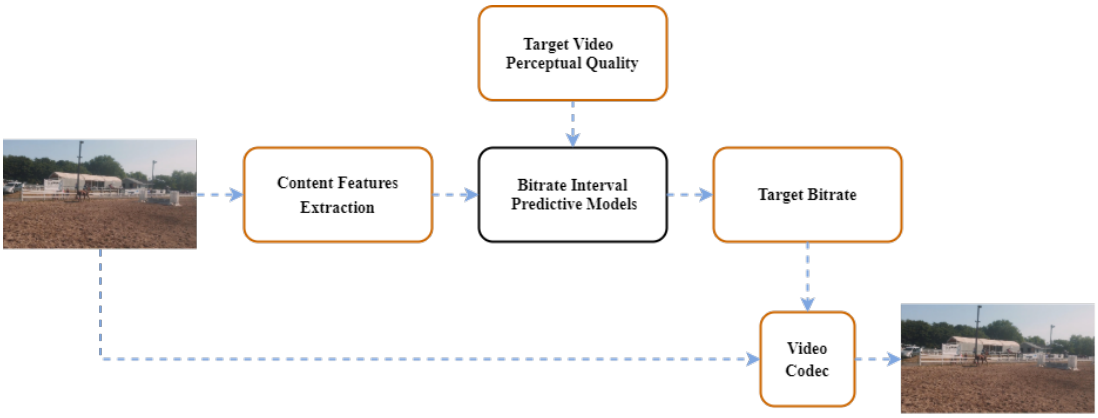
However, these methods still have their limitations. Most of the above methods do not consider the impact of visual characteristics. The two-classification model [11] takes visual factors into consideration, but it needs to be judged serially by multiple binary classification models to get the accurate coding parameters for a specific video. This not only requires building multiple datasets for training, but also spends a lot of coding time.

Regarding the issue above, we propose a multi-classification model to accurately predict the target bitrate interval of the input video to meet the visual requirements. Firstly, a dataset is constructed to generate multi-category models through machine learning. The quality score ladders based on the Just Noticeable Distortion (JND) level and the corresponding bitrate-interval categories are defined. Secondly, based on the JND model, several types of spatial-temporal features related to VMAF evaluation metrics and visual factors are extracted and processed statistically for classification based on machine learning. Finally, perception based bitrate-interval predictive models trained on the dataset by RandomForest classifier are used to determine the target bitrates of videos with target score.

The remainder of this paper is organized as follows. Section 2 introduces the establishment of the dataset, content features extraction, training and testing process. Section 3 presents our experimental setting and results. Finally, content summary and works to be improved are given in section 4.



(a) Predictive model generation framework.



(b) Algorithm application framework.

Fig. 2. Overview of the proposed multi-category prediction framework.

2. Proposed Method

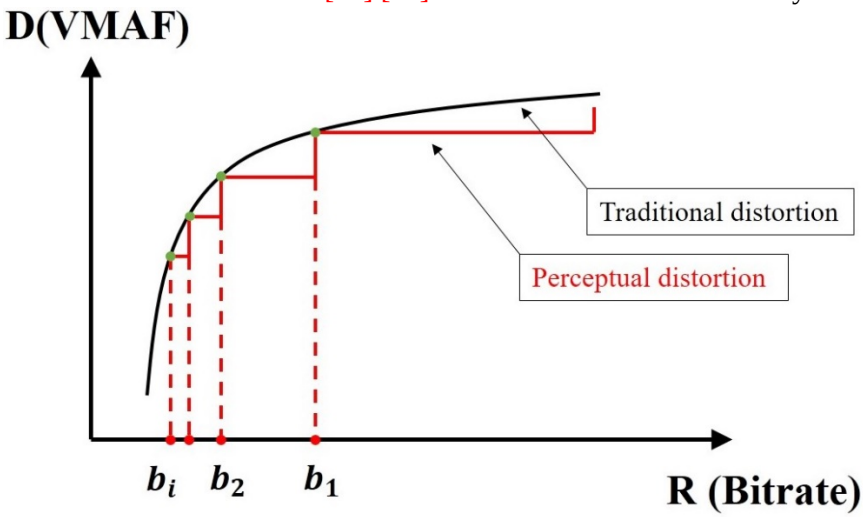
In this paper, our target is to predict the bitrate interval with the target quality of a given video. Through our algorithm, the target bitrate range can be accurately determined and no precoding is required. Block diagrams of prediction and application summarizing the proposed method are illustrated in **Fig. 2**.

