Telepresence Video Quality Assessment

Zhenqiang Ying[1] , Deepti Ghadiyaram[2] , and Alan Bovik[1]
[1]University of Texas at Austin [2]Facebook AI
zqying@utexas.edu, deeptigp@fb.com, bovik@ece.utexas.edu

Abstract

Global Internet traffic of video conferencing has dramatically increased because of the pandemic, efficient and accurate video quality tools are needed to monitor and perceptually optimize telepresence traffic streamed via Zoom, Webex, Meet, etc. However, existing models are limited in their prediction capabilities on multi-modal, live streaming telepresence content. Here we address the significant challenges of Telepresence Video Quality Assessment (TVQA) in several ways. First, we mitigated the dearth of subjectively labeled data by collecting ~2k telepresence videos from different countries, on which we crowdsourced ~80k subjective quality labels. Using this new resource, we created a firstof-a-kind online video quality prediction framework for live streaming, using a multi-modal learning framework with separate pathways to compute visual and audio quality predictions. Our all-in-one model is able to provide accurate quality predictions at the patch, frame, clip, and audiovisual levels. Our model achieves state-of-the-art performance on both existing quality databases and our new TVQA database, at a considerably lower computational expense, making it an attractive solution for mobile and embedded systems.

Dataset Collection

We collected 78, 880 ratings (34 ratings on each video) on 2320 videos from 526 subjects.

- Includes videos uploaded from 80 countries.
- representative of telepresence content (including grid-views of multiple speakers, single speaker views, slide sharing, screen content, etc.)
- Diverse resolution, aspect ratios, and distortions.

Study Interface Design

- Platform: Amazon Mechanical Turk (AMT).
- Test method: continuous rating scale instead of Absolute Category Rating (ACR) scale.
- quality control: use repeated and "golden" videos for which the highly reliable subjective scores were previously obtained, which may then be used to compare with the worker's inputs.

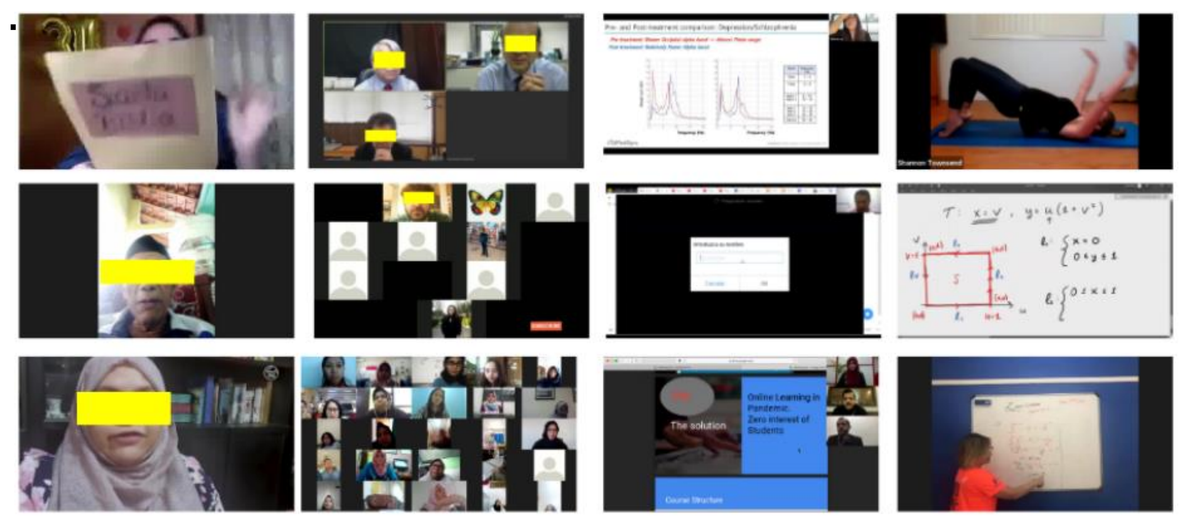

Figure: Sample video frames from the proposed database, each resized to fit. The actual videos are of highly diverse sizes and resolutions. Faces are masked to ensure privacy.

