ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097274

A DATABASE FOR MULTI-MODAL SHORT VIDEO QUALITY ASSESSMENT

Yukun Zhang[1,2], *Chuan Wang*[*1,2], *Sanyi Zhang*[1,2], *Xiaochun Cao*[*3]

SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences
School of Cyberspace Security, Chinese Academy of Sciences
School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University

ABSTRACT

The short video has gained increasing attention in information sharing and commercial promotions due to the fast development of social platforms. Accompanying, it introduces great requirements for assessing the quality of short videos for efficient information acquirement and propagation. However, existing video quality assessment researches focus on assessing video content with five rating scores, limiting the assessment to a one-dimension and simplified criterion. In this paper, we establish a novel database dubbed MMSVD-Douyin for assessing multi-modal short video quality under consideration of three evaluation criteria. It includes 4,684 short videos, three kinds of modalities, six kinds of data formats, and three assessment criteria. To conduct the short video quality assessment, we set up an all-around multi-modal short video quality assessment benchmark (MulSVQA) that dynamically fuses representations from three modalities and produces numbers of "likes", "shares" and "comments" of short videos.

Index Terms— Short Video Quality Assessment, Multi-modal database, Multi-criteria Evaluation.

1. INTRODUCTION

As portable devices and social media applications become widely accessible, users are getting used to capturing information with short videos, which possess characteristics of multiple modalities, limited-time series, and multi-facet assessments. A short video always consists of multiple modalities including image (cover image), video (video content), and text language (video title, author information, etc.), and be tagged as "clicks", "likes", "shares", "comments", etc. According to statistics, the total number of monthly active users existing in TikTok and Douyin (two popular video-sharing and social media platforms) achieves 1.2 billion [1] and 934 million [2], respectively. With a huge user community, short videos have accounted for a large portion of internet traffic and become a commercially important carrier, thus bringing about contributions to commercial promotions and gains. A practical issue arising from short videos is assessing values that may be potentially produced, i.e., Short Video Quality Assessment (SVQA).

Fig. 1. Data sample of the MMSVD-Douyin dataset. "*" is used to block the author's information.

Current VQA datasets, e.g., CVD2014 [3], LIVE-VQC [4], and YouTube-UGC [5] are constructed based on the video modality. The quality estimation is conducted with the Mean Opinion Score (MOS) and Differential Mean Opinion Score (DMOS), both of which describe the rating scores of videos. Most VQA methods [6–14] consider uni-modal video content and fail to handle videos with multiple modalities. Although Min et.al. [15] includes audio modality considerations, it still cannot provide multi-facet evaluations.

To conduct the short video quality assessment, we first build a novel multi-modal short video dataset for SVQA dubbed MMSVD-Douyin, including 4,684 short videos collected from the Douyin platform, three multimedia modalities consisting of text (author's name, personalized signature, and video caption), image (author profile and video cover image) and video, as shown in Fig. 1. The design of the evaluation metric considers three evaluation indicators including "likes", "shares" and "comments". Our collected SVD-Douyin is the first complete dataset for comprehensively assessing short videos with multiple modalities and assessment indicators. Further, to conduct the multi-modal

∗ is for the corresponding author.

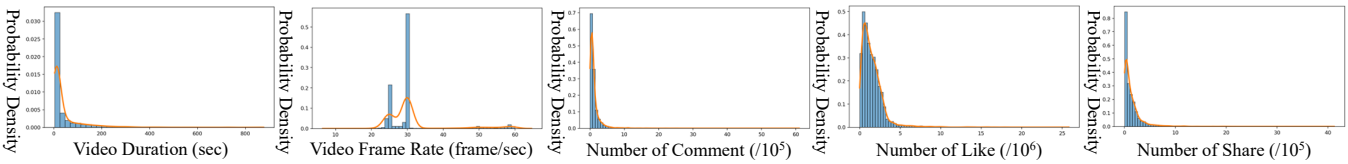

Fig. 2. Histograms and the fitted kernel distributions of video duration, frame rate, "comment", "like", and "share".