2.1 Dataset Construction Based on JND Level

The traditional R-D curve that approximates to a continuous hyperbolic function is shown in **Fig. 3 (a)**. In HEVC [12], it was expressed in the following [13] [14],

$$D(R) = CR^{-K} \tag{1}$$

where C and K are model parameters related to the characteristic of the video content. However, for the human eye perception, the image quality distortion caused by the bitrate reduction cannot be sensitively detected by the visual system. Only when the distortion accumulates to a certain level that is a critical point, the perceptual distortion can be detected by the human eyes. This forms a stepped R-D function, which is different from the traditional continuous R-D curve. Recent studies [15] [16] have also confirmed this theory.



(a) The perceptual R-D function and the traditional R-D function

(b) The frame of pristine video.

(c) The frame of video in the first ladder, bitrate = 4172.24 kbps, VMAF = 91.86

(d) The frame of video in the first ladder, bitrate = 3261.54 kbps, VMAF = 88.12

(e) The frame of video in the fourth ladder, bitrate = 1643.09 kbps, VMAF = 74.36.

(f) The frame of video in the fourth ladder, bitrate = 1308.85 kbps, VMAF = 68.64.

Fig. 3. Diagram of video perceptual R-D relationship.

A perceptual R-D function is shown in **Fig. 3(a)**. In the perceptual distortion function, the visual quality of compressed video within a perceptual distortion ladder is approximately the same and the bitrates of the compressed videos with the similar quality span a certain bitrate segment.

In this work, we use VMAF as the video quality assessment method which combines human vision model with machine learning. Experiments [17] have shown that VMAF evaluates the perceptual quality achieving a higher connection than other objective quality evaluation methods including PSNR and SSIM in many scenes. Ozer [18] [19] informed that**:**

1) when the VMAF score reaches 93 and below, the encoded video will generate noticeable artifact.

2) when the video quality rating evaluated by VMAF drops 6 points, a JND level will produce. In a JND level, humans do not perceive the difference in video quality sensitively, which can be approximately regarded as the same visual quality. A stair step function about bitrate and perceptual distortion like **Fig. 3(a)** will form.

Figs. 3(c)-(f) show frames of the source video coded with different QPs. The perceptual quality of **Fig. 3(d)** is similar to that of **Fig. 3(c)** and the bitrates of videos span from 2000kbps to 4000kbps. We can observe the similar phenomenon from **Figs. 3(e)** and **(f)**.

Based on the above theory, we divide JND levels every 6 points starting from VMAF 93 points, and totally five levels are divided including {93 to 87}, {87 to 81}, ..., {69 to 63}. In

the same visual quality ladder, the bitrate consumption range of different videos varies greatly. In high-perceptual-quality area such as VMAF score in range 93 to 87, we use the similar bitrate division method in [20]. Target bitrates for HD1080p sequences are divided including 1000kb/s, 2000kb/s, 4000kb/s and 6000kb/s. With the decreasing quality of the encoded video compressed by larger QP, the drop in bitrate consumption of each video shows a different trend. To keep the data balanced, the bitrate intervals must be adjusted correspondingly, which is discussed below. The final division scheme is shown in **Table 1**.

Table 1. The level division of bitrate and video quality based on JND level

Video Quality Ladder (VMAF)	Bitrate (kbps)				
	Interval 1	**Interval 2**	**Interval 3**	**Interval 4**	**Interval 5**
93-87	<1000	1000-2000	2000-4000	4000-6000	>6000
87-81	<550	550-1200	1200-2400	2400-3600	>3600
81-75	<400	400-850	850-1700	1700-2600	>2600
75-69	<325	325-700	700-1400	1400-2000	>2000
69-63	<275	275-550	550-1100	1100-1700	>1700

To achieve classification based on machine learning, we need to construct datasets about relationship between bitrate and quality. We choose videos in HD1080p of Kinetics700 dataset [21] and encode all the videos with the HEVC reference software HM16.8 [22] with the Random Access configuration. The sequences are encoded in the following QP ranges: QP = {18, 19...45}. The first nine frames of a video where no scene change has occurred are compressed. In a sequence the first frame is I frame and the remaining eight frames are B frames. The version of VMAF used for quality evaluation is 1.3.15. According to the time point annotation in the dataset [21], we use FFMpeg to cut the video from the time point of the action.

