November 2023

Meeting Summarization and Alerts for Video Conference Participants

Shiblee Hasan

Kathleen Bryan

Meeting Summarization and Alerts for Video Conference Participants

During video conferences, participants can get distracted and miss a part of a conversation. Though video conferencing platforms may provide a recording of the video conference for reference after the meeting has ended, this feature does not help participants during the video conference. Getting distracted and missing a part of the conversation can make it difficult for distracted participants to reenter the conversation during the video conference. In some instances, other participants may have to spend time repeating or summarizing what was discussed while the participant was distracted.

Therefore, a technique is proposed for providing video conference summaries for video conference participants during the video conference (also known as a virtual meeting, a video chat, etc.). The technique can include generating a running summary of the video conference while it is being held. Then, when a participant is identified as distracted, the technique can provide a portion of the running summary to the distracted participant. The portion can correspond to the period during which the participant was identified as distracted. This technique enables distracted participants to easily reenter the conversation during a video conference.

The disclosed technique can be implemented as processing logic (e.g., of a component such as a computer program or a module of a computer program) that may be integrated into a video conferencing platform and/or a client device. The processing logic may receive multiple frames and/or audio representing a video recording of a participant during a video conference. If a participant's camera and/or microphone are turned off, the processing logic may also receive a notification that such data is unavailable.

The processing logic may analyze frames and/or audio data to determine if the participant is distracted during a video conference. For example, the processing logic can analyze the frames to determine the participant’s body movements. If the participant’s body movements indicate that the participant is looking away from the camera or gesturing towards something outside the range of the camera, the processing logic can determine that the participant is distracted. As another example, the processing logic can determine if the participant is distracted based on the participant’s camera settings. For example, processing logic can determine that a participant is distracted if they turn off their camera after having it on for at least a threshold period of time during the video conference. In some instances, the determination of whether a participant is distracted may be done using a machine learning model, as described in detail below with respect to figure 1.

Once the processing logic has determined that a participant is distracted, the processing logic may retrieve a summary of the meeting for the period of time that the participant is distracted. For example, if the participant is determined to be distracted, the processing logic will begin retrieving a summary of the video conference to display on the participant’s device from the moment the participant is determined to be distracted. The participant will continue to receive the running summary until the processing logic determines the participant is no longer distracted or the participant manually stops the summary from being displayed. The summary may include alerts of action items for the participant to address, such as questions directed to the participant. The processing logic may utilize a generative artificial intelligence (AI) model to generate the summary of the video conference, as described in detail below with respect to figure 2.

Figure 1 illustrates a flow diagram of a method 100 for providing video conference summaries to distracted participants. At block 110, processing logic may receive, from a user device of a participant, multiple frames and audio representing a video stream of the participant during a specific time period of a video conference. In some embodiments, the processing logic may be separate from the video conference platform and can receive the multiple frames and audio from the video conference platform and/or the user device.

At block 120, the processing logic may feed the multiple frames and audio data into a machine learning model trained to assess if the participant is distracted based on factors such as eye movement, body movement, camera settings, and/or audio. In some instances, a training engine can train the machine learning model using training data that includes training inputs and corresponding target outputs (correct answers for respective training inputs). For example, the training data may include examples of attentive or distracted behavior (e.g.. nodding to show attention, looking away to show distraction, etc.) to predict whether a participant is distracted. The training engine may find patterns in the training data that map the training input to the target output (the prediction of whether the participant is distracted) and train the machine learning model according to these patterns.

In an illustrative example, multiple frames fed into the machine learning model can depict that a participant is looking down and typing while nodding occasionally to what is said. While looking down and typing may be seen as distracted behavior, the act of nodding while looking down and typing shows that a participant is paying attention. Accordingly, the machine learning model can determine that the participant is not distracted.

The machine learning model can be composed of, e.g., a single level of linear or non-linear operations (e.g. a support vector machine (SVM)) or a deep network, e.g., a machine learning model that is composed of multiple levels of non-linear operations. An example of a deep network is a neural network with one or more hidden layers, and such a machine learning model may be trained by, for example, adjusting weights of a neural network in accordance with a back propagation learning algorithm or the like. Once the model is trained, it can be used to compare the behavior of a participant (e.g. eye movement, body movement, camera settings, and audio) to behavior in the training dataset to determine if the participant is showing distracted behavior.

At block 130, based on the output of the machine learning model, the processing logic can determine if the participant is distracted for the period of time correlating to the multiple frames. If the participant is determined to be distracted, the processing logic can retrieve a summary of the meeting for the time period that the participant is distracted, as depicted in block 140. The summary can be retrieved from a data store or using an output of a generative AI model, as discussed below in figure 2.

At block 150, the processing logic provides the summary for display on the user device of the participant. The summary may first be displayed as a pop-up notification that the participant can expand to view the full summary of the time they were distracted. In some embodiments, the summary can include alerts that notify the participant of any action items they need to address. For example, if some participants introduced themselves while a participant was distracted, the summary can include an action item notifying the distracted participant to give an introduction. An alert can include an audio notification, bolded or highlighted text, and/or some other method form of notification.