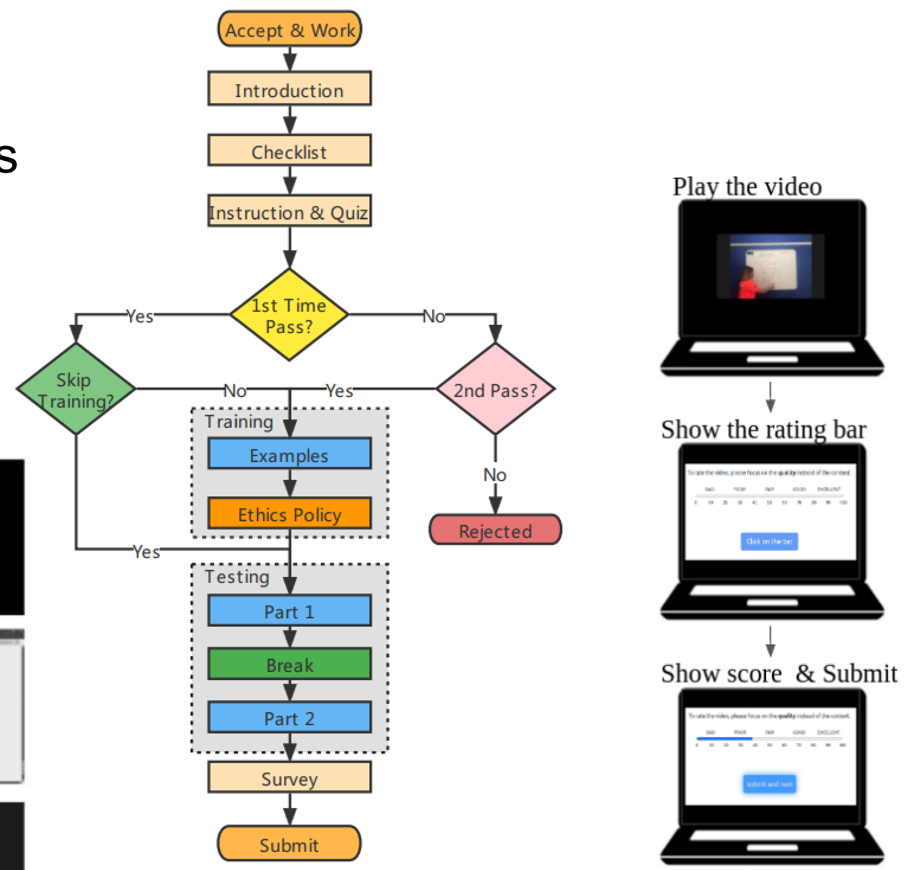

Figure: Left: flowchart of the AMT workflow experienced by crowd-sourced workers when rating telepresence videos. Right: workflow when rating a video.

Dataset Verification

We adopted a recent "soft" subject rejection model [42] that is designed to recover subjective quality scores from noisy measurements.

Rating Analysis

- Inter-subject consistency: correlation between the two sets of MOS from randomly divided subject groups. We arrived at an average SRCC of 0.765 over 50 random splits
- Intra-subject consistency: the median Linear Correlation Coefficient (LCC) between the collected MOS, against the original scores on the "golden" videos, is 0.845.

Modeling

Tele-IQA: our image model

- use pretrained MobileNetV3 backbone to extract features
- use PsRoIAlign to estimate local predictions on extracted feature maps instead of feeding patches to the network.
- view the extracted features as a multi-variate time series and feed them to a GRU-FCN.