Table 1. The statistics of our new dataset.

	Min	Max	Average
Video Duration (sec)	4	877	44.8
Frame Rate (frame/sec)	11	60	31
Number of Comments	0	6,096,949	114,286.8
Number of Likes	5,665	25,836,469	1,518,952.8
Number of Shares	550	4,123,251	134,917.8

short video quality assessment with various indicators, we present an all-around multi-modal short video quality assessment benchmark (MulSVQA) with the assessment in three aspects, i.e., the number of "likes", "comments" and "shares". The framework consists of three modules, 1) a multi-modal feature representation module for feature extraction of three modalities with six kinds of contents, 2) a multi-modal feature fusion module to generate a comprehensive representation for short videos, and 3) a multi-task video quality assessment module. For feature extraction, we separately utilize a Pre-trained BERT (PERT) model [16], a VGG16 model [17], and a CLIP-initialized pre-trained MCQ model [18] to extract text, image and video representations. We design a transformer encoder with gate fusion to conduct multi-modal fusion for generating short video representations. The short video representations are fed into three independent assessments for evaluating the number of "likes", "comments" and "shares", respectively.

To summarize, our contributions are summarized as 1) We set up a novel video quality assessment task considering short videos under realistic and diverse indicators. 2)We build a novel large-scale, multimodal short video quality assessment database dubbed MMSVD-Douyin, which includes 4,684 short videos, three kinds of modalities, 6 kinds of contents, and three assessment indicators. 3) We propose an all-around, multimodal short video quality assessment benchmark (MulSVQA) to conduct a short video assessment and discuss the game effect under multiple modalities.

2. DATASET

In this section, we first present detailed steps of constructed multi-modal short video quality assessment dataset MMSVD-Douyin and manifest statistics including video duration, frame rate, number of "likes", "comments" and "shares". Subsequently, we present differences and diversity compared with existing VQA datasets.

2.1. Dataset Construction and Data Analysis

We collect short videos from Douyin, a popular short video-sharing platform with active users in China. For short video selection, We choose 50 pieces of data under the Douyin Short Video Ranking List - Today's Popular Video List every day [2], and finally a total of 4,684 pieces of data are collected. For each short video, we download its complete information covering both the author and the short video. For author information, we choose the author's name, personalized signature, and author profile as parts of the short video data. For video information, we select cover images and videos to encapsulate short video content. For assessment, we also record the number of "likes", "comments" and "shares" along with each video. Considering that these records of short videos change after each click, we postpone the record time to a few days after downloading short videos. The statistics of collected videos are presented in Table 1. As shown in Table 1, the MMSVD-Douyin dataset covers large variances in video duration and the number of indicators. We plot histograms and fitted kernel distributions of video duration, frame rate, "comments", "likes", and "shares" in Figure 2. The distributions for these data are mostly long-tailed distributions, which match actual social media data. These long-tailed distributions of data make it more challenging to create models that can predict the value of short videos at different levels.

2.2. Dataset Comparison

We present a detailed comparison with popular "in-the-wild" video quality databsets [4, 5, 19] in Table 2. our MMSVD-Douyin dataset distinguishes itself in several ways. Firstly, in addition to the video data, we add another two modalities that cover text (author's name, personalized signature, and video caption) and image data (author profile and video cover image). Besides, we collect the number of "comments", "likes", and "shares" of each video, which are used as three criteria for evaluating the value of short videos. Finally, instead of sampling the videos to a fixed duration, we keep their original content and have a larger distribution range. The shortest duration is 4 seconds, and the longest can reach 877 seconds (14 minutes 37 seconds). Likewise, we did not restrict the collected videos to have fixed resolutions or aspect ratios, making the proposed dataset much more representative of real-world content.

Table 2. Summary of popular video quality datasets and our dataset.

