May 2020

Generating Condensed Videos

Victor Carbune

Alexandru Damian

GENERATING CONDENSED VIDEOS

ABSTRACT

Disclosed herein is an improved mechanism for generating condensed videos. The mechanism can identify a video for which a condensed video is to be generated. The mechanism can generate a summary transcript of the video. The mechanism can identify frames of the video that are deemed most relevant to portions of the generated summary transcript. Based on the identified frames of the video that are deemed most relevant to portions of the generated summary transcript, the mechanism can generate additional frames of video content that are used to connect the identified relevant frames of the video. The mechanism can then generate a condensed video by stitching together the identified frames of the video and the generated additional frames of video content.

BACKGROUND

A user viewing a video may want to view a video faster, for example, to consume a video in less time than the video's length. Options for allowing for faster playback may include allowing a user to skip scenes (e.g., to skip forward a particular duration of time, such as ten seconds) or increasing a playback rate of the video. In some cases, these options can have disadvantages. For example, skipping forward in a video can require continuous user input to continue skipping ahead in the video. As another example, increasing a playback rate of the video can lead to unnatural video content, such as a higher voice pitch, people moving too fast, etc. Thus, there is a need for a better method to generate condensed videos.

DESCRIPTION

The systems and techniques described in this disclosure relate to generating condensed videos. The system can be implemented on a server, such as a server associated with a video sharing service that hosts videos and/or provides videos to user devices for viewing.

Note that the systems and techniques described in this disclosure can be used in any suitable application. For example, in generating condensed videos, the system can enable a user consuming the video to consume a video in less time than the original length of the video with a fluent video at normal speed.

FIG. 1 illustrates an example process for generating condensed videos.