Tele-VQA: our video model

At each time step:

- Input: one frame (F_t), one video clip (C_t), and one audio clip (A_t)
- Output: timely visual ($S^{(v)}_t$), audio ($S^{(a)}_t$) and combined audio-visual quality predictions ($S^{(a/v)}_t$).
- 1D audio signal is transformed into a 2D spectrogram via the short-time Fourier transform.
- use MobileNetV3, R(2+1)D, and YAMNet to extract features at the frame/patch, clip, and audio levels, respectively.
- jointly trained two separate pathways to process the visual and audio information.
- refer to the KPN model in ITU-T Rec. P.911 to combine visual and audio quality predictions into a single audio-visual quality prediction.

References

[19] Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing vol. 25, no. 1, pp. 372-387 (2016)
[30] Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no reference image quality assessment. In: IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR). pp. 1733–1740 (2014)
[38] Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Transactions on Image Processing 28(12), 5923–5938 (2019)
[42] Li, Z., Bampis, C.G.: Recover subjective quality scores from noisy measurements. In: 2017 Data compression conference (DCC). pp. 52–61. IEEE (2017)
[44] Lin, H., Hosu, V., Saupe, D.: Koniq-10K: Towards an ecologically valid and large-scale IQA database. arXiv preprint arXiv:1803.08489 (2018)
[54] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing vol. 21, no. 12, pp. 4695–4708 (2012)
[55] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "Completely blind" image quality analyzer. IEEE Signal Processing Letters vol. 20, pp. 209-212 (2013)
[66] Sinno, Z., Bovik, A.C.: Large-scale study of perceptual video quality. IEEE Transactions on Image Processing vol. 28, no. 2, pp. 612-627 (2019)
[69] Talebi, H., Milanfar, P.: NIMA: Neural image assessment. IEEE Transactions on Image Processing vol. 27, no. 8, pp. 3998-4011 (2018)
[72] Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: UGC-VQA: Benchmarking blind video quality assessment for user generated content (2020)
[87] Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.C.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3572–3582 (2020)

Acknowledgements

LIVE is the entity that conducted all of the data collection/experimentation. This work was supported by Meta Platforms, Inc. A.C. Bovik was supported in part by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844.