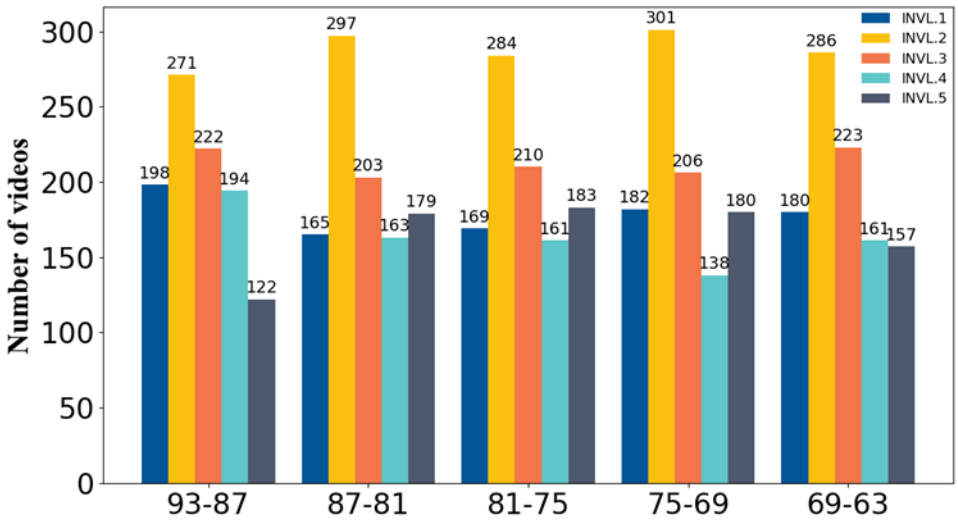


Fig. 4. Data distribution in different bit rate intervals under different visual quality levels.

In actual experiments, there are more encoded videos with medium bitrate consumption than videos with low bitrate consumption and high bitrate consumption in a same visual-quality category. In order to ensure the balance of data distribution and make the prediction more accurate, we balance the distribution of data after conscious selection. Totally 1007 videos are selected and their distributions are shown in **Fig. 4**.

2.2 Content Features Extraction

The JND model proposed in [23] mainly considers three independent visual factors: luminance, contrast, and structure. [24] shows the effect of color and brightness on VMAF quality evaluation score. In the evaluation process of VMAF, motion information and subjective evaluation are introduced.

According to the prior analysis, the spatial-temporal features related to evaluation metrics and visual factors are extracted and processed statistically. Video features used in this work are shown in **Table 2**. Totally four types of features are extracted to characterize video content. Three types of characteristics are spatial features including texture feature [25], contrast feature [26] and chromaticity feature [26]. The remaining type is temporal feature [25]. We adopt the similar statistical processing method for features in [8], and eliminate the useless processed-features for classification.

In the **Table 2**, F_1, F_2 correspond to statistical processing methods. "std" is standard deviation, "mean" is mean calculation. "skew" is skewness, "kurt" is the kurtosis and "max" is maximum. For example, $Texture = F_1\{F_2[Sobel(Y)]\}$ with $\{F_1, F_2\} = \{std, mean\}$ is calculated as, (1) The Y channel component of each frame is convolved with the Sobel filter. Then the pixels of each Sobel-filtered Y channel are calculated by F_2 where F_2 represents the mean calculation. (2) The mean value obtained after the calculation process in (1) is computed by F_1 where F_1 represents the standard deviation calculation in multiple frames. (F_1 is calculated on the output result of F_2).

Table 2. Summary of video spatial-temporal features.

