August 2020

Bandwidth-efficient Video Conferencing

N/A

Bandwidth-efficient Video Conferencing

ABSTRACT

Video conferencing (VC) typically requires a high bandwidth connection, which may not be available to users in certain situations or locations. There is also sometimes an expectation, not always met, that all participants in a VC switch on their video feed so as to participate equally. This disclosure describes the use of machine-learning techniques to reduce video-conferencing bandwidth by enabling participants to display synthesized videos of themselves to other participants. Since the bulk of the task of generating a participant's video is performed on the remote device, the videoconference can be carried out with bandwidth utilization comparable to that required for audio transmission alone. With user permission, the synthetic video obtained per the techniques can be operative at the receiver end even if the sender has turned off their camera, such that the video-conferencing anxiety felt by some users is mitigated. The described techniques are implemented with specific user permission and users are provided with options to turn off video generation features.

KEYWORDS

- Video conferencing
- Generative Adversarial Networks (GAN)
- Machine learning
- Synthesized video
- Lip synchronization
- Facial expression matching
- Simulated video

BACKGROUND

Video conferencing (VC) typically requires a high bandwidth connection which may not be available to users in certain situations or locations. There is also sometimes an expectation that all participants in a VC switch on their video feed so as to participate equally. This expectation is not always met. For example, certain participants may prefer to not switch on their video feed, or may be in a situation where it may be difficult or inappropriate to do so, e.g., they may be walking while participating in the VC.

Video compression and transmission are typically optimized for general-purpose video signals that involve high levels of movement and dynamic range of colors. A video call, on the other hand, typically features a nearly static background, and a nearly static speaker, with only the lips and facial expressions of the speaker varying. In addition, many video conferences take place between individuals who know each other and conduct video conferences regularly. While current video compression algorithms leverage the fact that parts of a video change little across frames and encode only the differential motion between consecutive frames, there is still substantial redundancy in the resulting video.

Machine learning techniques used thus far to perform video compression are designed to support live-streaming, and cannot supplant the live feed of a participant who may not be in a position to transmit their own video. Machine learning has been successfully applied to image compression; however, static images are unsatisfactory in a live VC environment. Certain apps insert (on the sender's side) synthetically-generated videos of public personalities into the VC for the purposes of entertainment. However, these are not designed to and do not reduce bandwidth utilized for the video feed in a video conference.

DESCRIPTION

This disclosure describes the use of machine-learning techniques to reduce video-conferencing bandwidth by enabling participants to display synthesized videos of themselves to other participants. The described techniques are implemented with specific user permission and users are provided with options to turn off video generation features.

With user permission, A machine-learning model such as a generative adversarial network (GAN) is utilized to generate (in real-time) at a receiver a synthesized video of a sending user, synchronized with the sending user's speech. Rather than transmitting a video signal over the network, the bulk of the task of generating the sending user's video is performed on the remote device. Video conferencing is thereby carried out with bandwidth utilization that is similar to that for audio transmission alone.

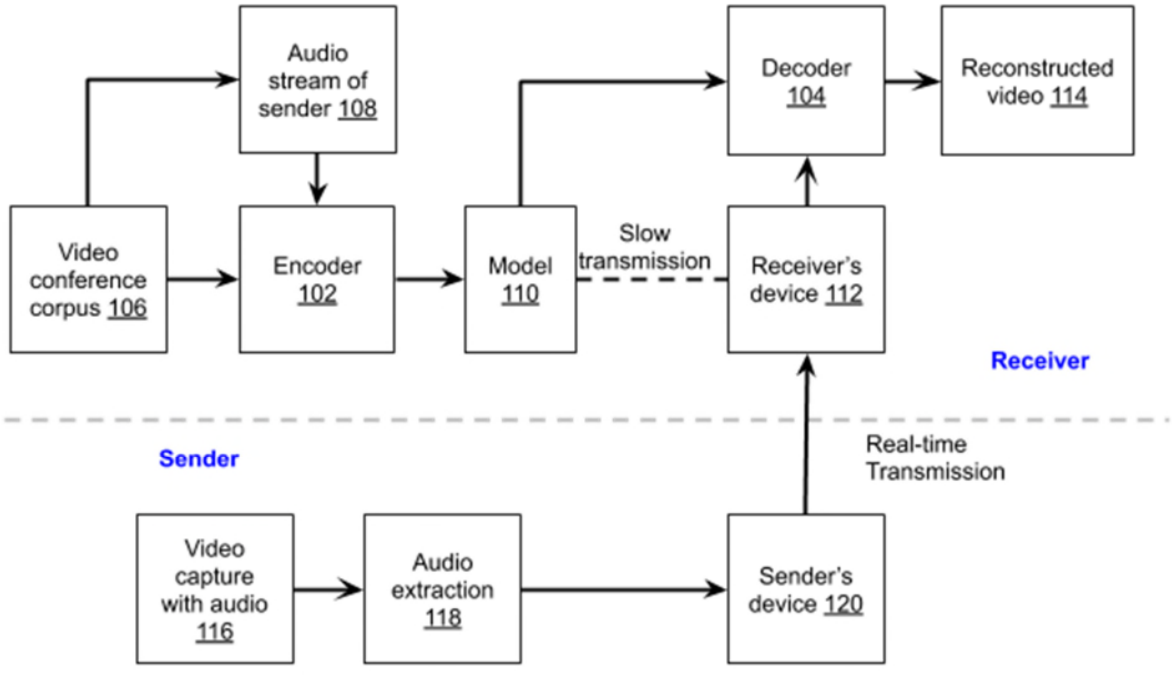

Fig. 1: Bandwidth efficient video conferencing

Fig. 1 illustrates bandwidth efficient video conferencing, per the techniques of this disclosure. At the receiver end, a matched encoder (102) and decoder (104) neural network are provided that together form a GAN. The GAN is a trained model, trained over a corpus of video conferencing feeds (106) and corresponding audio streams (108). After training, the GAN is able to generate a video stream aligned to an input audio stream.