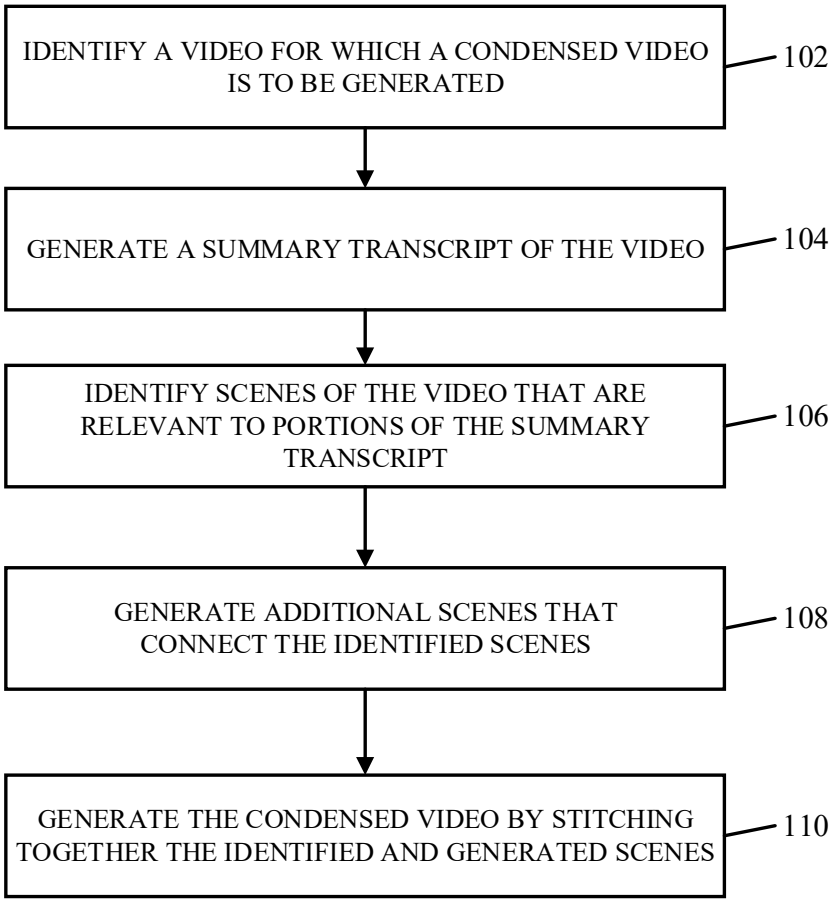

FIG. 1

At 102, the system can identify a video for which a condensed video is to be generated. The system can identify the video based on any suitable information. For example, the system can iterate through a group of videos (e.g., a group of videos that have been uploaded to a video sharing service, a group of videos for which highlights of the video are to be generated, and/or any other suitable group of videos). As another example, the system can identify the video based on a user input that indicates the video. As a more particular example, a creator of the video who has uploaded the video to a video sharing service can indicate that a condensed video is to be generated.

At 104, the system can generate a summary transcript of the video. The system can generate the summary transcript in any suitable manner. For example, in some instances, the system can use a neural network to generate the summary transcript. As a more particular example, the system can use a neural network that has been trained to take, as an input, a video that includes any suitable audio content and/or video content, and to identify portions of the video that are particularly relevant. As a specific example, the neural network can identify particular scenes of the video or groups of frames of the video that contain characters of interest, action, and/or any other suitable types of content. As another more particular example, in some instances, the neural network can take, as an input, a video that includes any suitable audio content and/or video content, and can produce, as an output, a condensed video that has been condensed to include scenes of the video identified by the neural network as important to the video.

The system can then generate the summary transcript of the video using the condensed portions of the video in any suitable manner. For example, the system can identify a portion of audio content that corresponds to each portion of the video identified by the neural network

at 102. As a more particular example, in an instance in which a portion of the video identified by the neural network corresponds to a particular group of frames of the video, the system can identify audio content that corresponds to the group of frames of the video. The system can then generate the summary transcript based on the identified audio content that corresponds to the portion of the video. For example, the system can use any suitable speech recognition techniques to identify spoken language included in the audio content. As another example, the system can use any suitable audio processing techniques to identify portions of the audio content that include non-speech sounds, such as music, laughter, etc. In some instances, the system can then generate a transcript that represents the audio content included in each identified portion of the video. For example, the generated transcript can indicate speech included in the audio content, times at which laughter or music occur in the audio content, etc. In some instances, the system can then generate the summary transcript by aggregating the generated transcripts for each portion of audio content corresponding to each identified portion of the video.

At 106, the system can identify or otherwise extract scenes of the video that are relevant to portions of the summary transcript. For example, in instances in which the summary transcript includes a conversation between two or more characters, the system can identify any suitable number of frames of the video that include the two or more characters participating in the conversation. As another example, in an instance in which the summary transcript indicates that a particular action or event occurred that is likely to be relatively important, the system can identify any suitable number of frames of the video that include characters participating in the action or event or video content corresponding to occurrence of the action or the event. Note that, in some instances the system can additionally identify any suitable frames of the video that include reactions of any suitable characters to the action or event.

Note that the system can identify the relevant scenes in any suitable manner. For example, in some instances, the system can identify scenes in which particular words, phrases, or sentences that appear in the summary transcript are spoken by one or more characters. In some instances, the system can identify particular words or phrases that are particularly important to the summary transcript before identifying the relevant scenes. For example, in some instances, the system can identify particularly common words (e.g., particularly common words related to location, particularly common character names, etc.) that occur in the summary transcript. As another example, in some instances, the system can identify words that correspond to a particular grammar object (e.g., a subject of a sentence, a predicate of a sentence, etc.) as being particularly relevant. Note that, in some instances, the system can identify particular words or phrases that are deemed relevant to the summary transcript based on any suitable combination of factors, such as a frequency that words appear in the summary transcript, a type of grammar object corresponding to the word, and/or any other suitable combination.