Feature	Formula	Dimension
Texture	$Texture = F_1\{F_2[Sobel(Y)]\}$, where $\{F_1, F_2\} = \{std, std\}$, $\{mean, std\}$, $\{skew, std\}$, $\{mean, mean\}$, $\{skew, mean\}$, $\{std, mean\}$, $\{kurt, mean\}$, $\{kurt, std\}$, $\{max, mean\}$, $\{max, std\}$	10
Temporal	$Temporal = F_1\{F_2(Y_2 - Y_1)\}$, where $\{F_1, F_2\} = \{std, std\}$, $\{mean, std\}$, $\{mean, max\}$, $\{skew, std\}$, $\{mean, mean\}$, $\{skew, mean\}$, $\{std, mean\}$,$\{kurt, mean\}$, $\{kurt, std\}$, $\{max, mean\}$, $\{max, std\}$, $\{std, max\}$, $\{max, max\}$	13
Chromaticity	$Chromaticity_U = F_1\{F_2(U)\}$,$Chromaticity_V = F_1\{W_R * F_2(V)\}$, $W_R = 5$, where $\{F_1, F_2\} = \{std, std\}$, $\{mean, std\}$, $\{mean, max\}$, $\{skew, std\}$, $\{mean, mean\}$, $\{skew, mean\}$, $\{std, mean\}$, $\{kurt, mean\}$, $\{kurt, std\}$, $\{max, mean\}$, $\{max, std\}$, $\{std, max\}$, $\{max, max\}$	26
Contrast	$Contrast = F_1\{F_2(Y)\}$, where $\{F_1, F_2\} = \{std, std\}$, $\{mean, std\}$, $\{mean, max\}$, $\{skew, std\}$, $\{mean, mean\}$, $\{skew, mean\}$, $\{std, mean\}$, $\{kurt, mean\}$, $\{kurt, std\}$, $\{max, mean\}$, $\{max, std\}$, $\{std, max\}$, $\{max, max\}$	13

2.3 Training and Testing processing

The extracted features are used to represent different contents and to predict the bitrate-interval category. We train video classification models with different quality requirements on multiple machine learning based classifiers, which take the extracted features as input and bitrate-interval label as output. We choose the model obtained by the classifier with the best predictive performance as the final application model. All five independent prediction models are trained. Each model corresponds to a perceptual quality requirement.

3. Experimental Results

3.1 Experiment implementation

Fig. 5 is the schematic diagram of our application platform. For a video to be compressed, we import the video to the platform, and then select the desired compressed video quality. The platform will calculate the spatial-temporal characteristics of the original video and obtain the bitrate interval through the classification prediction model. The average bitrate of the interval is used to compress the original video. The target VMAF score, the actual bitrate and the actual VMAF score of the compressed video will display in the 'Information' section. The details of the video can be observed through the zoom control. All our coding, prediction and model applications are performed on the platform Intel(R) Core (TM) i5-4590 CPU @ 3.30GHz with 8 GB of RAM.

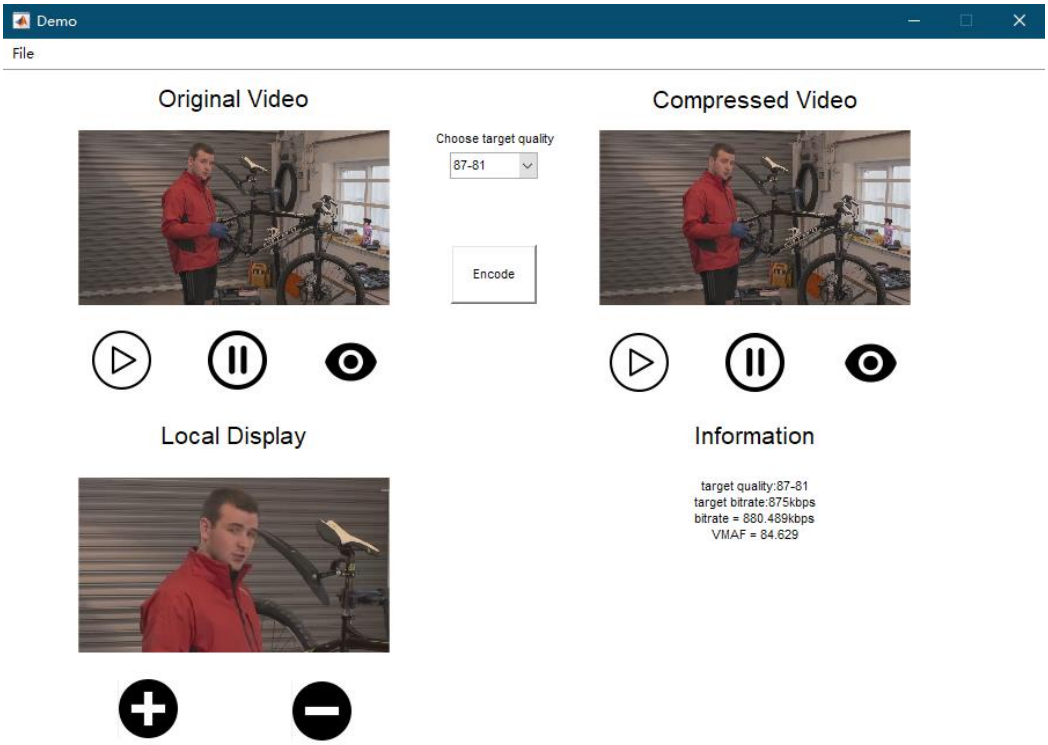


Fig. 5. Schematic diagram of algorithm application platform.