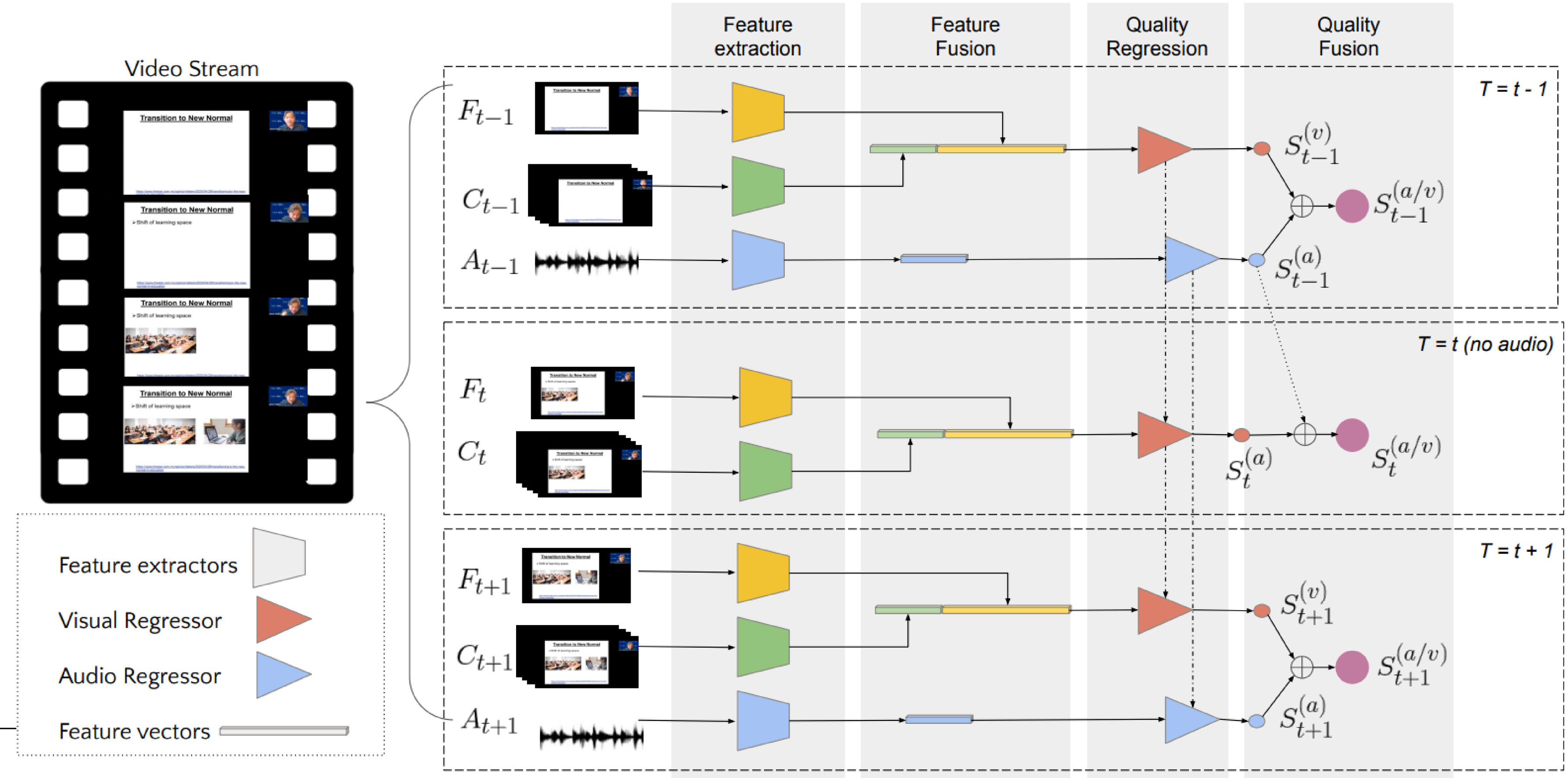

Figure: Our Tele-VQA model which involves 4 sequential steps: feature extraction, feature fusion, quality regression, and quality fusion. For video conferences, the audio signal is not guaranteed to be always available. Here we describe how we handle the case of missing audio (T = t)

Results

	CLIVE [19]		KonIQ [44]	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
NIQE [55]	0.052	0.154	0.534	0.509
BRISQUE [54]	0.495	0.494	0.641	0.596
CNNIQA [30]	0.580	0.481	0.596	0.403
NIMA [69]	0.395	0.411	0.666	0.721
P2P-FM [87]	0.756	0.783	**0.788**	**0.808**
Tele-IQA	**0.767**	**0.795**	0.772	0.800

Table: Picture quality predictions: Performance of picture quality models on different databases. A higher value indicates superior performance.

	Our database		LIVE-VQC [66]	
	SRCC	**LCC**	**SRCC**	**LCC**
IQA models				
BRISQUE [54]	0.411	0.482	0.592	0.638
TeleVQA (p)	0.476	0.488	0.621	0.603
TeleVQA (f)	0.609	0.590	0.710	0.716
VQA models				
VSFA [40]	0.601	0.655	0.773	0.795
TLVQM [38]	0.565	0.617	0.799	0.803
VIDEVAL [72]	0.536	0.560	0.752	0.751
TeleVQA (c)	0.475	0.467	0.792	0.730
TeleVQA (f+c)	0.621	0.652	0.811	0.801
TeleVQA (p+f+c)	0.633	0.672	**0.811**	**0.829**
AVQA models				
TeleVQA (a)	0.114	0.136	-	-
TeleVQA (f+a)	0.622	0.686	-	-
TeleVQA (f+c+a)	0.639	0.686	-	-
TeleVQA (p+f+c+a)	**0.663**	**0.715**	-	-

Table: Video quality predictions: Performance when all models are separately trained and tested on our database and LIVE-VQC. Here p, f, c, a means patch, frame, clip, and audio features, respectively.

The University of Texas at Austin
LIVE
Laboratory for Image & Video Engineering