Dataset	Source	Unique Contents	Resolution	Frame Rate	Video Duration	Format	Data Type	Evaluation Indicator
CVD2014 [3]	Captured	5	480p,720p	9-30	10-25	AVI	Video	Mean Opinion Score (MOS)
KoNViD-1k [19]	Flickr	1,200	540p	24-30	8	MP4	Video	Mean Opinion Score (MOS)
LIVE-VQC [4]	Captured	585	240p-1080p	19-30	10	MP4	Video	Mean Opinion Score (MOS)
YouTube-UGC [5]	YouTube	1,500	360p-4k	15-60	20	MKV	Video	Mean Opinion Score (MOS)
MMSVD-Douyin (Ours)	Douyin	4,684	480p-720p	11-60	4-877	MP4	Text + Image + Video	Comments, Likes, Shares

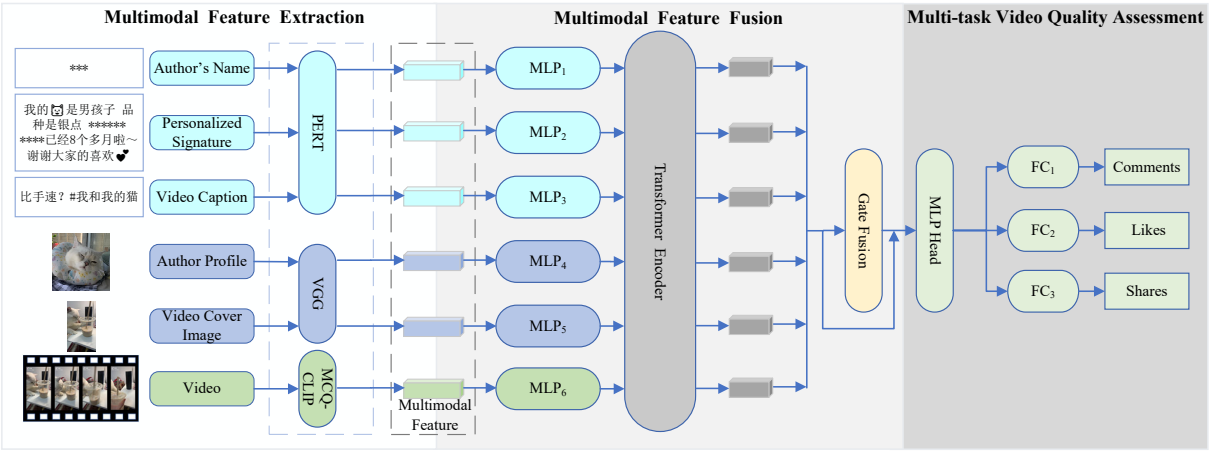

Fig. 3. Multi-modal short video quality assessment benchmark framework.

3. METHODOLOGY

In this section, we introduce the benchmark for multimodal short video quality assessment. As illustrated in Figure 3, the benchmark consists of three modules: multimodal representation extraction, multimodal feature fusion and multi-task short video quality assessment.

3.1. Multimodal Representation Extraction

In our dataset, each piece of data contains an author's name, a personalized signature, a short video caption, an author profile, a video cover image, and a short video.

Text Feature Representation. The input text messages are first tokenized into a token sequence $\boldsymbol{s}$. To fit the PERT [16] encoding procedure, we add the token [CLS] to the head of the sequence as $\boldsymbol{s}' = [[\text{CLS}], w_1, ..., w_N]$, where w is the token after tokenization. $\boldsymbol{s}'$ is converted into a contextualized representation $\boldsymbol{H} \in \mathcal{R}^{N \times d_T}$ (N is the maximum sequence length and $d_T = 768$ is the dimension of hidden layers) through an embedding layer, consisting of word embeddings, position embeddings, and segment embeddings, and a consecutive L-layer transformer as,

$$\boldsymbol{H}^{(0)} = \text{Embedding}(s'), \tag{1}$$

$$\boldsymbol{H}^{(i)} = \text{Transformer}(\boldsymbol{H}^{(i-1)}), i \in 1, ..., L, \boldsymbol{H} = \boldsymbol{H}^{(L)}. \tag{2}$$

The author's name $\boldsymbol{s}_{name}$, the personalized signature $\boldsymbol{s}_{sign}$ and the short video caption $\boldsymbol{s}_{caption}$ are fed into PERT [16], respectively and author's name feature $\boldsymbol{f}_{name} \in \mathcal{R}^{d_T}$, personalized signature feature $\boldsymbol{f}_{sign} \in \mathcal{R}^{d_T}$ and short video title feature $\boldsymbol{f}_{title} \in \mathcal{R}^{d_T}$ are provided.