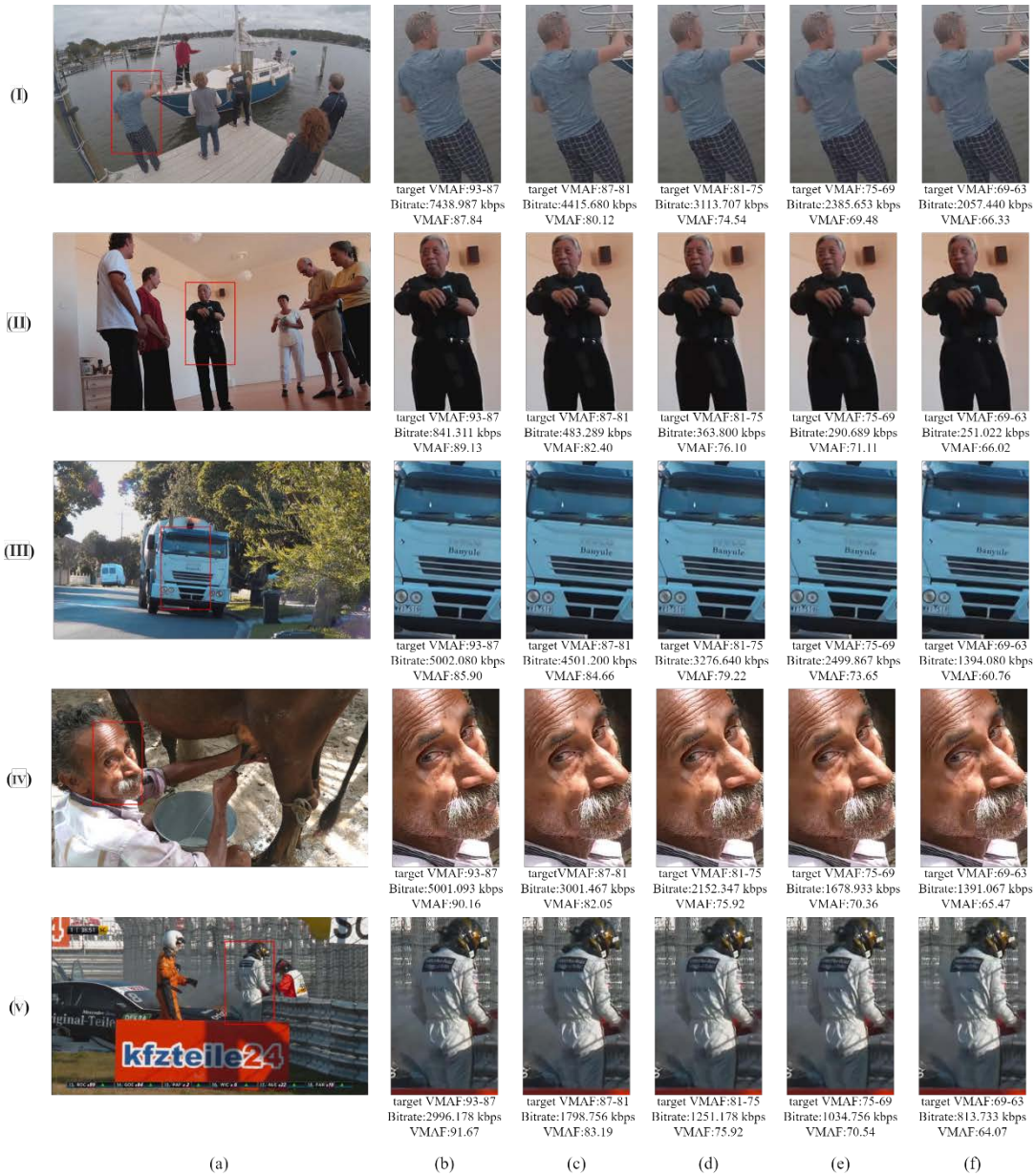
3.2 Classify prediction and result

90% of the videos in the dataset are used for training and remaining 10% are used for testing. Classification prediction results of all classifiers are shown in **Table 3**. We apply four machine learning-based classifiers including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Decision Tree (DT). Compared with other machine learning-based classifiers, random forest-based classifier is able to process high-dimensional features without feature selection. For unbalanced datasets, random forest-based classifier can balance errors and has strong anti-interference ability. Among various machine-learning based classification algorithms, the predictive performance of the random forest-based classifier is the best. We choose random forest-based classifier for predicting the bitrate interval of a video in target perceptual quality. In each model, an accuracy of about 70% can be achieved, and an average accuracy of 70.5% can be achieved.