Note also that, in some instances, the system can use any suitable type of machine learning techniques to further condense a summary transcript prior to identifying the relevant scenes. For example, in some instances, the system can use any suitable type of machine learning algorithms to re-write the summary transcript in a more condensed format. As another example, in some instances, the system can use any suitable type of machine learning algorithms to identify portions of the summary transcript that are redundant or relatively unimportant and that can therefore be removed from the summary transcript. As a more particular example, in some instances the system can identify portions of the summary transcript that correspond to speech that includes pauses or filler words. As another more particular example, in some instances the system can identify portions of the summary transcript that correspond to a closing

credits scene or other scene that is relatively unimportant to a plot of the video. In some such instances, the system can then remove portions of the summary transcript that correspond to the identified redundant or relatively unimportant portions prior to identifying the scenes of the video relevant to the summary transcript.

At 108, the system can generate additional scenes that connect the scenes identified as relevant to the summary transcript. In some instances, the generated additional scenes can include any suitable number of frames of video content that can allow one identified scene to connect to another identified smoothly, that is, without jumps or gaps in the video content. Note that, in some instances, the system can generate any suitable audio content that corresponds to the generated additional video content that allows audio content corresponding to one identified scene to connect to audio content of another identified scene. For example, in some instances, the system can generate audio content that corresponds to generated video content. As another example, in some instances, the system can generate connecting audio content that allows audio content from an identified scene to connect to audio content of another identified scene. As a more particular example, the system can generate audio content that connects audio content of two identified scenes such that the generated audio content includes missing words in speech between the two identified scenes.

In some instances, the system can generate the additional scenes in any suitable manner and using any suitable technique(s). For example, in some instances, the system can generate any suitable number of frames corresponding to the additional scenes using a trained Generative Adversarial Network (GAN). In some instances, the GAN can include a generator network that is trained, using examples of real video content and/or real audio content, to generate fake video content and/or fake audio content, and a discriminator network that is trained to discriminate real

video content and/or real audio content from fake video content and/or fake audio content generated by the generator network. In some instances, by training the generator network and the discriminator network in connection with each other, the trained GAN can generate realistic video content and/or audio content corresponding to the additional scenes to be used to connect the identified scenes of the video.

Note that, in some instances, the system can modify any of the scenes identified at 106 or the additional scenes generated at 108 in any suitable manner. For example, in some instances, the system can modify any of the scenes such that the video content in the scenes and audio content associated with the video content line up. As a more particular example, in some instances, the system can time shift audio content such that speech included in the audio content lines up with movement of a character's mouth as included in the video content.

At 110, the system can generate the condensed video by stitching together the scenes identified at 106 with the additional scenes generated at 108. In some instances, the system can stitch the scenes together in any suitable manner. For example, in some instances, the system can append a generated scene to a first identified scene (e.g., a first scene identified at 106 as relevant to the summary transcript), and can append a second identified scene (e.g., a second scene identified at 106 as relevant to the summary transcript) such that the condensed video includes the first identified scene, the generated scene, and the second identified scene. In some instances, the system can then iterate through all of the scenes identified at 106 and all of the scenes generated at 108 to stich together the scenes in a correct ordering.

In some instances, the system can store the condensed video in connection with the video. For example, in some instances, the system can assign the condensed video an identifier that links the condensed video to an identifier associated with the video.

Note that, in some instances, the system can make the condensed video available in any suitable manner and for any suitable purpose. For example, in some instances, the system can cause the condensed video to be presented to a user viewing the video who indicates that they are interested in viewing the video at a faster playback rate. As another example, in some instances, the system can cause the condensed video to be presented as a highlight or preview of the video, for example, in a page associated with a channel of content, in a social networking post that includes a link to the video etc. Note that, in some instances, the system can generate multiple condensed versions of the video. For example, in some instances the system can generate a first condensed version of the video with a particular duration (e.g., ten seconds, thirty seconds, two minutes, and/or any other suitable duration), and a second condensed version of the video with a different duration. In some such instances, the system can cause the different versions of the condensed video to be presented at different times. For example, a shorter version can be presented as a highlight or preview of the video, whereas a longer version can be presented in response to a request to view the video at a faster playback rate.

Accordingly, a mechanism for generating condensed videos is provided.