Figure 2 illustrates a flow diagram of a method 200 for generating a summary for a video conference. The summary of the video conference can be continuously generated throughout the video conference. At block 210, the processing logic may identify conference data such as audio input, chat comments, and/or participant facial expressions. The audio input, chat comments, or participant expressions can correspond to reactions and statements made by participants during a specific time period of the video conference. For example, audio input, chat comments, and participant expressions from a 10 second duration of the video conference can be fed into the generative AI model, which generates a summary for that 10 second duration of the video conference. The chat comments can include text comments, emojis, message reactions, gifs, and/or images.

At block 220, the processing logic feeds the conference data as input to a generative AI model trained to generate a summary of the meeting based on inputs including audio input, chat comments (e.g., text comments, emojis, etc.), and/or participant expressions. In some embodiments, the processing logic also provides a prompt with a request to generate a summary of the meeting as input to the generative AI model. In some embodiments, the generative AI model can use a transformer-based model architecture with a self-attention mechanism. The generative AI model can comprise an artificial neural network, composed of artificial neurons or nodes connected by weights. A positive weight reflects a relevant connection, while a negative weight reflects irrelevant connections. Through training, the generative AI model can adjust the weights to minimize the difference between predicted and desired outputs.

The generative AI model can be trained by gathering text data from various sources and tokenizing the data. The generative AI model can learn the probabilities of token sequences by using unsupervised learning to predict the next token in a sequence given the context of previous tokens with the goal of reducing the difference between predicted token probabilities and actual tokens. Once the model is trained, it can be used to process video conference data (e.g., transcript of the audio input, chat comments, etc.) and generate an accurate summary of the video conference. The generative AI model may be hosted by a video conferencing platform, a separate server, or a client device.

Participant expressions can be used by the generative AI model to summarize the reaction of specific participants to certain topics or statements during the video conference. For example, if a participant frowns in response to a statement, the summary can include that the participant was upset or concerned by the statement. Alternatively, if a participant smiles at something, the summary can include that the participant was delighted by the statement. In some embodiments, the generative AI model can record a participant's expression instead of interpreting the participant's emotion based off of their expression. For example, the generative AI model can include in the summary that a participant smiled at a statement instead of stating in the summary that the participant was delighted by the statement.

In some embodiments, the generative AI model can transcribe the audio input into a written transcript for the purpose of generating the video conference summary. The transcript can be combined with the participant's expressions and/or chat comments to provide a comprehensive summary. In some embodiments, the generative AI model will generate a summary by paraphrasing and using its own words to capture the main ideas of the video conference. In other embodiments, the generative AI model can select the most important sentences and phrases from the inputs and combine them to form the summary. The generated summary can include timestamps to easily identify what portion of the video conference is being summarized. In some embodiments, the generated summary can be stored in a separate data store.

At block 230, the processing logic receives a request to provide a summary to a participant device. In some embodiments, the request to provide a summary can come from a determination that a participant is distracted, as described in block 140 of figure 1. In other embodiments, the request may come from a participant manually requesting, via the UI of the video conference platform, a summary of a video conference for a specified period of time. The request can include a portion of time during which the participant is identified as being distracted. Responsive to receiving the request, the processing logic identifies the portion of the generated summary based on the time period indicated in the request. For example, the portion of the generated summary to provide to the participant may be identified by timestamps in the summary. The generated summary may be retrieved from the generative AI model or a data store.

By determining when a participant is distracted and summarizing what topics were discussed during the period of time the participant was distracted, the disclosed technique helps distracted participants more easily transition back to conversations during video conferences.

Abstract

A technique is proposed for summarizing meetings to help distracted participants transition back into conversations during video conferences. Processing logic may receive, from a video conferencing platform, a video file including multiple frames and audio. The processing logic may determine a period of time where a participant is distracted based on the participant's body movement, eye movement, and camera settings. This determination can be made using a machine learning model. The processing logic can also generate a summary of the meeting based on received voice input, chat comments, and participant expressions. This summary may be generated using a generative artificial intelligence model. The processing logic can send a summary of the meeting to a participant for the time period during which they were distracted. This results in helping distracted participants integrate back into the conversation more easily during a video conference.