Table 3. Performance of the predictive model

VMAF-Ladder	Classification algorithm	Accuracy	Recall	F1 score
93-87	RF	**0.703**	**0.696**	**0.700**
	SVM	0.515	0.420	0.417
	LR	0.554	0.620	0.475
	DT	0.624	0.656	0.634
87-81	RF	**0.683**	**0.696**	**0.691**
	SVM	0.554	0.568	0.470
	LR	0.554	0.501	0.482
	DT	0.673	0.689	0.673
81-75	RF	**0.733**	**0.691**	**0.694**
	SVM	0.584	0.574	0.502
	LR	0.564	0.503	0.486
	DT	0.624	0.643	0.611
75-69	RF	**0.703**	**0.640**	**0.628**
	SVM	0.584	0.463	0.468
	LR	0.584	0.498	0.478
	DT	0.644	0.619	0.619
69-63	RF	**0.703**	**0.708**	**0.702**
	SVM	0.564	0.538	0.468
	LR	0.534	0.543	0.468
	DT	0.644	0.653	0.634
Average	**RF**	**0.705**	**0.687**	**0.683**

Fig. 6 shows the results of the videos compressed by the target bitrate predicted by the models. We test 10 videos that are not part of the dataset and import them into the platform for experimentation. After the classification prediction by the trained model, the target bitrate category of the video under the target visual quality can be obtained. Then the video is compressed by the average bitrate of the predictive interval. **Fig. 6 (a)** is the pristine video. **Figs. 6 (b)-(f)** correspond to the compressed videos with the target bitrate which is predicted by the model under different visual quality. We can observe that the compressed video quality score meets the visual requirement.