Image Feature Representation. We utilize the VGG16 [17] to extract image features, which include 13 convolutional layers, 3 fully-connected layers, and one softmax layer. We use the output of the second fully-connected layer as the features. The author profile photo and the short video cover image are separately fed into VGG16 and obtain features $\boldsymbol{f}_{prof} \in \mathcal{R}^{d_I \times 1}$ and $\boldsymbol{f}_{cover} \in \mathcal{R}^{d_I \times 1}$, where $d_I = 4096$.

Video Feature Representation. We utilize the CLIP-initialized pre-trained MCQ [18] to extract video feature representations. Given a video $\boldsymbol{V} \in \mathcal{R}^{M \times H \times W \times C}$, where M, H, W, C denote its number of frames, height, width, and the number of channels, respectively. The video is split into $M \times N$ patches with $P \times P$ patch size and $N = HW/P^2$. The video patches $\{\boldsymbol{x}_p^j \in \mathcal{R}^{P \times P \times C} | j = 1, 2, \cdots, M \times N\}$ are flattened into a sequence of tokens $\boldsymbol{z} \in \mathcal{R}^{M \times N \times d_V}$ ($d_V = 512$) and mapped to d_V dimension with a linear projection head. A learnable [CLS] token denoted as $\boldsymbol{x}_{cls}$ is concatenated to the head of the token sequence. The output [CLS] token serves as the final video representations $\boldsymbol{f}_{video} \in \mathcal{R}^{d_V}$. Learnable spatial positional embeddings $\boldsymbol{E}_{pos} \in \mathcal{R}^{(1+M \times N) \times d_V}$ are added to each video tokens. In different frames, patches with the same spatial position share the same spatial positional embedding. The final input token sequence can be expressed as:

$$\boldsymbol{z}^0 = [\boldsymbol{x}_{cls}; F(\boldsymbol{x}_p^1); F(\boldsymbol{x}_p^2); \cdots; F(\boldsymbol{x}_p^{M \times N})] + \boldsymbol{E}_{pos}, \tag{3}$$

where F is the linear projection head.

3.2. Multimodal Feature Fusion

To make full use of the obtained multimodal representation, we use a 3-layer transformer encoder to fuse the text, image, and video features. Since the dimensions of the previously obtained multimodal features are not consistent, we feed these 6 features into 6 different MLPs, which consist of two fully-connected layers, to adjust their dimensions to $d_M = 512$. Then we concatenate six adjusted features to form the input sequence $\boldsymbol{f}^0 \in \mathcal{R}^{6\times d_M}$ of the transformer encoder. The output $\boldsymbol{f}^3$ of the transformer encoder is fed into a gate layer, which consists of two convolutional layers and one softmax layer, to obtain the weight α_k for each feature. The overall feature $\boldsymbol{f}$ is obtained by weighted summation over all multimodal feature representations as $\boldsymbol{f} = \sum_{k=1}^{6} \alpha_k \boldsymbol{f}_k^3$.

3.3. Prediction

We apply an MLP head and 3 different fully-connected layers to estimate the number of "comments", "likes", and "shares", respectively. The MLP head consists of two fully-connected layers, which reduces the feature dimension to half of the original. The fully-connected layers are adopted to reduce the feature dimension to 1. The network is optimized with the L1 loss calculating the absolute value between the prediction and the ground truth on three indicators.