Keywords: virtual meeting, video conference, eye tracking, head tracking, body movement tracking, notification system, meeting summarization

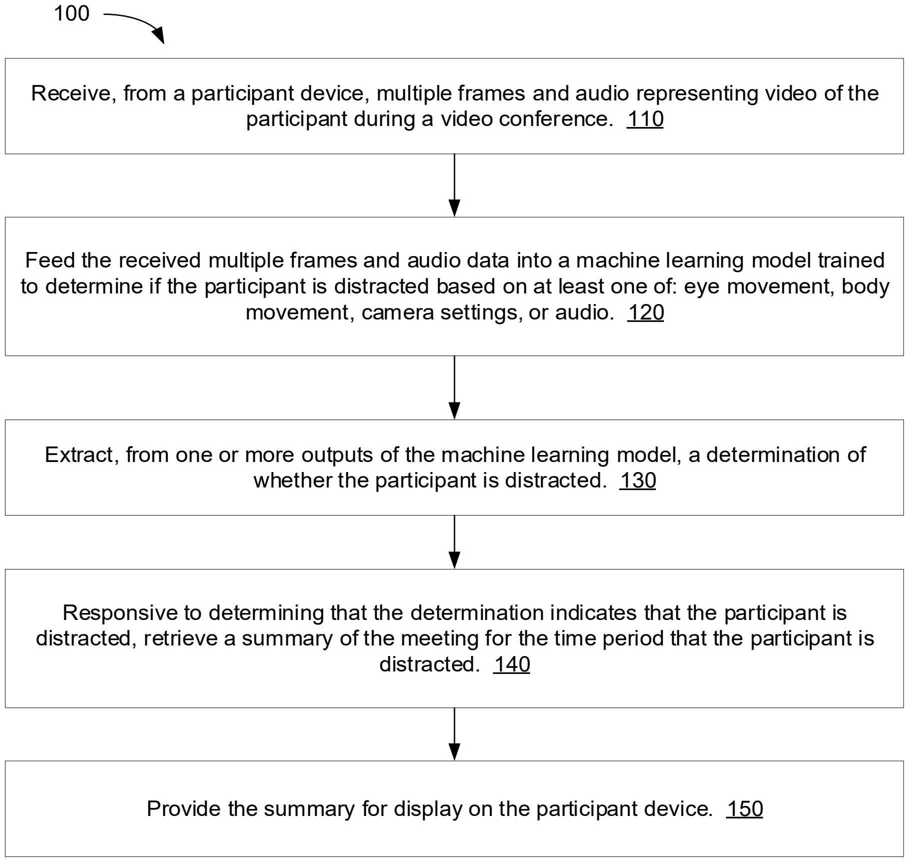
100
Receive, from a participant device, multiple frames and audio representing video of the participant during a video conference. 110
Feed the received multiple frames and audio data into a machine learning model trained to determine if the participant is distracted based on at least one of: eye movement, body movement, camera settings, or audio. 120
Extract, from one or more outputs of the machine learning model, a determination of whether the participant is distracted. 130
Responsive to determining that the determination indicates that the participant is distracted, retrieve a summary of the meeting for the time period that the participant is distracted. 140
Provide the summary for display on the participant device. 150

Figure 1

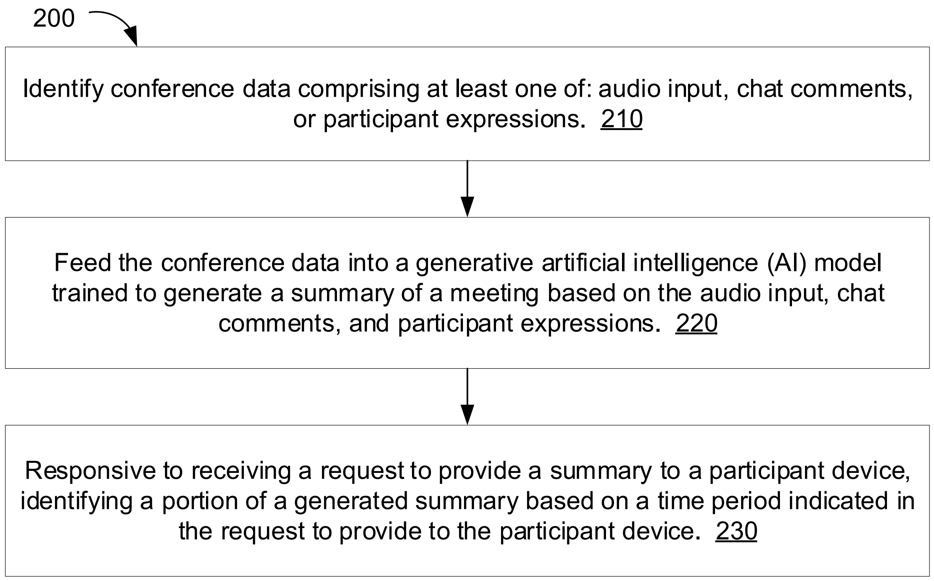
200
Identify conference data comprising at least one of: audio input, chat comments, or participant expressions. 210
Feed the conference data into a generative artificial intelligence (AI) model trained to generate a summary of a meeting based on the audio input, chat comments, and participant expressions. 220
Responsive to receiving a request to provide a summary to a participant device, identifying a portion of a generated summary based on a time period indicated in the request to provide to the participant device. 230

Figure 2