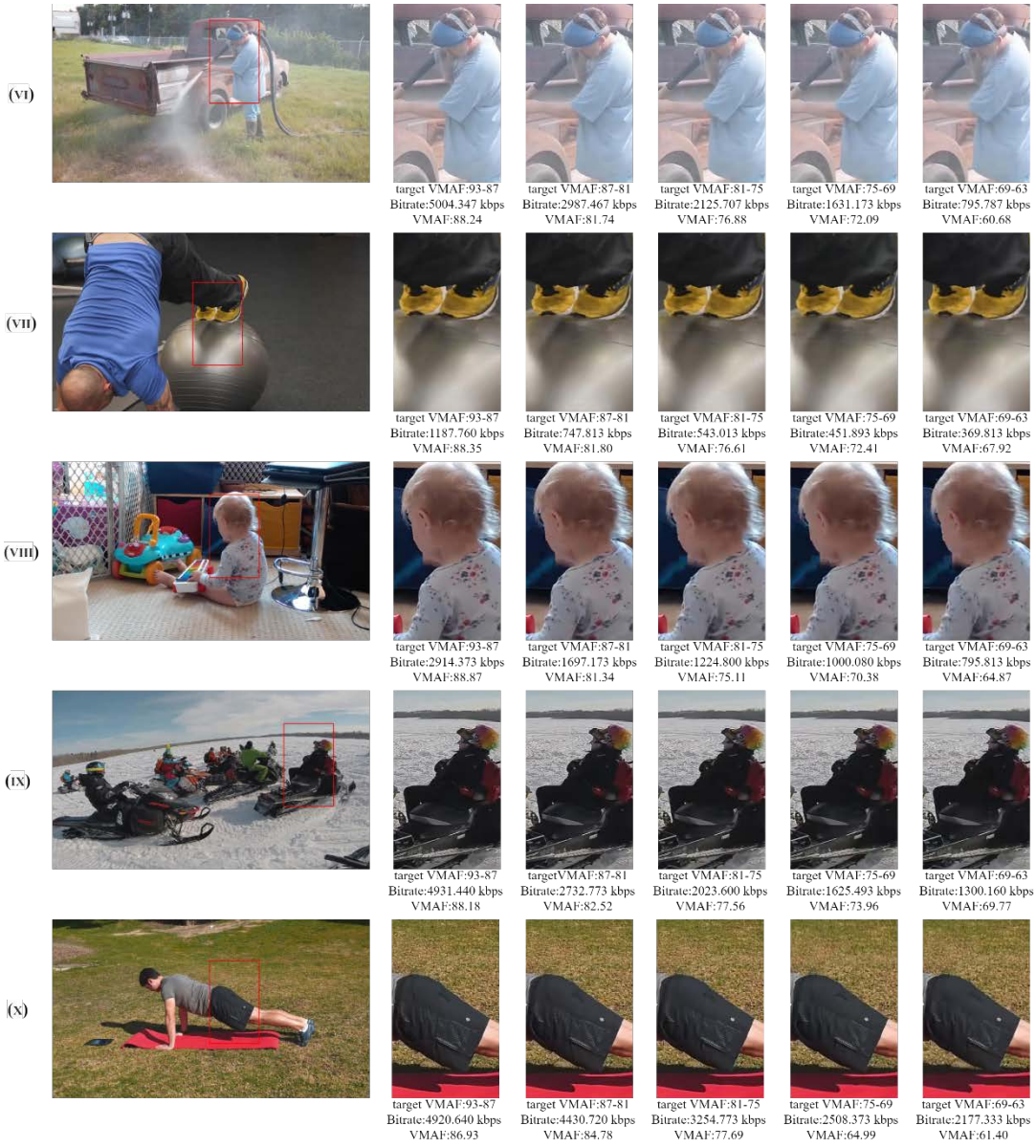


Fig. 6. Videos encoded with the target bitrate which is accurately predicted from the model.

4. Conclusion

In this paper, we propose a multi-category model to predict the target bitrate interval and obtain the target video perceptual quality based on JND level. A dataset is constructed to generate multi-category models through machine learning. The quality score ladders based on JND level and the corresponding bitrate-interval categories are defined. Secondly, several types of spatial-temporal features related to the JND model, VMAF evaluation metrics and visual factors are extracted and processed statistically. Finally, several perception based bitrate-interval predictive models trained on the dataset by random forest based classifier are used to

accurately determine the target bitrate with target video quality. In our experiments, classification accuracy reaches 0.705 and the encoded video can achieve the target perceptual visual requirement. The predicted result is acceptable but there is still room for improvement. We will carry on enlarging our dataset and try to introduce more perception-related features to make the results better.

References

[1] N. Kim and B. Lee, "Analysis and Improvement of MPEG-DASH-based Internet Live Broadcasting Services in Real-world Environments," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 5, pp. 2544-2557, May 2019. Article (CrossRef Link)

[2] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Netflix, Los Gatos, CA, USA, The Netflix Tech Blog, 2016. [Online]. Available:https://medium.com/netflix-techblog/toward-a-practicalperceptualvideoquality-metric-653f208b9652

[3] H. Yang, L. Shen, X. Dong, Q. Ding, P. An and G. Jiang, "Low-Complexity CTU Partition Structure Decision and Fast Intra Mode Decision for Versatile Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1668-1682, Jun. 2020. Article (CrossRef Link)

[4] L. Shen, Z. Zhang and Z. Liu, "Adaptive Inter-Mode Decision for HEVC Jointly Utilizing Inter-Level and Spatiotemporal Correlations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1709-1722, Oct. 2014. Article (CrossRef Link)

[5] J. De Cock, Z. Li, M. Manohara and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 1484-1488, Sep. 2016. Article (CrossRef Link)

[6] I. Katsavounidis, "Dynamic optimizer - a perceptual video encoding optimization framework," Netflix, Los Gatos, CA, USA, The Netflix Tech Blog, Mar.2018. [Online].Available:https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f

[7] A. V. Katsenou, J. Sole and D. R. Bull, "Content-gnostic Bitrate Ladder Prediction for Adaptive Video Streaming," in *Proc. of Picture Coding Symposium (PCS)*, pp. 1-5, Nov. 2019. Article (CrossRef Link)

[8] C. Chen, Y. Lin, S. Benting, and A. Kokaram, "Optimized Transcoding for Large Scale Adaptive Streaming Using Playback Statistics," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 3269–3273, Oct 2018. Article (CrossRef Link)