4. EXPERIMENTS

4.1. Implementation and Evaluation

We randomly select 4000 videos for the training and leave 684 videos for the testing. All images and videos are resized to 224 x 224. Each video is divided into M equal segments, from which one frame is uniformly sampled. VGG16 [17] is pre-trained on ImageNet [20]. Chinese-PERT-base model [16] is pre-trained on the same data as MacBERT [21]. Our model is trained by the SGD optimizer with 150 epochs, where the batch size is 100. The learning rate is initialized to 1×10^{-3} and is scaled by 0.1 every 50 epochs. All experiments are performed with one NVIDIA 3060 GPU using PyTorch. To evaluate predictions, we adopt commonly used metrics: Mean Absolute Error (MAE).

4.2. Experimental Analysis

We conduct an experimental comparison of the proposed MulSVQA benchmark under different modalities and show results in Table 3. Firstly, we only conduct the quality assessment with video content. Secondly, we add text content including the author's name, personalized signature, and video caption into the training. We observe that adding text modality does help the overall assessment. We argue that text modality has much-contributed information for assessment since text contents describe the most outstanding information in short videos. Thirdly, when adding image modality containing the author profile and video cover image, the contribution is limited. We conclude the inferiority of the image modality is that the information in images may be included in the video with a high probability. Finally, when conducting with all modalities, we observe that adding image modality hurts performance. We argue that the reason for this result is that the gap between the three modalities may sometimes introduce redundant and noisy information. Specifically, in addition to the information in image modality that may be included in the video, the information in text modality may also be contained in images, such as the video title (video cover) and video author (author profile). We conclude that the long-tailed distribution heavily disturbs the quality assessment. Besides, gaps among multiple modalities also deteriorate the representation ability of fused features. We argue that in short video quality assessment, the video modality dominates the optimization process, and may suppress the optimization of the text and image modality, resulting in worse performance for prediction using the three modalities.

Table 3. Experimental Results.

Video	Text	Image	Criterion	MAE	Mean Error
✓			Comments	87,198	50,040
			Likes	855,611	361,993
			Shares	108,382	70,116
✓	✓		Comments	86,792	53,134
			Likes	832,529	377,522
			Shares	107,918	69,817
✓		✓	Comments	87,095	46,679
			Likes	854,247	245,329
			Shares	107,781	61,447
✓	✓	✓	Comments	88,314	49,856
			Likes	846,417	254,633
			Shares	106,556	63,080

5. CONCLUSION

In this paper, we tackle a novel short video quality assessment task that conducts multiple assessments with multi-modal short video data based on a novel multi-modal short video quality assessment dataset MMSVD-Douyin. The proposed all-around multi-modal short video quality assessment benchmark MulSVQA presents the assessment performance.

6. ACKNOWLEDGE

This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3103504, in part by the National Science Foundation of China Under Grant 62132006 and 62202461, in part by the China Postdoctoral Science Foundation under Grant 2021M703472 and Grant 2022M723364, and in part by the Shenzhen Science and Technology Program (No. 20220016).

7. REFERENCES

[1] Omnicore, "Tiktok statistics," [Online] Available: https://www.omnicoreagency.com/tiktok-statistics/.

[2] Douyin, "Douyin short video," [Online] Available: https://www.douyin.com.

[3] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen, "Cvd2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[4] Zeina Sinno and Alan Conrad Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[5] Yilin Wang, Sasi Inguva, and Balu Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[6] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] Michele A Saad, Alan C Bovik, and Christophe Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[8] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[9] Jari Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[10] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[11] Dingquan Li, Tingting Jiang, and Ming Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[12] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," *arXiv preprint arXiv:2207.02595*, 2022.

[13] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik, "Patch-vq:'patching up'the video quality problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14019–14029.

[14] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[15] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylene CQ Farias, and Alan Conrad Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing*, vol. 29, pp. 6054–6068, 2020.

[16] Yiming Cui, Ziqing Yang, and Ting Liu, "Pert: Pre-training bert with permuted language model," *arXiv preprint arXiv:2203.06906*, 2022.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo, "Bridging video-text retrieval with multiple choice questions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16167–16176.

[19] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.