The encoder of the GAN produces an embedding or model (110) that is provided to the decoder as its initial state. The decoder can run over a time-dependent audio input stream to produce video that is lip and face synchronized with a given audio stream.

At the sending end, a video of the sender is captured along with audio (116). Audio is extracted (118). The extracted audio signal is sent over the network from the sender's device (120) to the receiver's device (112). The decoder uses the audio signal, which originates from the sender, to construct a video (114) of the sender's face that is aligned, e.g., lip and face-movement synchronized, to the audio signal.

The described techniques of synthetic video generation to reduce transmission bandwidth can be applied bi-directionally, such that each user's device locally generates a reconstruction of the other user's video. If the users permit, synthetic video at the receiver can be operative even if the sender has turned off their camera, such that the video-conferencing anxiety (or unpreparedness to face the camera) felt by some users is mitigated.

A decrease in transmission bandwidth may be offset by an increase in the size of the model of the sender's face that is made available to the receiver. However, per the techniques, the model is not transmitted in real-time. Rather, it can be transmitted offline from the VC and relatively infrequently, possibly just on a one-time basis, at a low bitrate, over long periods of time, and in preparation for the next video call. The nature of video conferencing is that VCs

frequently occur between the same groups of individuals, such that a one-time, infrequently updated model-transmission can serve a large number of VCs.

An application of the techniques is video conferencing with recipients with devices of low processing power. Such devices may be unable to encode video in real time. Rather, per the techniques, they can locally, e.g., at the receiving end, synthesize the sender's video based on the audio received from the sender.

Another possible application of the techniques is to produce a photorealistic video conference out of a text chat by simulating the video of both or either of the participants. Further, by synthesizing audio on the receiver's device using only the sender's text messages, video conferencing effectively becomes possible using only text chats. The bandwidth required for a VC that is based off text messaging is extremely low, perhaps an order of magnitude (or more) below a traditional VC.

In implementing the described synthesis of video, users are provided with clear indicators that the video is a generated video, not an as-captured original video received from a sender's device. Further, when bandwidth is sufficient, the sender's video is received and displayed, rather than synthesizing the video locally. Still further, the generated video is marked (e.g., via watermarking techniques) or otherwise identified clearly as being a generated video. Both the receiver and the sender are provided indications when a video is synthesized and used during a videoconference. The model used for synthesizing videos is obtained and used with specific user permission. Users are provided options to control the generation, training, and use of a model that can generate a synthesized video. The model, if stored, is stored in a secure manner, and is only made available for video synthesis if permitted by the user. If any user denies permission, video generation techniques are not used in the VC.

Further to the descriptions above, a user is provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's audio/video feed, text messages, participation in a VC, a user's preferences), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user. Thus, the user has control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes the use of machine-learning techniques to reduce video-conferencing bandwidth by enabling participants to display synthesized videos of themselves to other participants. Since the bulk of the task of generating a participant's video is performed on the remote device, the videoconference can be carried out with bandwidth utilization comparable to that required for audio transmission alone. With user permission, the synthetic video obtained per the techniques can be operative at the receiver end even if the sender has turned off their camera, such that the video-conferencing anxiety felt by some users is mitigated. The described techniques are implemented with specific user permission and users are provided with options to turn off video generation features.

REFERENCES

[1] Gopalakrishnan, Manikandan, "Video Stream Simulator," Technical Disclosure Commons, (July 17, 2018) https://www.tdcommons.org/dpubs_series/1322 accessed on Aug. 3, 2020.

[2] Ben Munson, "Disney unveils AI-powered video compression research" https://www.fiercevideo.com/tech/disney-unveils-ai-powered-video-compression-research accessed on Aug. 3, 2020.

[3] Joon Ian Wong, "Netflix's new AI tweaks each scene individually to make video look good even on slow internet" https://qz.com/920857/netflix-nflx-uses-ai-in-its-new-codec-to-compress-video-scene-by-scene/ accessed on Aug. 3, 2020.

[4] Nick Johnston and David Minner, "Image compression with neural networks," https://ai.googleblog.com/2016/09/image-compression-with-neural-networks.html accessed on Aug. 3, 2020.

[5] "Avatars for Zoom, Skype, and other video-conferencing apps," https://github.com/alievk/avatarify accessed on Aug. 3, 2020.

[6] Nguyen, Thanh Thi, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. "Deep learning for deepfakes creation and detection." *arXiv preprint arXiv:1909.11573* (2019).

[7] Pablo Barrera and Florian Stimberg,"Improving audio quality in Duo with WaveNet," https://ai.googleblog.com/2020/04/improving-audio-quality-in-duo-with.html accessed on Aug. 3, 2020.