[9] S. John, A. Gadde and B. Adsumilli, "Rate Distortion Optimization Over Large Scale Video Corpus With Machine Learning," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 1286-1290, Oct. 2020. Article (CrossRef Link)

[10] S. Ling, Y. Baveye, P. L. Callet, J. Skinner and I. Katsavounidis, "Towards Perceptually-Optimized Compression of User Generated Content (UGC): Prediction Of UGC Rate-Distortion Category," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, Jul. 2020. Article (CrossRef Link)

[11] S. Meng, Y. Li, Y. Liao, J. Li and S. Wang, "Learning to encode usergenerated short videos with lower bitrate and the same perceptual quality," in *Proc. of IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 383-386, Dec. 2020. Article (CrossRef Link)

[12] G. J. Sullivan, J. M. Boyce, Y. Chen, J. Ohm, C. A. Segall and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC)," *IEEE Journal of selected topics in Signal Processing*, vol. 7, no. 6, pp. 1001-1016, Dec. 2013. Article (CrossRef Link)

[13] N. Kamaci, Y. Altunbasak and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 8, pp. 994-1006, Aug. 2005. Article (CrossRef Link)

[14] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 1027-1042, April 1998. Article (CrossRef Link)
[15] S. Hu, H. Wang and C. -. J. Kuo, "A GMM-based stair quality model for human perceived JPEG images," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1070-1074, Mar. 2016. Article (CrossRef Link)
[16] Wang H, Katsavounidis I, Zhou J, et al. "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *Journal of Visual Communication and Image Representation*, vol 46, pp. 292-302, 2017. Article (CrossRef Link)
[17] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. Cock, "Vmaf: The journey continues," Netflix, Los Gatos, CA, USA, The Netflix Tech Blog, Oct. 2018. [Online]. Available:https://netflixtechblog.com/vmaf-the-journeycontinues-44b51ee9ed12
[18] Ozer J. "Finding the Just Noticeable Difference with Netflix VMAF," Sep. 2017. [Online]. Available: https://streaminglearningcenter.com/codecs/finding-the-just-noticeable-difference-with-netflix-vmaf.html
[19] Ozer J. "Fine-Tune Your Encoding With Objective Quality Metrics – Video and Handout," Dec. 2019.[Online]. Available: https://streaminglearningcenter.com/learning/fine-tune-your-encoding-with-objective-quality-metrics-video-and-handout.html
[20] Z. Liu, L. Wang, X. Li and X. Ji, "Optimize x265 Rate Control: An Exploration of Lookahead in Frame Bit Allocation and Slice Type Decision," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2558-2573, May 2019. Article (CrossRef Link)
[21] Carreira J, Noland E, Hillier C, et al. "A short note on the kinetics-700 human action dataset," *arXiv preprint*, 2019. Article (CrossRef Link)
[22] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012. Article (CrossRef Link)
[23] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang and S. Kwong, "Just Noticeable Distortion Profile Inference: A Patch-Level Structural Visibility Learning Approach," *IEEE Transactions on Image Processing*, vol. 30, pp. 26-38, 2021. Article (CrossRef Link)
[24] A. Zvezdakova, S. Zvezdakov, D. Kulikov, and D. Vatolin, "Hacking vmaf with video color and contrast distortion," *arXiv preprint*, 2019. Article (CrossRef Link)
[25] "Subjective Video Quality Assessment Methods for Multimedia Applications," ITU-R Rec. P.910, 1999.
[26] S. Wolf and M. Pinson, "Video quality measurement techniques," NTIA, Washington D.C., Tech. Rep. 02-392, Jun. 2002.

Zhenwei Yang received the B.S. degree from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2019, where he is currently pursuing the M.S. degree. His research interests include video coding, rate control.

Liquan Shen received the B.S. degree in automation control from Henan Polytechnic University, Jiaozuo, China, in 2001, and the M.E. and Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2005 and 2008, respectively. Since 2008, he has been a Faculty Member with the School of Communication and Information Engineering, Shanghai University, where he is currently a Professor. From November 2013 to November 2014, he was a Visiting Professor with the Department of Electrical and Computer Engineering, University of Florida, Gainesville. He has authored or coauthored more than 100 refereed technical papers in international journals and conferences in the field of video coding and image processing. He holds ten patents in the areas of image/video coding and communications. His major research interests include high efficiency video coding, perceptual coding, video codec optimization, 3DTV, and video quality assessment.